ACADEMIC PRESS SERIES IN BIOMEDICAL ENGINEERING

Introduction to BIOMEDICAL ENGINEERING

Third Edition

JOHN ENDERLE
JOSEPH BRONZINO

INTRODUCTION TO BIOMEDICAL ENGINEERING

THIRD EDITION

This is a volume in the
ACADEMIC PRESS SERIES IN BIOMEDICAL ENGINEERING

JOSEPH BRONZINO, SERIES EDITOR
Trinity College—Hartford, Connecticut

INTRODUCTION TO BIOMEDICAL ENGINEERING

THIRD EDITION

JOHN D. ENDERLE
University of Connecticut
Storrs, Connecticut

JOSEPH D. BRONZINO
Trinity College
Hartford, Connecticut

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier

30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

MATLAB® and Simulink® are trademarks of The MathWorks, Inc. and are used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® and Simulink® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® and Simulink® software.

Library of Congress Cataloging-in-Publication Data
Introduction to biomedical engineering / [edited by] John Enderle, Joseph Bronzino. – 3rd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-12-374979-6 (alk. paper)
1. Biomedical engineering. I. Enderle, John D. (John Denis) II. Bronzino, Joseph D., 1937-
[DNLM: 1. Biomedical Engineering. QT 36]
R856.I47 2012
610.28–dc22 2010046267

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

For information on all Academic Press publications
visit our Web site at www.elsevierdirect.com

Printed in the United State of America
11 12 13 14 9 8 7 6 5 4 3 2 1

This book is dedicated to our families.

This page intentionally left blank

Contents

6. Tissue Engineering

RANDALL E. MCCLELLAND, ROBERT DENNIS, LOLA M. REID, JAN P. STEGEMANN, BERNARD PALSSON, AND JEFFREY M. MACDONALD

7. Compartmental Modeling

JOHN D. ENDERLE

8. Biochemical Reactions and Enzyme Kinetics

JOHN D. ENDERLE

9. Bioinstrumentation

JOHN D. ENDERLE

10. Biomedical Sensors

YITZHAK MENDELSON

11. Biosignal Processing

MONTY ESCABI

This page intentionally left blank

Preface

The purpose of the third edition remains the same as the first and second editions, that is, to serve as an introduction to and overview of the field of biomedical engineering. Many chapters have undergone major revision from the previous editions with new end-of-chapter problems added. Some chapters were eliminated completely, with several new chapters added to reflect changes in the field.

Over the past fifty years, as the discipline of biomedical engineering has evolved, it has become clear that it is a diverse, seemingly all-encompassing field that includes such areas as bioelectric phenomena, bioinformatics, biomaterials, biomechanics, bioinstrumentation, biosensors, biosignal processing, biotechnology, computational biology and complexity, genomics, medical imaging, optics and lasers, radiation imaging, tissue engineering, and moral and ethical issues. Although it is not possible to cover all of the biomedical engineering domains in this textbook, we have made an effort to focus on most of the major fields of activity in which biomedical engineers are engaged.

The text is written primarily for engineering students who have completed differential equations and a basic course in statics. Students in their sophomore year or junior year should be adequately prepared for this textbook. Students in the biological sciences, including those in the fields of medicine and nursing can also read and understand this material if they have the appropriate mathematical background.

Although we do attempt to be fairly rigorous with our discussions and proofs, our ultimate aim is to help students grasp the nature of biomedical engineering. Therefore, we have compromised when necessary and have occasionally used less rigorous mathematics in order to be more understandable. A liberal use of illustrative examples amplifies concepts and develops problem-solving skills. Throughout the text, MATLAB® (a matrix equation solver) and SIMULINK® (an extension to MATLAB® for simulating dynamic systems) are used as computer tools to assist with problem solving. The Appendix provides the necessary background to use MATLAB® and SIMULINK®. MATLAB® and SIMULINK® are available from:

The Mathworks, Inc.
24 Prime Park Way
Natick, Massachusetts 01760
Phone: (508) 647-7000
Email: info@mathworks.com
WWW: http://www.mathworks.com

Chapters are written to provide some historical perspective of the major developments in a specific biomedical engineering domain as well as the fundamental principles that underlie biomedical engineering design, analysis, and modeling procedures in that domain. In addition, examples of some of the problems encountered, as well as the techniques used to solve them, are provided. Selected problems, ranging from simple to difficult, are presented at the end of each chapter in the same general order as covered in the text.

The material in this textbook has been designed for a one-semester, two-semester, or three-quarter sequence depending on the needs and interests of the instructor. Chapter 1 provides necessary background to understand the history and appreciate the field of biomedical engineering. Chapter 2 presents the vitally important chapter on biomedically based morals and ethics. Basic anatomy and physiology are provided in Chapter 3. Chapters 4–11 provide the basic core biomedical engineering areas: biomechanics, biomaterials, tissue engineering, compartmental modeling, biochemical reactions, bioinstrumentation, biosensors, and biosignal processing. To assist instructors in planning the sequence of material they may wish to emphasize, it is suggested that the chapters on bioinstrumentation, biosensors and biosignal processing should be covered together as they are interdependent on each other. The remainder of the textbook presents material on biomedical systems and biomedical technology (Chapters 12–17).

Readers of the text can visit http://www.elsevierdirect.com/9780123749796 to view extra material that may be posted there from time to time.

Instructors can register at http://www.textbooks.elsevier.com for access to solutions and additional resources to accompany the text.

ACKNOWLEDGMENTS

Many people have helped us in writing this textbook. Well deserved credit is due to the many contributors who provided chapters and worked under a very tight timeline. Special thanks go to our publisher, Elsevier, especially for the tireless work of the Publisher, Joseph Hayton and Associate Editor, Steve Merken. In addition, we appreciate the work of Lisa Lamenzo, the Project Manager.

A great debt of gratitude is extended to Joel Claypool, the editor of the first edition of the book and Diane Grossman from Academic Press, and Christine Minihane, the editor of the second edition. Also, we wish to acknowledge the efforts of Jonathan Simpson, the first editor of this edition, who moved onto to other assignments before this project was complete.

A final and most important note concerns our co-author of the first two editions of this book, Susan Blanchard. She decided that she wanted to devote more time to her family and not to continue as a co-author.

Contributors to the Third Edition

Susan M. Blanchard Florida Gulf Coast University, Fort Meyers, Florida

Joseph D. Bronzino Trinity College, Hartford, Connecticut

Stanley A. Brown Food and Drug Administration, Gaithersburg, Maryland

Gerard L. Coté Texas A&M University, College Station, Texas

Robert Dennis University of North Carolina, Chapel Hill, North Carolina

John Enderle University of Connecticut, Storrs, Connecticut

Monty Escabí University of Connecticut, Storrs, Connecticut

Liisa T. Kuhn University of Connecticut Health Center, Farmington, Connecticut

Jeffrey M. Macdonald University of North Carolina-Chapel Hill, Chapel Hill, North Carolina

Randall McClelland University of North Carolina, Chapel Hill, North Carolina

Yitzhak Mendelson Worcester Polytechnic Institute, Worcester, Massachusetts

Katharine Merritt Food and Drug Administration, Gaithersburg, Maryland

Gerald E. Miller Virginia Commonwealth University, Richmond, Virginia

Joseph Palladino Trinity College, Hartford, Connecticut

Bernard Palsson University of California at San Diego, San Diego, California

Sohi Rastegar National Science Foundation, Arlington, Virginia

Lola M. Reid University of North Carolina, Chapel Hill, North Carolina

Kirk K. Shung University of Southern California, Los Angeles, California

Jan P. Stegemann University of Michigan, Ann Arbor, Michigan

Thomas Szabo Boston University, Boston, Massachusetts

LiHong V. Wang Washington University in St. Louis, St. Louis, Missouri

Contributors to the Second Edition

Susan M. Blanchard Florida Gulf Coast University, Fort Meyers, Florida

Joseph D. Bronzino Trinity College, Hartford, Connecticut

Stanley A. Brown Food and Drug Administration, Gaithersburg, Maryland

Gerard L. Coté Texas A&M University, College Station, Texas

Charles Coward Drexel University, Philadelphia, Pennsylvania

Roy B. Davis III Shriners Hospital for Children, Greenville, South Carolina

Robert Dennis University of North Carolina, Chapel Hill, North Carolina

John Enderle University of Connecticut, Storrs, Connecticut

Monty Escabí University of Connecticut, Storrs, Connecticut

Robert J. Fisher University of Massachusetts, Amherst, Massachusetts

Liisa T. Kuhn University of Connecticut Health Center, Farmington, Connecticut

Carol Lucas University of North Carolina-Chapel Hill, Chapel Hill, North Carolina

Jeffrey M. Macdonald University of North Carolina-Chapel Hill, Chapel Hill, North Carolina

Amanda Marley North Carolina State University, Raleigh, North Carolina

Randall McClelland University of North Carolina, Chapel Hill, North Carolina

Yitzhak Mendelson, PhD Worcester Polytechnic Institute, Worcester, Massachusetts

Katharine Merritt Food and Drug Administration, Gaithersburg, Maryland

Spencer Muse North Carolina State University, Raleigh, North Carolina

H. Troy Nagle North Carolina State University, Raleigh, North Carolina

Banu Onaral Drexel University, Philadelphia, Pennsylvania

Joseph Palladino Trinity College, Hartford, Connecticut

Bernard Palsson University of California at San Diego, San Diego, California

Sohi Rastegar National Science Foundation, Arlington, Virginia

Lola M. Reid University of North Carolina, Chapel Hill, North Carolina

Kirk K. Shung University of Southern California, Los Angeles, California

Anne-Marie Stomp North Carolina State University, Raleigh, North Carolina

Thomas Szabo Boston University, Boston, Massachusetts

Andrew Szeto San Diego State University, San Diego, California

LiHong V. Wang Washington University in St. Louis, St. Louis, Missouri

Melanie T. Young North Carolina State University, Raleigh, North Carolina

Contributors to the First Edition

Susan M. Blanchard Florida Gulf Coast University, Fort Meyers, Florida

Joseph D. Bronzino Trinity College, Hartford, Connecticut

Stanley A. Brown Food and Drug Administration, Gaithersburg, Maryland

Gerard L. Coté Texas A&M University, College Station, Texas

Roy B. Davis III Shriners Hospital for Children, Greenville, South Carolina

John Enderle University of Connecticut, Storrs, Connecticut

Robert J. Fisher University of Massachusetts, Amherst, Massachusetts

Carol Lucas University of North Carolina-Chapel Hill, Chapel Hill, North Carolina

Amanda Marley North Carolina State University, Raleigh, North Carolina

Yitzhak Mendelson, PhD Worcester Polytechnic Institute, Worcester, Massachusetts

Katharine Merritt Food and Drug Administration, Gaithersburg, Maryland

H. Troy Nagle North Carolina State University, Raleigh, North Carolina

Joseph Palladino Trinity College, Hartford, Connecticut

Bernard Palsson University of California at San Diego, San Diego, California

Sohi Rastegar National Science Foundation, Arlington, Virginia

Daniel Schneck Virginia Polytechnic Institute & State University, Blacksburg, Virginia

Kirk K. Shung University of Southern California, Los Angeles, California

Anne-Marie Stomp North Carolina State University, Raleigh, North Carolina

Andrew Szeto San Diego State University, San Diego, California

LiHong V. Wang Washington University in St. Louis, St. Louis, Missouri

Steven Wright Texas A&M University, College Station, Texas

Melanie T. Young North Carolina State University, Raleigh, North Carolina

This page intentionally left blank

CHAPTER

1

Biomedical Engineering: A Historical Perspective

Joseph D. Bronzino, PhD, PE

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Identify the major role that advances in medical technology have played in the establishment of the modern health care system.
- Define what is meant by the term *biomedical engineering* and the roles biomedical engineers play in the health care delivery system.
- Explain why biomedical engineers are professionals.

In the industrialized nations, technological innovation has progressed at such an accelerated pace that it has permeated almost every facet of our lives. This is especially true in the area of medicine and the delivery of health care services. Although the art of medicine has a long history, the evolution of a technologically based health care system capable of providing a wide range of effective diagnostic and therapeutic treatments is a relatively new phenomenon. Of particular importance in this evolutionary process has been the establishment of the modern hospital as the center of a technologically sophisticated health care system.

Since technology has had such a dramatic impact on medical care, engineering professionals have become intimately involved in many medical ventures. As a result, the discipline of *biomedical engineering* has emerged as an integrating medium for two dynamic professions—medicine and engineering—and has assisted in the struggle against illness and disease by providing tools (such as biosensors, biomaterials, image processing, and artificial intelligence) that health care professionals can use for research, diagnosis, and treatment.

Thus, biomedical engineers serve as relatively new members of the health care delivery team that seeks new solutions for the difficult problems confronting modern society. The purpose of this chapter is to provide a broad overview of technology's role in shaping our modern health care system, highlight the basic roles biomedical engineers play, and present a view of the professional status of this dynamic field.

1.1 THE EVOLUTION OF THE MODERN HEALTH CARE SYSTEM

Primitive humans considered diseases to be "visitations"—the whimsical acts of affronted gods or spirits. As a result, medical practice was the domain of the witch doctor and the medicine man and medicine woman. Yet even as magic became an integral part of the healing process, the cult and the art of these early practitioners were never entirely limited to the supernatural. Using their natural instincts and learning from experience, these individuals developed a primitive science based upon empirical laws. For example, through acquisition and coding of certain reliable practices, the arts of herb doctoring, bone setting, surgery, and midwifery were advanced. Just as primitive humans learned from observation that certain plants and grains were good to eat and could be cultivated, the healers and shamans observed the nature of certain illnesses and then passed on their experiences to other generations.

Evidence indicates that the primitive healer took an active, rather than simply intuitive, interest in the curative arts, acting as a surgeon and a user of tools. For instance, skulls with holes made in them by trephiners have been collected in various parts of Europe, Asia, and South America. These holes were cut out of the bone with flint instruments to gain access to the brain. Although one can only speculate the purpose of these early surgical operations, magic and religious beliefs seem to be the most likely reasons. Perhaps this procedure liberated from the skull the malicious demons that were thought to be the cause of extreme pain (as in the case of migraines) or attacks of falling to the ground (as in epilepsy). That this procedure was carried out on living patients, some of whom actually survived, is

evident from the rounded edges on the bone surrounding the hole, which indicate that the bone had grown again after the operation. These survivors also achieved a special status of sanctity so that, after their death, pieces of their skull were used as amulets to ward off convulsive attacks. From these beginnings, the practice of medicine has become integral to all human societies and cultures.

It is interesting to note the fate of some of the most successful of these early practitioners. The Egyptians, for example, have held Imhotep, the architect of the first pyramid (3000 BC), in great esteem through the centuries, not as a pyramid builder but as a doctor. Imhotep's name signified "he who cometh in peace" because he visited the sick to give them "peaceful sleep." This early physician practiced his art so well that he was deified in the Egyptian culture as the god of healing.

Egyptian mythology, like primitive religion, emphasized the interrelationships between the supernatural and one's health. For example, consider the mystic sign *Rx*, which still adorns all prescriptions today. It has a mythical origin: the legend of the Eye of Horus. It appears that as a child Horus lost his vision after being viciously attacked by Seth, the demon of evil. Then Isis, the mother of Horus, called for assistance to Thoth, the most important god of health, who promptly restored the eye and its powers. Because of this intervention, the Eye of Horus became the Egyptian symbol of godly protection and recovery, and its descendant, Rx, serves as the most visible link between ancient and modern medicine.

The concepts and practices of Imhotep and the medical cult he fostered were duly recorded on papyri and stored in ancient tombs. One scroll (dated c. 1500 BC), which George Elbers acquired in 1873, contains hundreds of remedies for numerous afflictions ranging from crocodile bites to constipation. A second famous papyrus (dated c. 1700 BC), discovered by Edwin Smith in 1862, is considered to be the most important and complete treatise on surgery of all antiquity. These writings outline proper diagnoses, prognoses, and treatment in a series of surgical cases. These two papyri are certainly among the outstanding writings in medical history.

As the influence of ancient Egypt spread, Imhotep was identified by the Greeks with their own god of healing: Aesculapius. According to legend, the god Apollo fathered Aesculapius during one of his many earthly visits. Apparently Apollo was a concerned parent, and, as is the case for many modern parents, he wanted his son to be a physician. He made Chiron, the centaur, tutor Aesculapius in the ways of healing (Figure 1.1). Chiron's student became so proficient as a healer that he soon surpassed his tutor and kept people so healthy that he began to decrease the population of Hades. Pluto, the god of the underworld, complained so violently about this course of events that Zeus killed Aesculapius with a thunderbolt and in the process promoted Aesculapius to Olympus as a god.

Inevitably, mythology has become entangled with historical facts, and it is not certain whether Aesculapius was in fact an earthly physician like Imhotep, the Egyptian. However, one thing is clear: by 1000 BC, medicine was already a highly respected profession. In Greece, the Aesculapia were temples of the healing cult and may be considered the first hospitals (Figure 1.1). In modern terms, these temples were essentially sanatoriums that had strong religious overtones. In them, patients were received and psychologically prepared, through prayer and sacrifice, to appreciate the past achievements of Aesculapius and his physician priests. After the appropriate rituals, they were allowed to enjoy "temple sleep." During

FIGURE 1.1 A sick child brought to the Temple of Aesculapius. *Courtesy of http://www.nouveaunet.com/images/art/84.jpg.*

the night, "healers" visited their patients, administering medical advice to clients who were awake or interpreting dreams of those who had slept. In this way, patients became convinced that they would be cured by following the prescribed regimen of diet, drugs, or bloodletting. On the other hand, if they remained ill, it would be attributed to their lack of faith. With this approach, patients, not treatments, were at fault if they did not get well. This early use of the power of suggestion was effective then and is still important in medical treatment today. The notion of "healthy mind, healthy body" is still in vogue today.

One of the most celebrated of these "healing" temples was on the island of Cos, the birthplace of Hippocrates, who as a youth became acquainted with the curative arts through his father, also a physician. Hippocrates was not so much an innovative physician as a collector of all the remedies and techniques that existed up to that time. Since he viewed the physician as a scientist instead of a priest, Hippocrates also injected an essential ingredient into medicine: its scientific spirit. For him, diagnostic observation and clinical treatment began to replace superstition. Instead of blaming disease on the gods, Hippocrates taught that disease was a natural process, one that developed in logical steps, and that symptoms were reactions of the body to disease. The body itself, he emphasized, possessed its own means of recovery, and the function of the physician was to aid these natural forces. Hippocrates treated each patient as an original case to be studied and documented. His shrewd

descriptions of diseases are models for physicians even today. Hippocrates and the school of Cos trained many individuals, who then migrated to the corners of the Mediterranean world to practice medicine and spread the philosophies of their preceptor. The work of Hippocrates and the school and tradition that stem from him constitute the first real break from magic and mysticism and the foundation of the rational art of medicine. However, as a practitioner, Hippocrates represented the spirit, not the science, of medicine, embodying the good physician: the friend of the patient and the humane expert.

As the Roman Empire reached its zenith and its influence expanded across half the world, it became heir to the great cultures it absorbed, including their medical advances. Although the Romans themselves did little to advance clinical medicine (the treatment of the individual patient), they did make outstanding contributions to public health. For example, they had a well-organized army medical service, which not only accompanied the legions on their various campaigns to provide "first aid" on the battlefield but also established "base hospitals" for convalescents at strategic points throughout the empire. The construction of sewer systems and aqueducts were truly remarkable Roman accomplishments that provided their empire with the medical and social advantages of sanitary living. Insistence on clean drinking water and unadulterated foods affected the control and prevention of epidemics and, however primitive, made urban existence possible. Unfortunately, without adequate scientific knowledge about diseases, all the preoccupation of the Romans with public health could not avert the periodic medical disasters, particularly the plague, that mercilessly befell its citizens.

Initially, the Roman masters looked upon Greek physicians and their art with disfavor. However, as the years passed, the favorable impression these disciples of Hippocrates made upon the people became widespread. As a reward for their service to the peoples of the Empire, Julius Caesar (46 BC) granted Roman citizenship to all Greek practitioners of medicine in his empire. Their new status became so secure that when Rome suffered from famine that same year, these Greek practitioners were the only foreigners not expelled from the city. On the contrary, they were even offered bonuses to stay!

Ironically, Galen, who is considered the greatest physician in the history of Rome, was himself a Greek. Honored by the emperor for curing his "imperial fever," Galen became the medical celebrity of Rome. He was arrogant and a braggart and, unlike Hippocrates, reported only successful cases. Nevertheless, he was a remarkable physician. For Galen, diagnosis became a fine art; in addition to taking care of his own patients, he responded to requests for medical advice from the far reaches of the empire. He was so industrious that he wrote more than 300 books of anatomical observations, which included selected case histories, the drugs he prescribed, and his boasts. His version of human anatomy, however, was misleading because he objected to human dissection and drew his human analogies solely from the studies of animals. However, because he so dominated the medical scene and was later endorsed by the Roman Catholic Church, Galen actually inhibited medical inquiry. His medical views and writings became both the "bible" and "the law" for the pontiffs and pundits of the ensuing Dark Ages.

With the collapse of the Roman Empire, the Church became the repository of knowledge, particularly of all scholarship that had drifted through the centuries into the Mediterranean. This body of information, including medical knowledge, was literally scattered through the monasteries and dispersed among the many orders of the Church.

The teachings of the early Roman Catholic Church and the belief in divine mercy made inquiry into the causes of death unnecessary and even undesirable. Members of the Church regarded curing patients by rational methods as sinful interference with the will of God. The employment of drugs signified a lack of faith by the doctor and patient, and scientific medicine fell into disrepute. Therefore, for almost a thousand years, medical research stagnated. It was not until the Renaissance in the 1500s that any significant progress in the science of medicine occurred. Hippocrates had once taught that illness was not a punishment sent by the gods but a phenomenon of nature. Now, under the Church and a new God, the older views of the supernatural origins of disease were renewed and promulgated. Since disease implied demonic possession, monks and priests would treat the sick through prayer, the laying on of hands, exorcism, penances, and exhibition of holy relics—practices officially sanctioned by the Church.

Although deficient in medical knowledge, the Dark Ages were not entirely lacking in charity toward the sick poor. Christian physicians often treated the rich and poor alike, and the Church assumed responsibility for the sick. Furthermore, the evolution of the modern hospital actually began with the advent of Christianity and is considered one of the major contributions of monastic medicine. With the rise in 335 AD of Constantine I, the first of the Roman emperors to embrace Christianity, all pagan temples of healing were closed, and hospitals were established in every cathedral city. (The word *hospital* comes from the Latin *hospes,* meaning "host" or "guest." The same root has provided *hotel* and *hostel*.) These first hospitals were simply houses where weary travelers and the sick could find food, lodging, and nursing care. The Church ran these hospitals, and the attending monks and nuns practiced the art of healing.

As the Christian ethic of faith, humanitarianism, and charity spread throughout Europe and then to the Middle East during the Crusades, so did its "hospital system." However, trained "physicians" still practiced their trade primarily in the homes of their patients, and only the weary travelers, the destitute, and those considered hopeless cases found their way to hospitals. Conditions in these early hospitals varied widely. Although a few were well financed and well managed and treated their patients humanely, most were essentially custodial institutions to keep troublesome and infectious people away from the general public. In these establishments, crowding, filth, and high mortality among both patients and attendants were commonplace. Thus, the hospital was viewed as an institution to be feared and shunned.

The Renaissance and Reformation in the fifteenth and sixteenth centuries loosened the Church's stronghold on both the hospital and the conduct of medical practice. During the Renaissance, "true learning," the desire to pursue the true secrets of nature including medical knowledge, was again stimulated. The study of human anatomy was advanced, and the seeds for further studies were planted by the artists Michelangelo, Raphael Durer, and, of course, the genius Leonardo da Vinci. They viewed the human body as it really was, not simply as a text passage from Galen. The painters of the Renaissance depicted people in sickness and pain, sketched in great detail and, in the process, demonstrated amazing insight into the workings of the heart, lungs, brain, and muscle structure. They also attempted to portray the individual and to discover emotional as well as physical qualities. In this stimulating era, physicians began to approach their patients and the pursuit of medical knowledge in similar fashion. New medical schools, similar to the most famous of such institutions at

Salerno, Bologna, Montpelier, Padua, and Oxford, emerged. These medical training centers once again embraced the Hippocratic doctrine that the patient was human, disease was a natural process, and commonsense therapies were appropriate in assisting the body to conquer its disease.

During the Renaissance, fundamentals received closer examination, and the age of measurement began. In 1592, when Galileo visited Padua, Italy, he lectured on mathematics to a large audience of medical students. His famous theories and inventions (the thermoscope and the pendulum, in addition to the telescopic lens) were expounded upon and demonstrated. Using these devices, one of his students, Sanctorius, made comparative studies of the human temperature and pulse. A future graduate of Padua, William Harvey, later applied Galileo's laws of motion and mechanics to the problem of blood circulation. This ability to measure the amount of blood moving through the arteries helped to determine the function of the heart.

Galileo encouraged the use of experimentation and exact measurement as scientific tools that could provide physicians with an effective check against reckless speculation. Quantification meant theories would be verified before being accepted. Individuals involved in medical research incorporated these new methods into their activities. Body temperature and pulse rate became measures that could be related to other symptoms to assist the physician in diagnosing specific illnesses or diseases. Concurrently, the development of the microscope amplified human vision, and an unknown world came into focus. Unfortunately, new scientific devices had little impact upon the average physician, who continued to blood-let and to disperse noxious ointments. Only in the universities did scientific groups band together to pool their instruments and their various talents.

In England, the medical profession found in Henry VIII a forceful and sympathetic patron. He assisted the doctors in their fight against malpractice and supported the establishment of the College of Physicians, the oldest purely medical institution in Europe. When he suppressed the monastery system in the early sixteenth century, church hospitals were taken over by the cities in which they were located. Consequently, a network of private, nonprofit, voluntary hospitals came into being. Doctors and medical students replaced the nursing sisters and monk physicians. Consequently, the professional nursing class became almost nonexistent in these public institutions. Only among the religious orders did "nursing" remain intact, further compounding the poor lot of patients confined within the walls of the public hospitals. These conditions were to continue until Florence Nightingale appeared on the scene years later.

Still another dramatic event was to occur. The demands made upon England's hospitals, especially the urban hospitals, became overwhelming as the population of these urban centers continued to expand. It was impossible for the facilities to accommodate the needs of so many. Therefore, during the seventeenth century two of the major urban hospitals in London—St. Bartholomew's and St. Thomas—initiated a policy of admitting and attending to only those patients who could possibly be cured. The incurables were left to meet their destiny in other institutions such as asylums, prisons, or almshouses.

Humanitarian and democratic movements occupied center stage primarily in France and the American colonies during the eighteenth century. The notion of equal rights finally began, and as urbanization spread, American society concerned itself with the welfare of many of its members. Medical men broadened the scope of their services to include the

"unfortunates" of society and helped to ease their suffering by advocating the power of reason and spearheading prison reform, child care, and the hospital movement. Ironically, as the hospital began to take up an active, curative role in medical care in the eighteenth century, the death rate among its patients did not decline but continued to be excessive. In 1788, for example, the death rate among the patients at the Hotel Dru in Paris, thought to be founded in the seventh century and the oldest hospital in existence today, was nearly 25 percent. These hospitals were lethal not only to patients but also to the attendants working in them, whose own death rate hovered between 6 and 12 percent per year.

Essentially the hospital remained a place to avoid. Under these circumstances, it is not surprising that the first American colonists postponed or delayed building hospitals. For example, the first hospital in America, the Pennsylvania Hospital, was not built until 1751, and the city of Boston took over two hundred years to erect its first hospital, the Massachusetts General, which opened its doors to the public in 1821.

A major advancement in the history of modern medicine came in the mid-nineteenth century with the development of the now well-known Germ Theory. Germ Theory simply states that infectious disease is caused by microorganisms living within the body. A popular example of early Germ Theory demonstration is that of John Snow and the Broad Street pump handle. When Cholera reached epidemic levels in the overcrowded Industrial Era streets of London, local physician John Snow was able to stop the spread of the disease with a street map. Snow plotted the cases of Cholera in the city, and he discovered an epicenter at a local water pump. By removing the handle, and thus access to the infected water supply, Snow illustrated Germ Theory and saved thousands of lives at the same time. French chemist Louis Pasteur is credited with developing the foundations of Germ Theory throughout the mid-nineteenth century.

Not until the nineteenth century could hospitals claim to benefit any significant number of patients. This era of progress was due primarily to the improved nursing practices fostered by Florence Nightingale (Figure 1.2) on her return to England from the Crimean War. She demonstrated that hospital deaths were caused more frequently by hospital conditions than by disease. During the latter part of the nineteenth century, she was at the height of her influence, and few new hospitals were built anywhere in the world without her advice. During the first half of the nineteenth century, Nightingale forced medical attention to focus once more on the care of the patient. Enthusiastically and philosophically, she expressed her views on nursing: "Nursing is putting us in the best possible condition for nature to restore and preserve health.... The art is that of nursing the sick. Please mark, not nursing sickness."

Although these efforts were significant, hospitals remained, until the twentieth century, institutions for the sick poor. In the 1870s, for example, when the plans for the projected Johns Hopkins Hospital were reviewed, it was considered quite appropriate to allocate 324 charity and 24 pay beds. Not only did the hospital population before the turn of the century represent a narrow portion of the socioeconomic spectrum, but it also represented only a limited number of the types of diseases prevalent in the overall population. In 1873, for example, roughly half of America's hospitals did not admit contagious diseases, and many others would not admit incurables. Furthermore, in this period, surgery admissions in general hospitals constituted only 5 percent, with trauma (injuries incurred by traumatic experience) making up a good portion of these cases.

FIGURE 1.2 A portrait of Florence Nightingale. *Courtesy of http://ginnger.topcities.com/cards/computer/nurses/765x525nightengale.gif.*

American hospitals a century ago were rather simple in that their organization required no special provisions for research or technology and demanded only cooking and washing facilities. In addition, since the attending and consulting physicians were normally unsalaried, and the nursing costs were quite modest, the great bulk of the hospital's normal operating expenses were for food, drugs, and utilities. Not until the twentieth century did "modern medicine" come of age in the United States. As we shall see, technology played a significant role in its evolution.

1.2 THE MODERN HEALTH CARE SYSTEM

Modern medical practice actually began at the turn of the twentieth century. Before 1900, medicine had little to offer the average citizen, since its resources were mainly physicians, their education, and their little black bags. At this time physicians were in short supply, but for different reasons than exist today. Costs were minimal, demand was small, and many of the services provided by the physician could also be obtained from experienced amateurs residing in the community. The individual's dwelling was the major site for treatment and recuperation, and relatives and neighbors constituted an able and willing nursing staff.

Midwives delivered babies, and those illnesses not cured by home remedies were left to run their fatal course. Only in the twentieth century did the tremendous explosion in scientific knowledge and technology lead to the development of the American health care system, with the hospital as its focal point and the specialist physician and nurse as its most visible operatives.

In the twentieth century, the advances made in the basic sciences (chemistry, physiology, pharmacology, and so on) began to occur much more rapidly. Discoveries in the physical sciences enabled medical researchers to take giant strides forward. For example, in 1903, William Einthoven devised the first electrocardiograph and measured the electrical changes that occurred during the beating of the heart (Figure 1.3). In the process, Einthoven initiated a new age for both cardiovascular medicine and electrical measurement techniques.

Of all the new discoveries that followed one another like intermediates in a chain reaction, the most significant for clinical medicine was the development of x-rays. When W. K. Roentgen described his "new kinds of rays," the human body was opened to medical inspection. Initially these x-rays were used in the diagnosis of bone fractures and dislocations. In the United States, x-ray machines brought this "modern technology" to most urban hospitals. In the process, separate departments of radiology were established, and the influence of their activities spread with almost every department of medicine (surgery, gynecology, and so forth) advancing with the aid of this new tool. By the 1930s, x-ray visualization

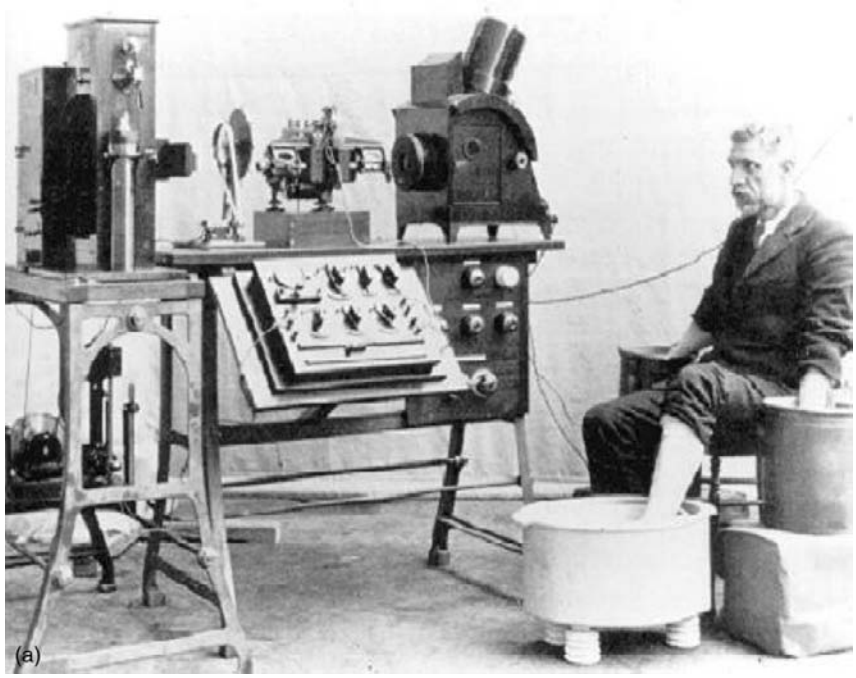


FIGURE 1.3 (a) An early electrocardiograph machine and

Continued

FIGURE 1.3, cont'd (b) a modern ECG setup. Computer technology and electronics advances have greatly simplified and strengthened the ECG as a diagnosis tool.

of practically all the organ systems of the body was possible by the use of barium salts and a wide variety of radiopaque materials.

The power this technological innovation gave physicians was enormous. The x-ray permitted them to diagnose a wide variety of diseases and injuries accurately. In addition, being within the hospital, it helped trigger the transformation of the hospital from a passive receptacle for the sick poor to an active curative institution for all the citizens of American society.

The introduction of sulfanilamide in the mid-1930s and penicillin in the early 1940s significantly reduced the main danger of hospitalization: cross-infection among patients. With these new drugs in their arsenals, surgeons were able to perform their operations without prohibitive morbidity and mortality due to infection. Also, despite major early-twentieth-century advancements in the field of hematology (including blood type differentiation and the use of sodium citrate to prevent clotting), blood banks were not fully developed until the 1930s, when technology provided adequate refrigeration. Until that time, "fresh" donors were bled, and the blood was transfused while it was still warm.

As technology in the United States blossomed, so did the prestige of American medicine. From 1900 to 1929, Nobel Prize winners in physiology or medicine came primarily from Europe, with no American among them. In the period 1930 to 1944, just before the end of World War II, 19 Americans were honored as Nobel Prize Laureates. During the postwar period (1945–1975), 102 American life scientists earned similar honors, and from 1975 to 2009, the number was 191. Thus, since 1930 a total of 312 American scientists, including some born abroad, have performed research that was significant enough to warrant the

distinction of a Nobel Prize. Most of these efforts were made possible by the technology that was available to these clinical scientists.

The employment of the available technology assisted in advancing the development of complex surgical procedures. The Drinker respirator was introduced in 1927, and the first heart-lung bypass was performed in 1939. In the 1940s, cardiac catheterization and angiography (the use of a cannula threaded through an arm vein and into the heart with the injection of radiopaque dye for the x-ray visualization of lung and heart vessels and valves) were developed. Accurate diagnoses of congenital and acquired heart disease (mainly valve disorders due to rheumatic fever) also became possible, and a new era of cardiac and vascular surgery began. The development and implementation of robotic surgery in the first decade of the twenty-first century have even further advanced the capabilities of modern surgeons. Neurosurgery, both peripheral and central, and vascular surgery have seen significant improvements and capabilities with this new technology (Figure 1.4).

Another child of this modern technology, the electron microscope, entered the medical scene in the 1950s and provided significant advances in visualizing relatively small cells. Body scanners using early PET (positron-emission tomography) technology to detect tumors arose from the same science that brought societies reluctantly into the atomic age. These "tumor detectives" used radioactive material and became commonplace in newly established departments of nuclear medicine in all hospitals.

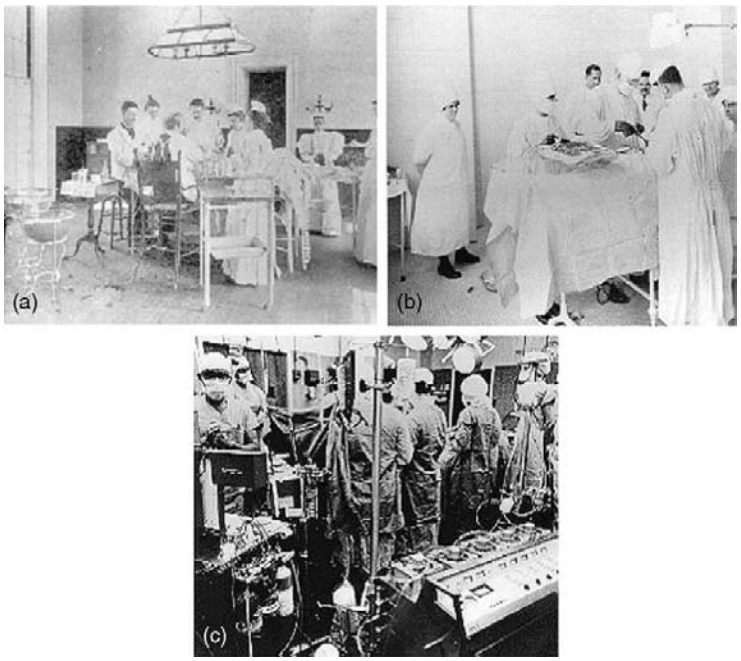


FIGURE 1.4 Changes in the operating room: (a) the surgical scene at the turn of the century, (b) the surgical scene in the late 1920s and early 1930s, and (c) the surgical scene today *From J. D. Bronzino,* Technology for Patient Care, *St. Louis: Mosby, 1977;* The Biomedical Engineering Handbook, *CRC Press, 1995; 2000; 2005.*

The impact of these discoveries and many others was profound. The health care system that consisted primarily of the "horse and buggy" physician was gone forever, replaced by the doctor backed by and centered around the hospital, as medicine began to change to accommodate the new technology.

Following World War II, the evolution of comprehensive care greatly accelerated. The advanced technology that had been developed in the pursuit of military objectives now became available for peaceful applications, with the medical profession benefiting greatly from this rapid surge of technological "finds." For instance, the realm of electronics came into prominence. The techniques for following enemy ships and planes, as well as providing aviators with information concerning altitude, air speed, and the like, were now used extensively in medicine to follow the subtle electrical behavior of the fundamental unit of the central nervous system—the neuron—or to monitor the beating heart of a patient.

The Second World War also brought a spark of innovation in the rehabilitation engineering and prosthetics fields. With advances in medical care technologies, more veterans were returning home alive—and disabled. This increase in need, combined with a surge in new materials development in the late 1940s, assisted the growth of assistive technologies during the post-WWII era.

Science and technology have leapfrogged past each other throughout recorded history. Anyone seeking a causal relation between the two was just as likely to find technology the cause and science the effect, with the converse also holding true. As gunnery led to ballistics and the steam engine transformed into thermodynamics, so did powered flight lead to aerodynamics. However, with the advent of electronics this causal relation has been reversed; scientific research is systematically exploited in the pursuit of technical advancement.

Just as World War II sparked an advancement in comprehensive care, the 1960s enjoyed a dramatic electronics revolution, compliments of the first lunar landing. What was considered science fiction in the 1930s and 1940s became reality. Devices continually changed to incorporate the latest innovations, which in many cases became outmoded in a very short period of time. Telemetry devices used to monitor the activity of a patient's heart freed both the physician and the patient from the wires that previously restricted them to the four walls of the hospital room. Computers, similar to those that controlled the flight plans of the *Apollo* capsules, now completely inundate our society.

Since the 1970s, medical researchers have put these electronic brains to work performing complex calculations, keeping records (via artificial intelligence), and even controlling the very instrumentation that sustains life. The development of new medical imaging techniques such as computerized tomography (CT) and magnetic resonance imaging (MRI) totally depended on a continually advancing computer technology. New imaging developments include functional MRI (Figure 1.5), a tool capable of illustrating active neural areas by quantifying oxygen consumption and blood flow in the brain. The citations and technological discoveries are so myriad that it is impossible to mention them all.

"Spare parts" surgery is now routine. With the first successful transplantation of a kidney in 1954, the concept of "artificial organs" gained acceptance and officially came into vogue in the medical arena (Figure 1.6). Technology to provide prosthetic devices, such as artificial heart valves and artificial blood vessels, developed. Even an artificial heart program to develop a replacement for a defective or diseased human heart began.

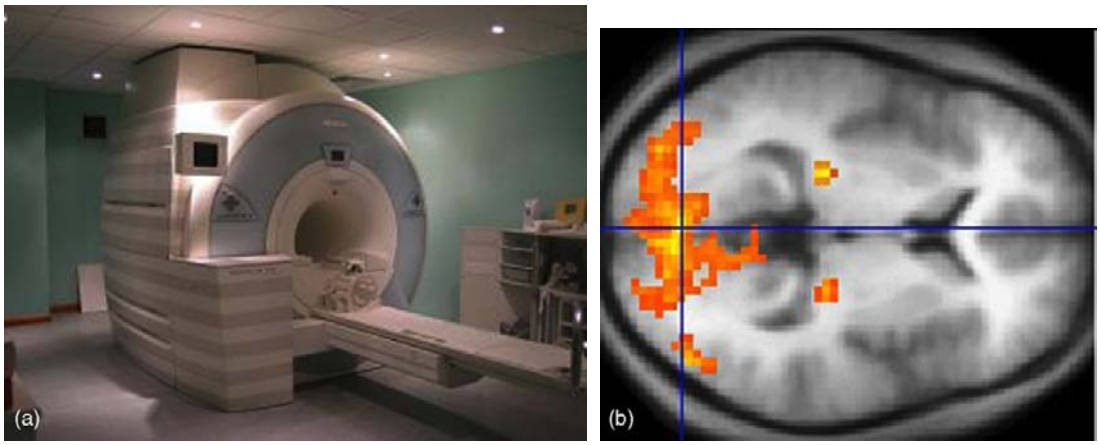


FIGURE 1.5 (a) A modern fMRI medical imaging facility and (b) an fMRI scan image. *http://neurophilosophy.wordpress.com.*

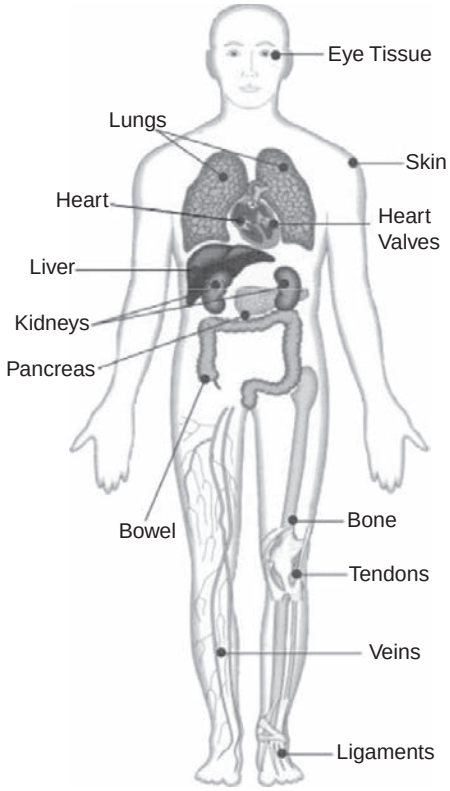


FIGURE 1.6 Transplantations performed today. *http://www.transplant.bc.ca/images/what_organs.gif.*

With the neural function, resilience, and incredible mechanical strength and endurance of the human heart, complete replacement prosthetics have been only marginally successful. Left ventricular assist devices (LVAD), however, have seen success as a replacement for the "workhorse" region of the heart and are a popular temporary option for those waiting on a full heart transplant. Future directions for heart failure solutions will most likely involve more tissue and cellular level treatments, as opposed to macromechanical systems. These technological innovations have vastly altered surgical organization and utilization, even further enhancing the radical evolution hospitals have undergone from the low-tech institutions of just 100 years ago to the modern advanced medical centers of tomorrow.

In recent years, technology has struck medicine like a thunderbolt. The Human Genome Project was perhaps the most prominent scientific and technological effort of the 1990s. Some of the engineering products vital to the effort included automatic sequencers, robotic liquid handling devices, and software for databasing and sequence assembly (See Figure 1.7). As a result, a major transition occurred, moving biomedical engineering to focus on the cellular and molecular level rather than solely on the organ system level. With the success of the "genome project," completed in 2003 after a 13-year venture, new vistas have been opened. Stem cell research highlights this chemical and molecular level focus and has been on the

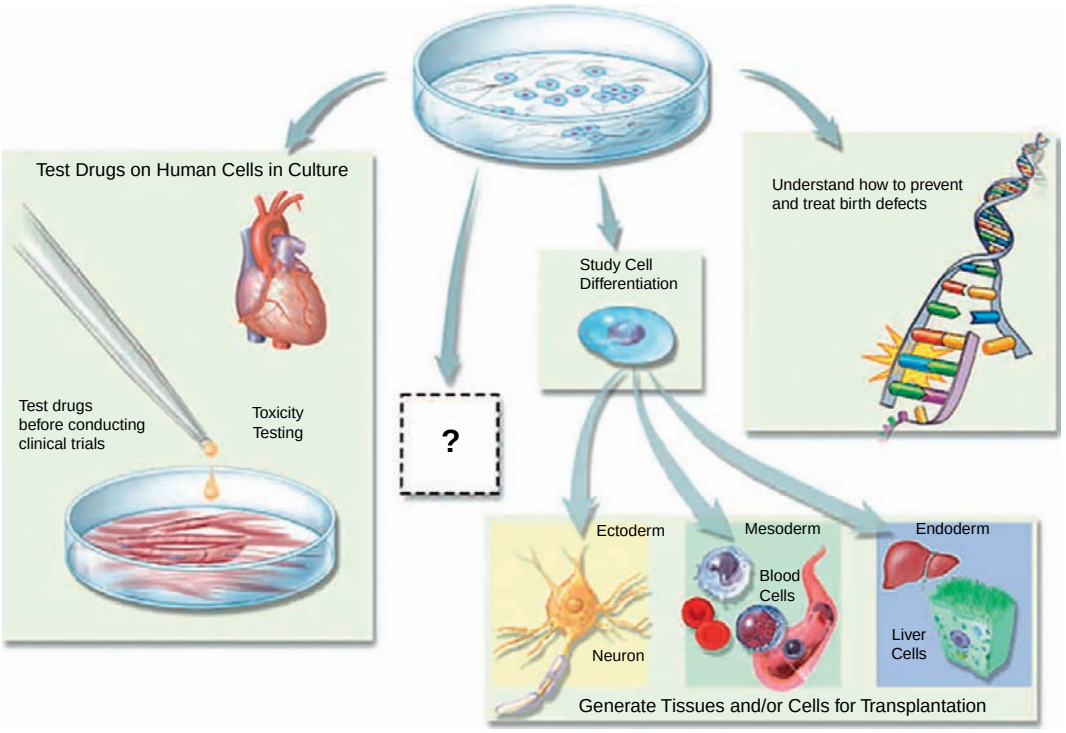


FIGURE 1.7 Stem cell research—potential applications made possible. *http://stemcells.nih.gov/info/media/promise.htm.*

FIGURE 1.8 Robotic surgery—a new tool in the arsenal of the physician. *http://library.thinkquest.org/03oct/00760/steve.jpg.*

forefront of controversial scientific research since its conception. While the multitudes of possibilities defy imagination, the moral issues accompanying stem cells have received equal attention in recent years.

Furthermore, advances in nanotechnology, tissue engineering, and artificial organs are clear indications that science fiction will continue to become reality. However, the social and economic consequences of this vast outpouring of information and innovation must be fully understood if this technology is to be exploited effectively and efficiently.

As one gazes into the crystal ball, technology offers great potential for affecting health care practices (Figure 1.8). It can provide health care for individuals in remote rural areas by means of closed-circuit television health clinics with complete communication links to a regional health center. Development of multiphasic screening systems can provide preventative medicine to the vast majority of our population and restrict hospital admissions to those requiring the diagnostic and treatment facilities housed there. With the creation of a central medical records system, anyone moving or becoming ill away from home can have records made available to the attending physician easily and rapidly. These are just a few of the possibilities that illustrate the potential of technology in creating the type of medical care system that will indeed be accessible, high quality, and reasonably priced for all. (For an extensive review of major events in the evolution of biomedical engineering, see Nebeker, 2002.)

1.3 WHAT IS BIOMEDICAL ENGINEERING?

Many of the problems confronting health professionals today are of extreme importance to the engineer because they involve the fundamental aspects of device and systems analysis, design, and practical application—all of which lie at the heart of processes that are fundamental to engineering practice. These medically relevant design problems can range

from very complex large-scale constructs, such as hospital information systems, to the creation of relatively small and "simple" devices, such as recording electrodes and transducers used to monitor the activity of specific physiological processes.

The American health care system, therefore, encompasses many problems that represent challenges to certain members of the engineering profession, called biomedical engineers. Since biomedical engineering involves applying the concepts, knowledge, and approaches of virtually all engineering disciplines (e.g., electrical, mechanical, and chemical engineering) to solve specific health care–related problems, the opportunities for interaction between engineers and health care professionals are many and varied.

Although what is included in the field of biomedical engineering is considered by many to be quite clear, many conflicting opinions concerning the field can be traced to disagreements about its definition. For example, consider the terms *biomedical engineering*, *bioengineering*, *biological engineering*, and *clinical* (or *medical*) *engineer*, which are defined in the *Bioengineering Education Directory*. While Pacela defined *bioengineering* as the broad umbrella term used to describe this entire field, bioengineering is usually defined as a basic-research-oriented activity closely related to biotechnology and genetic engineering—that is, the modification of animal or plant cells or parts of cells to improve plants or animals or to develop new microorganisms for beneficial ends. In the food industry, for example, this has meant the improvement of strains of yeast for fermentation. In agriculture, bioengineers may be concerned with the improvement of crop yields by treatment plants with organisms to reduce frost damage. It is clear that bioengineers for the future will have tremendous impact on the quality of human life. The full potential of this specialty is difficult to image. Typical pursuits include the following:

- The development of improved species of plants and animals for food production
- The invention of new medical diagnostic tests for diseases
- The production of synthetic vaccines from clone cells
- Bioenvironmental engineering to protect human, animal, and plant life from toxicants and pollutants
- The study of protein-surface interactions
- Modeling of the growth kinetics of yeast and hybridoma cells
- Research in immobilized enzyme technology
- The development of therapeutic proteins and monoclonal antibodies

The term *biomedical engineering* appears to have the most comprehensive meaning. Biomedical engineers apply electrical, chemical, optical, mechanical, and other engineering principles to understand, modify, or control biological (i.e., human and animal) systems. When a biomedical engineer works within a hospital or clinic, he or she is more properly called a *clinical engineer*. However, this theoretical distinction is not always observed in practice, since many professionals working within U.S. hospitals today continue to be called biomedical engineers.

The breadth of activity of biomedical engineers is significant. The field has moved significantly from being concerned primarily with the development of medical devices in the 1950s and 1960s to include a more wide-ranging set of activities. As shown in Figure 1.9, the field of biomedical engineering now includes many new career areas:

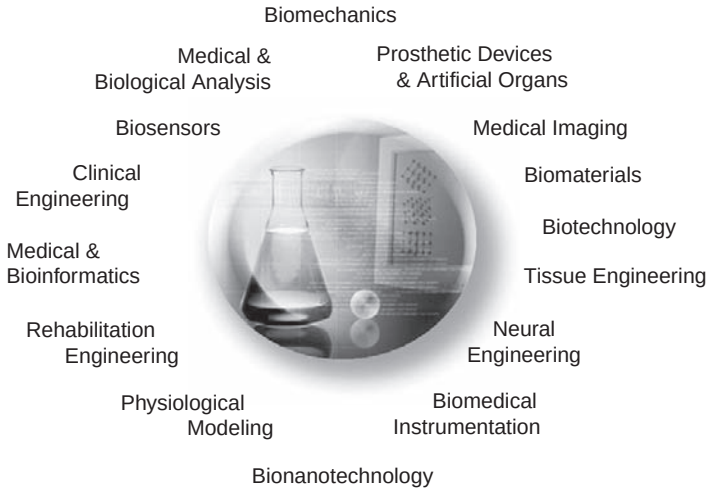


FIGURE 1.9 The world of biomedical engineering.

- Application of engineering system analysis (physiologic modeling, simulation, and control to biological problems)
- Detection, measurement, and monitoring of physiologic signals (i.e., biosensors and biomedical instrumentation)
- Diagnostic interpretation via signal-processing techniques of bioelectric data
- Therapeutic and rehabilitation procedures and devices (*rehabilitation engineering*)
- Devices for replacement or augmentation of bodily functions (*artificial organs*)
- Computer analysis of patient-related data and clinical decision making (i.e., medical informatics and artificial intelligence)
- Medical imaging—that is, the graphical display of anatomic detail or physiologic function
- The creation of new biologic products (i.e., biotechnology and tissue engineering)

Typical pursuits of biomedical engineers include the following:

- Research in new materials for implanted artificial organs
- Development of new diagnostic instruments for blood analysis
- Writing software for analysis of medical research data
- Analysis of medical device hazards for safety and efficacy
- Development of new diagnostic imaging systems
- Design of telemetry systems for patient monitoring
- Design of biomedical sensors
- Development of expert systems for diagnosis and treatment of diseases
- Design of closed-loop control systems for drug administration
- Modeling of the physiologic systems of the human body
- Design of instrumentation for sports medicine
- Development of new dental materials
- Design of communication aids for individuals with disabilities

- Study of pulmonary fluid dynamics
- Study of biomechanics of the human body
- Development of material to be used as replacement for human skin

The preceding list is not intended to be all-inclusive. Many other applications use the talents and skills of the biomedical engineer. In fact, the list of activities of biomedical engineers depends on the medical environment in which they work. This is especially true for the clinical engineers—biomedical engineers employed in hospitals or clinical settings. Clinical engineers are essentially responsible for all the high-technology instruments and systems used in hospitals today, the training of medical personnel in equipment safety, and the design, selection, and use of technology to deliver safe and effective health care.

Engineers were first encouraged to enter the clinical scene during the late 1960s in response to concerns about electrical safety of hospital patients. This safety scare reached its peak when consumer activists, most notably Ralph Nader, claimed, "At the very least, 1,200 Americans are electrocuted annually during routine diagnostic and therapeutic procedures in hospitals." This concern was based primarily on the supposition that catheterized patients with a low-resistance conducting pathway from outside the body into blood vessels near the heart could be electrocuted by voltage differences well below the normal level of sensation. Despite the lack of statistical evidence to substantiate these claims, this outcry served to raise the level of consciousness of health care professionals with respect to the safe use of medical devices.

In response to this concern, a new industry—hospital electrical safety—arose almost overnight. Organizations such as the National Fire Protection Association (NFPA) wrote standards addressing electrical safety specifically for hospitals. Electrical safety analyzer manufacturers and equipment safety consultants became eager to serve the needs of various hospitals that wanted to provide a "safety fix" and of some companies, particularly those specializing in power distribution systems (most notably isolation transformers). To alleviate these fears, the Joint Commission on the Accreditation of Healthcare Organizations (then known as the Joint Commission on Accreditation of Hospitals) turned to NFPA codes as the standard for electrical safety and further specified that hospitals must inspect all equipment used on or near a patient for electrical safety at least every six months. To meet this new requirement, hospital administrators considered a number of options, including (1) paying medical device manufacturers to perform these electrical safety inspections, (2) contracting for the services of shared-services organizations, or (3) providing these services with in-house staff. When faced with this decision, most large hospitals opted for in-house service and created whole departments to provide the technological support necessary to address these electrical safety concerns.

As a result, a new engineering discipline—clinical engineering—was born. Many hospitals established centralized clinical engineering departments. Once these departments were in place, however, it soon became obvious that electrical safety failures represented only a small part of the overall problem posed by the presence of medical equipment in the clinical environment. At the time, this equipment was neither totally understood nor properly maintained. Simple visual inspections often revealed broken knobs, frayed wires, and even evidence of liquid spills. Many devices did not perform in accordance with manufacturers'

specifications and were not maintained in accordance with manufacturers' recommendations. In short, electrical safety problems were only the tip of the iceberg. By the mid-1970s, complete performance inspections before and after equipment use became the norm, and sensible inspection procedures were developed. In the process, these clinical engineering pioneers began to play a more substantial role within the hospital. As new members of the hospital team, they did the following:

- Became actively involved in developing cost-effective approaches for using medical technology
- Provided hospital administrators with advice regarding the purchase of medical equipment based on their ability to meet specific technical specifications
- Started using modern scientific methods and working with standards-writing organizations
- Became involved in the training of health care personnel regarding the safe and efficient use of medical equipment

Then, during the 1970s and 1980s, a major expansion of clinical engineering occurred, primarily due to the following events:

- The Veterans Administration (VA), convinced that clinical engineers were vital to the overall operation of the VA hospital system, divided the country into biomedical engineering districts, with a chief biomedical engineer overseeing all engineering activities in the hospitals in that district.
- Throughout the United States, clinical engineering departments were established in most large medical centers and hospitals and in some smaller clinical facilities with at least three hundred beds.
- Health care professionals—physicians and nurses—needed assistance in utilizing existing technology and incorporating new innovations.
- Certification of clinical engineers became a reality to ensure the continued competence of practicing clinical engineers.

During the 1990s, the evaluation of clinical engineering as a profession continued with the establishment of the American College of Clinical Engineering (ACCE) and the Clinical Engineering Division within the International Federation of Medical and Biological Engineering (IFMBE). Clinical engineers today provide extensive engineering services for the clinical staff and serve as a significant resource for the entire hospital (Figure 1.10). Possessing in-depth knowledge regarding available in-house technological capabilities as well as the technical resources available from outside firms, the modern clinical engineer enables the hospital to make effective and efficient use of most if not all of its technological resources.

Biomedical engineering is thus an interdisciplinary branch of engineering heavily based in both engineering and the life sciences. It ranges from theoretical, nonexperimental undertakings to state-of-the-art applications. It can encompass research, development, implementation, and operation. Accordingly, like medical practice itself, it is unlikely that any single person can acquire expertise that encompasses the entire field. As a result, there has been an explosion of biomedical engineering specialties to cover this broad field. Yet, because of the interdisciplinary nature of this activity, there are considerable interplay and overlapping of interest and effort between them. For example, biomedical engineers engaged in the

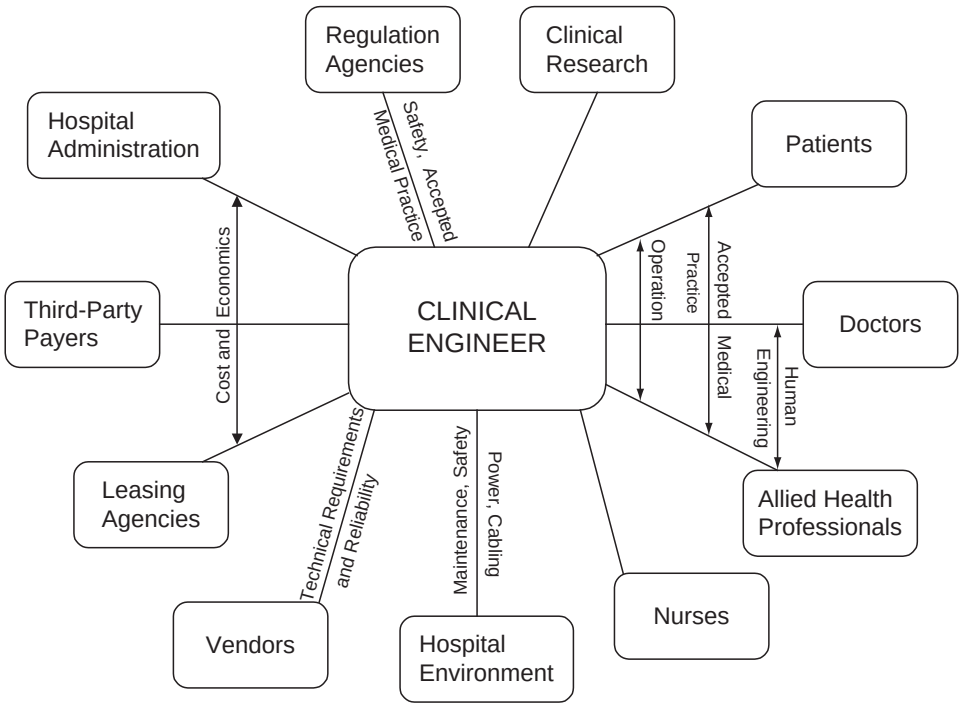


FIGURE 1.10 The range of interactions that a clinical engineer may be required to engage in a hospital setting.

development of biosensors may interact with those interested in prosthetic devices to develop a means to detect and use the same bioelectric signal to power a prosthetic device. Those engaged in automating the clinical chemistry laboratory may collaborate with those developing expert systems to assist clinicians in making clinical decisions based upon specific laboratory data. The possibilities are endless.

Perhaps an even greater benefit of the utilization of biomedical engineers lies in the potential for implementing existing technologies to identify and solve problems within our present health care system. Consequently, the field of biomedical engineering offers hope in the continuing battle to provide high-quality health care at a reasonable cost. If properly directed toward solving problems related to preventative medical approaches, ambulatory care services, and the like, biomedical engineers can provide the tools and techniques to make our health care system more effective and efficient.

1.4 ROLES PLAYED BY THE BIOMEDICAL ENGINEERS

In its broadest sense, biomedical engineering involves training essentially three types of individuals: the clinical engineer in health care, the biomedical design engineer for industry, and the research scientist. Presently, one might also distinguish among three specific roles these biomedical engineers can play. Each is different enough to merit a separate

description. The first type, the most common, might be called the "problem solver." This biomedical engineer (most likely the clinical engineer or biomedical design engineer) maintains the traditional service relationship with the life scientists who originate a problem that can be solved by applying the specific expertise of the engineer. For this problem-solving process to be efficient and successful, however, some knowledge of each other's language and a ready interchange of information must exist. Biomedical engineers must understand the biological situation to apply their judgment and contribute their knowledge toward the solution of the given problem, as well as to defend their methods in terms that the life scientist can understand. If they are unable to do these things, they do not merit the "biomedical" appellation.

The second type, which is less common, could be called the "technological entrepreneur" (most likely a biomedical design engineer in industry). This individual assumes that the gap between the technological education of the life scientist or physician and the present technological capability has become so great that the life scientist cannot pose a problem that will incorporate the application of existing technology. Therefore, technological entrepreneurs examine some portion of the biological or medical front and identify areas in which advanced technology might be advantageous. Thus, they pose their own problem and then proceed to provide the solution, at first conceptually and then in the form of hardware or software. Finally, these individuals must convince the medical community that they can provide a useful tool because, contrary to the situation in which problem solvers find themselves, the entrepreneur's activity is speculative at best and has no ready-made customer for the results. If the venture is successful, however, whether scientifically or commercially, then an advance has been made much earlier than it would have been through the conventional arrangement. Because of the nature of their work, technological entrepreneurs should have a great deal of engineering and medical knowledge as well as experience in numerous medical systems.

The third type of biomedical engineer—the "engineer-scientist" (most likely found in academic institutions and industrial research labs)—is primarily interested in applying engineering concepts and techniques to the investigation and exploration of biological processes. The most powerful tool at their disposal is the construction of an appropriate physical or mathematical model of the specific biological system under study. An example of this relationship can be found in the study of cardiac function. The engineer-scientist may be exploring the complexities of fluid flow through the incredible pump that is the human heart. Mathematical models may be created to model the kinematics of the heart during contraction and equations to define the behavior of fluid flow. Through simulation techniques and available computing machinery, they can use this model to understand features that are too complex for either analytical computation or intuitive recognition. In addition, this process of simulation facilitates the design of appropriate experiments that can be performed on the actual biological system. The results of these experiments can, in turn, be used to amend the model. Thus, increased understanding of a biological mechanism results from this iterative process.

This mathematical model can also predict the effect of these changes on a biological system in cases where the actual experiments may be tedious, very difficult, or dangerous. The researchers are thus rewarded with a better understanding of the biological system,

and the mathematical description forms a compact, precise language that is easily communicated to others. In the example of the cardiac researcher, the engineer must at all times consider the anatomical and physiological causes for the macro-model results—in this case, why the heart is pumping the way it is. The activities of the engineer-scientist inevitably involve instrument development because the exploitation of sophisticated measurement techniques is often necessary to perform the biological side of the experimental work. It is essential that engineer-scientists work in a biological environment, particularly when their work may ultimately have a clinical application. It is not enough to emphasize the niceties of mathematical analysis while losing the clinical relevance in the process. This biomedical engineer is a true partner of the biological scientist and has become an integral part of the research teams being formed in many institutes to develop techniques and experiments that will unfold the mysteries of the human organism. Each of these roles envisioned for the biomedical engineer requires a different attitude, as well as a specific degree of knowledge about the biological environment. However, each engineer must be a skilled professional with a significant expertise in engineering technology.

Therefore, in preparing new professionals to enter this field at these various levels, biomedical engineering educational programs are continually being challenged to develop curricula that will provide an adequate exposure to and knowledge about the environment, without sacrificing essential engineering skills. As we continue to move into a period characterized by a rapidly growing aging population, rising social and economic expectations, and a need for the development of more adequate techniques for the prevention, diagnosis, and treatment of disease, development and employment of biomedical engineers have become a necessity. This is true not only because they may provide an opportunity to increase our knowledge of living systems but also because they constitute promising vehicles for expediting the conversion of knowledge to effective action.

The ultimate role of the biomedical engineer, like that of the nurse and physician, is to serve society. This is a profession, not just a skilled technical service. To use this new breed effectively, health care practitioners and administrators should be aware of the needs for these new professionals and the roles for which they are being trained. The great potential, challenge, and promise in this endeavor offer not only significant technological benefits but humanitarian benefits as well.

1.5 RECENT ADVANCES IN BIOMEDICAL ENGINEERING

Biomedical engineering is a vast field with a multitude of concentrations and research initiatives. While the technicians affiliated with clinical engineering and a number of other concentrations focus mainly on preexisting technologies, researchers enjoy the exhilaration of innovating the new. Biomedical engineering has grown exponentially since its acceptance as a field less than a century ago, to the extent that today there is not a branch of medicine untouched by the problem-solving skill set of the engineer. The objective of this section is not to make the reader aware of every cutting-edge technology in development today but rather to provide an introduction to a sample of these new adventures.

1.5.1 Prosthetics

Prosthetics are one of the oldest innovations of biomedical engineering. The assistive technology field, prosthetics especially, became a true engineering discipline in itself in the period following World War II, when an unprecedented number of veterans returned home alive, but disabled, due to advances in medicine.

Prosthetics are defined as any "internal or external device(s) that *replace* lost parts or functions of the neuroskeletomotor system" and may be either orthopedic or externally controlled. Externally controlled devices may be powered by the body itself through myoelectricity or a separate power supply. Neural prosthetics represent the newest field in prosthetics and one of the fastest-developing topics in biomedical engineering today.

Orthopedic Prosthetics

In designing a "replacement" limb for the human body, an engineer is buried under an obscene amount of considerations and design constraints. The appendage must be functionally sufficient, a design unique to each individual, depending on the activities to be accomplished. It must be comfortable, aesthetically pleasing, convenient, and simple in attachment. Prosthetics and orthoses seeking to imitate the human body piece by piece tend to have a great amount of difficulty in development and implementation. Instead, the general application of the device should always be considered, with the user in mind. An example of this design strategy can be found in the flex foot, a prosthetic foot with no real resemblance to the natural appendage. Instead of struggling to recreate the biomechanics of the ankle, tarsals, metatarsals, and phalanges of the lower leg, designers created a prosthetic with a single contact piece, no joint, and consisting of only one material. The Cheetah Leg shown in Figure 1.11, is one type of such a prosthetic and has

FIGURE 1.11 Paralympic sprinter Oscar Pistorius with a prosthetic leg. Designing for overall function, as opposed to mirroring the human body, is often the more practical approach. *Compliments of http://www.thefinalsprint.com/images/2008/05/oscar-pistorius-double-amputee-sprinter.jpg.*

allowed paralympians like Oscar Pistorius to compete at a scale approaching that of able-bodied athletes. Actually, the Cheetah Leg allowed Pistorius, a double amputee, to compete at a level that became subject to controversy. In 2008, the South African sprinter battled courts for the opportunity to race with able-bodied athletes in the Beijing Olympic Games. While Pistorius ultimately did not qualify, his efforts fueled a debate as to whether his engineered prosthetics functioned better than a human leg, actually giving him an advantage over runners in the standard Olympic Games.

Externally controlled prosthetics use external motors to power their operation. The C-Leg is an example of such a device. This prosthetic leg has a microprocessor-controlled knee; has force sensors throughout for angle, swing, and velocity; and lasts 25 to 30 hours without charging. Uneven terrain is tackled with the C-Leg, as are changes in walking pace and direction. In recent years, sensor and minimally sized motor developments have made devices such as the C-Leg possible.

Neural Prosthetics

Neural prosthetics present one of the newest and perhaps most exciting concentrations of biomedical engineering. These devices may be powered by the human body—that is, they operate from electrical signals sent via electrodes from an external source to the peripheral muscle neuron—or they may be powered externally. These systems that use functional electrical stimulation (FES) to "restore sensory or motor function"are the definition of neural prostheses. These NPs have the potential to assist victims of spinal cord or cervical column injury (SCI and CI), restoring function to the muscle and lower extremities.

Stimulation via electrodes must reach a threshold frequency to achieve tetanization, or the smooth motion contraction of muscle. Stimulation below this frequency results in isolated twitches and muscle fatigue. Electrodes may be implanted transcutaneously (on the surface), percutaneously (stimulator outside the body connects to a stimulation point inside), or implanted.

As opposed to the leg, where a series of fairly simple joints and large motor units provide sufficient function, the upper extremities prove a significant challenge in fine-tuned control requirements. The incredible strength and flexibility of complex hand function are difficult to reproduce. The newest in prosthetic design hopes to overcome some of these challenges. The Luke Arm (Figure 1.12) is the brainchild of Segway inventor Dean Kamen. The arm has just as many degrees of freedom as the human arm and is capable of lifting above the user's head. The arm uses myoelectric signals originating from residual nerves in the upper body. Fine-tuned control is assisted by controls in the user's shoe; by activating different "pedals," the user can rotate the wrist or grasp or release an object. Sensory feedback, a constant issue with mechanical prosthetics, is provided via a pressure sensor on the fingertips, which feed back to a vibrating patch worn on the user's back. Increased pressure is felt by the user by changes in vibration intensity. Clinical trials are underway.

The design of prosthetics involves an intensive materials engineering background, as well as an in-depth understanding of kinematics modeling and physiology. The American Board for Certification in Orthotics, Prosthetics, and Pedorthics provides guidelines for certification as a licensed prosthetist. Those in the field are required to complete an accredited

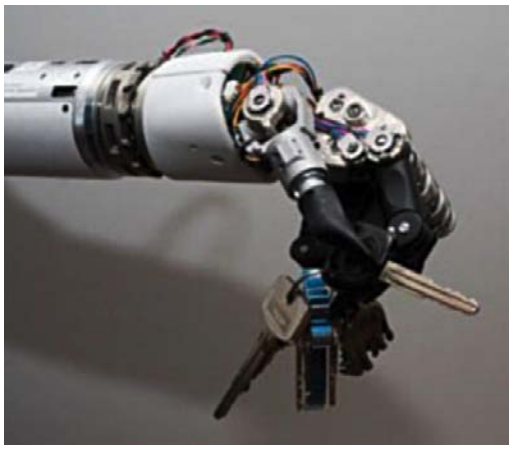


FIGURE 1.12 Dean Kamen's Luke Arm, the most advanced neural prosthetic to date, which uses myoelectric signals. Clinical trials are presently underway. *Courtesy of http://medgadget.com.*

undergraduate program in prosthetics or a graduate program specializing in the field with an appropriate undergraduate degree. Neural prosthetic development involves a team of members from various backgrounds, including biomechanics, electronics, and mathematical modeling.

While prosthetics provide a strong example of the evolution of assistive technology and present a number of interesting design innovations, in recent years the field of biomedical engineering has shifted from a focus on mechanical systems to biological and organic solutions. Whereas decades ago the primary objectives of the biomedical engineer consisted of device design, modern feats are more likely to involve biochemistry and gene therapy than screws, nuts, and bolts. Two prime examples of this shift in focus are tissue engineering and stem cell research.

1.5.2 Tissue Engineering

Tissue engineering, a relatively new field in biomedical engineering, consists of the manufacture of biological tissue either ex vivo or in vitro (outside the body), or the incorporation of new advancements to aid in the repair and growth of existing tissues in vivo (inside the body). In ex vivo applications, bioartificial tissues (those composed of both synthetic and natural materials) are used as an alternative to organ transplant or developed to study tissue behavior in vitro. Some important issues within the field include cell isolation, control of cell organization and function, upscaling to full bioartificial tissues, and biomaterial fabrication.

While the most well-known tissue engineering feats have been in epithelial tissues, clinical trials are also currently under way for reconstruction of cartilage, bone, neural, and liver tissues. Grafts are used for treatment of every type of skin damage, including burns, pressure sores, venous stasis ulcers, and diabetic ulcers. Polymeric tubes are implanted to assist in nerve regeneration due to central and peripheral nervous system damage or disorders.

Tissue engineering also covers joint replacements, including connective tissue recreation and bone grafts. Artificial heart valves implement bovine and porcine tissues along with bioartificial substances. Organ failure is treated with innovations in the field as well, with treatment for everything from liver cancer to breast reconstruction. Blood transfusions and dental surgery advancements are just two more examples of the wide range of applications of tissue engineering technologies.

Bone marrow transplantation works to regenerate the most prolific organ of the body. Marrow is responsible for the production of blood cells and is often damaged by myeloablative treatment regimens, such as chemotherapy and radiation. Modern methods involve harvesting patient samples of marrow prior to the therapy regimen and reinjecting them following treatment. The body regenerates its marrow supply, causing a temporary immunodeficiency.

In the case of pancreatic and liver tissue development, a bioreactor model is used. Bioreactors are systems consisting of a large number of cells that take in an input of reactants and output a set of products. Bioreactors have also been implemented for blood cell production from hematopoietic tissue. The two types of bioreactor systems are hollow fiber and microcarrier-based systems. In the hollow fiber system, a large number of small-diameter, hollow tubes are bundled together by a larger shell tube. The small tubes are injected with organ-specific cells that are suspended in a collagen-based matrix. The matrix will contract, leaving space within the small tubes. The patient's own blood or plasma is injected into the larger, encompassing tube and is allowed to nourish the hepatopoietic cells by flowing through the newly emptied space in the smaller tubes. In microcarrier-based systems, small beads (less than 500 μm) with surfaces specially treated for cell attachment are either positioned in a packed or fluidized bed or incorporated in hollow fiber cartridges. In the packed bed method, a column is filled with the beads and capped at each end with porous plates to allow perfusion. Success rates rely on fluid flow rate through the column, as well as the density of packed beads and dimension ratios of the column.

Biomaterials play a significant role in tissue engineering. In each of the previous examples, biomaterials prove an integral component of tissue regeneration and reconstruction. From the obvious application of artificial valve design to the less apparent role of injection needle design in bone marrow transplantation, biomaterial development is a necessary step in the advancement of tissue engineering. Devices must provide mechanical support, prevent undesirable tissue interactions, and potentially allow for timely biodegradation. Biomaterial devices can be broken down into two types, each existing on a scale as small as a few hundred microns. Immunoprotective devices contain semipermeable membranes that prevent specific host immune system elements from entering the device. Open devices, in contrast, are designed for systems to be fully integrated with the host and have large pores (greater than 10 μm), allowing for free transport of cells and molecules.

Pore sizes within a biomaterial directly correlate to the functions of the device. The structure of a pore is determined by the continuity of individual pores in the device, as well as the size and size distribution. The three classifications of porous materials are microporous, mesoporous, and macroporous. Microporous materials have pores with a diameter less than 2 nm and allow for transport of small molecules, including gases. Mesoporous materials allow for transport of small proteins and have pores with diameters ranging from 2 to 50 nm. Macroporous materials have pores with diameters greater than 50 nm and allow

for large proteins, and possibly even cells, to pass through. Fibrovascular tissue will pass through any material with pores greater than 10 nm. Pore size and distribution are rarely regular, and abnormalities will both change general material properties as well as fluid flow rate across the device.

A major focus in tissue engineering is controlling cell organization and regeneration. The more control the researcher has over cell development, the greater the capabilities and the wider the range of applications of the bioartificial tissue. Stem cells provide an opportunity for researchers to develop tissues essentially from scratch. Stem cells both build and maintain cells in vivo and possess the ability to be used for tissue generation ex vivo. Some background information on this new technology is provided following, and specific applications to tissue engineering are available in later chapters.

1.5.3 Stem Cell Research

In recent years, stem cells have become the topic of both intense controversy and incredible excitement within the research community. The potential for stem cell technology is apparently limitless, with some known possibilities shown in Figure 1.7. Cells may be used to test drugs on different types of tissues, to understand how to prevent birth defects, and to potentially replace and regenerate damaged tissue in the body. The possibilities truly seem endless.

In actuality there are two different types of stem cells. *Embryonic stem cells* come from embryos, which are mostly supplied by in vitro fertilization clinics four to five days following fertilization. At this point, stem cells will either self-regenerate or commit and differentiate. *Self-renewal* or *regeneration* means that the stem cell will reproduce with no developmental commitment. Essentially, the stem cell remains a stem cell. *Differentiation* is the expression of tissue or cell-specific genes.

For the majority of tissues in the human body, cells will differentiate terminally. In some cases, however, dynamic operation is required, and, as such, a population of *adult stem cells* is maintained for regeneration purposes. The two most common types of adult stem cells are those of the hematopoietic system (blood renewal) and the intestinal epithelia. These cell types are similar in that they both occur in very large numbers and have short life spans. Stem cells are required to maintain this dynamic population.

Researchers control stem cell development and differentiation within cultures by a number of means. For embryonic stem cells, the difference between self-regeneration and differentitation, surprisingly enough, is the concentration of a single essential protein, or growth factor. Leukemia inhibitory factor (LIF), in high enough concentrations, will cause embryonic stem cells to regenerate indefinitely in cultures. This is an interesting fact, because it proves that stem cell development is not an intrinsic predetermined state but rather is induced by extrinsic factors.

With an executive order in 2009, President Barack Obama lifted an eight-and-a-half-year ban on government-funded stem cell research, earning praise from the science community for opening the door to potential cures for some of mankind's most debilitating diseases. Both tissue engineering and stem cell research represent just a sampling of the breakthrough biologically focused ventures currently being explored by today's biomedical engineers.

1.6 PROFESSIONAL STATUS OF BIOMEDICAL ENGINEERING

Biomedical engineers are *professionals,* which are defined as an aggregate of people finding identity in sharing values and skills absorbed during a common course of intensive training. Whether individuals are professionals is determined by examining whether they have internalized certain given professional values. Furthermore, a professional is someone who has internalized professional values and is licensed on the basis of his or her technical competence. Professionals generally accept scientific standards in their work, restrict their work activities to areas in which they are technically competent, avoid emotional involvement, cultivate objectivity in their work, and put their clients' interests before their own.

The concept of a profession that is involved in the design, development, and management of medical technology encompasses three primary occupational models: science, business, and profession. Consider initially the contrast between science and profession. *Science* is seen as the pursuit of knowledge, its value hinging on providing evidence and communicating with colleagues. *Profession*, on the other hand, is viewed as providing a service to clients who have problems they cannot handle themselves. Scientists and professionals have in common the exercise of some knowledge, skill, or expertise. However, while scientists practice their skills and report their results to knowledgeable colleagues, professionals, such as lawyers, physicians, and engineers, serve lay clients. To protect both the professional and the client from the consequences of the layperson's lack of knowledge, the practice of the profession is often regulated through such formal institutions as state licensing. Both professionals and scientists must persuade their clients to accept their findings. Professionals endorse and follow a specific code of ethics to serve society. On the other hand, scientists move their colleagues to accept their findings through persuasion.

Consider, for example, the medical profession. Its members are trained in caring for the sick, with the primary goal of healing them. These professionals not only have a responsibility for the creation, development, and implementation of that tradition, but they are also expected to provide a service to the public, within limits, without regard to self-interest. To ensure proper service, the profession closely monitors the licensing and certification process. Thus, medical professionals themselves may be regarded as a mechanism of social control. However, this does not mean that other facets of society are not involved in exercising oversight and control of physicians in their practice of medicine.

A final attribute of professionals is that of integrity. Physicians tend to be both permissive and supportive in relationships with patients and yet are often confronted with moral dilemmas involving the desires of their patients and social interest. For example, how to honor the wishes of terminally ill patients while not facilitating the patients' deaths is a moral question that health professionals are forced to confront. A detailed discussion of the moral issues posed by medical technology is presented in Chapter 2.

One can determine the status of professionalization by noting the occurrence of six crucial events: the first training school, the first university school, the first local professional association, the first national professional association, the first state license law, and the first formal code of ethics. The early appearances of the training school and the university affiliation underscore the importance of the cultivation of a knowledge base. The strategic innovative role of the universities and early teachers lies in linking knowledge to practice and creating

a rationale for exclusive jurisdiction. Those practitioners pushing for prescribed training then form a professional association. The association defines the tasks of the profession: raising the quality of recruits; redefining their function to permit the use of less technically skilled people to perform the more routine, less involved tasks; and managing internal and external conflicts. In the process, internal conflict may arise between those committed to previously established procedures and newcomers committed to change and innovation. At this stage, some form of professional regulation, such as licensing or certification, surfaces because of a belief that it will ensure minimum standards for the profession, enhance status, and protect the layperson in the process.

The last area of professional development is the establishment of a formal code of ethics, which usually includes rules to exclude the unqualified and unscrupulous practitioners, rules to reduce internal competition, and rules to protect clients and emphasize the ideal service to society. A code of ethics usually comes at the end of the professionalization process.

In biomedical engineering, all six critical steps have been clearly taken. The field of biomedical engineering, which originated as a professional group interested primarily in medical electronics in the late 1950s, has grown from a few scattered individuals to a very well-established organization. There are approximately 48 international societies throughout the world serving an increasingly expanding community of biomedical engineers. Today, the scope of biomedical engineering is enormously diverse. Over the years, many new disciplines such as tissue engineering, artificial intelligence, and so on, which were once considered alien to the field, are now an integral part of the profession.

Professional societies play a major role in bringing together members of this diverse community to share their knowledge and experience in pursuit of new technological applications that will improve the health and quality of life. Intersocietal cooperation and collaborations, at both the national and international levels, are more actively fostered today through professional organizations such as the Biomedical Engineering Society (BMES), the American Institute for Medical and Biological Engineering (AIMBE), Engineering in Medicine and Biology Society (EMBS), and the Institute of Electrical and Electronic Engineers (IEEE).

1.7 PROFESSIONAL SOCIETIES

1.7.1 The American Institute for Medical and Biological Engineering

The United States has the largest biomedical engineering community in the world. Major professional organizations that address various cross sections of the field and serve biomedical engineering professionals include the American College of Clinical Engineering, the American Institute of Chemical Engineers, the American Medical Informatics Association, the American Society of Agricultural Engineers, the American Society for Artificial Internal Organs, the American Society of Mechanical Engineers, the Association for the Advancement of Medical Instrumentation, the Biomedical Engineering Society, the IEEE Engineering in Medicine and Biology Society, an interdisciplinary Association for the Advancement of

Rehabilitation and Assistive Technologies, and the Society for Biomaterials. In an effort to unify all the disparate components of the biomedical engineering community in the United States as represented by these various societies, the American Institute for Medical and Biological Engineering (AIMBE) was created in 1992. The primary goal of AIMBE is to serve as an umbrella organization in the United States for the purpose of unifying the bioengineering community, addressing public policy issues, and promoting the engineering approach in society's effort to enhance health and the quality of life through the judicious use of technology. (For information, contact AIMBE, 1701 K Street, Suite 510, Washington, DC, 20036; http://www.aimbe.org/; e-mail: aimbeoffice@gmail.com.)

1.7.2 IEEE Engineering in Medicine and Biology Society

The Institute of Electrical and Electronic Engineers (IEEE) is the largest international professional organization in the world and accommodates 37 different societies and councils under its umbrella structure. Of these 37, the Engineering in Medicine and Biology Society (EMBS) represents the foremost international organization, serving the needs of over 8,000 biomedical engineering members around the world. The major interest of the EMBS encompasses the application of concepts and methods from the physical and engineering sciences to biology and medicine. Each year, the society sponsors a major international conference while cosponsoring a number of theme-oriented regional conferences throughout the world. Premier publications consist of a monthly journal (*Transactions on Biomedical Engineering*), three quarterly journals (*Transactions on Neural Systems and Rehabilitation Engineering*, *Transactions on Information Technology in Biomedicine*, and *Transactions on Nanobioscience*), as well as a bimonthly magazine (*IEEE Engineering in Medicine and Biology Magazine*). Secondary publications, authored in collaboration with other societies, include *Transactions on Medical Imaging*, *Transactions on Neural Networks*, and *Transactions on Pattern Analysis and Machine Intelligence*. (For more information, contact the IEEE EMBS Executive Office, IEEE, 445 Hoes Lane, Piscataway, NJ, 08855-1331; http://www.embs.org/; e-mail: emb-exec@ieee.org.)

1.7.3 The Biomedical Engineering Society

Established in 1968, the Biomedical Engineering Society (BMES) was founded in order to address a need for a society that afforded equal status to representatives of both biomedical and engineering interests. With that in mind, the primary goal of the BMES, as stated in their Articles of Incorporation, is "to promote the increase of biomedical engineering knowledge and its utilization." Regular meetings are scheduled biannually in both the spring and fall. Additionally, special interest meetings are interspersed throughout the year and are promoted in conjunction with other biomedical engineering societies such as AIMBE and EMBS. The primary publications associated with the BMES include *Annals of Biomedical Engineering*, a monthly journal presenting original research in several biomedical fields; *BMES Bulletin*, a quarterly newsletter presenting a wider array of subject matter relating both to biomedical engineering as well as BMES news and events; and the *BMES Membership Directory*, an annual publication listing the contact information of the society's

individual constituents. (For more information, contact the BMES directly: BMES, 8401 Corporate Drive, Suite 140, Landover, MD 20785-2224; http://www.bmes.org/default.asp; e-mail: info@bmes.org.)

The activities of these biomedical engineering societies are critical to the continued advancement of the professional status of biomedical engineers. Therefore, all biomedical engineers, including students in the profession, are encouraged to become members of these societies and engage in the activities of true professionals.

1.8 EXERCISES

1. Select a specific "medical technology" from the historical periods indicated. Describe the fundamental principles of operation and discuss their impact on health care delivery: (a) 1900–1939; (b) 1945–1970; (c) 1970–1980; (d) 1980–2003.
2. Provide a review of the effect that computer technology has had on health care delivery, citing the computer application and the time frame of its implementation.
3. The term *genetic engineering* implies an engineering function. Is there one? Should this activity be included in the field of biomedical engineering?
4. Discuss in some detail the role the genome project has had and is anticipated having on the development of new medical technology.
5. Using your crystal ball, what advances in engineering and/or life science do you think will have the greatest impact on clinical care or biomedical research?
6. The organizational structure of a hospital involves three major groups: the Board of Trustees, the administrators, and the medical staff. Specify the major responsibilities of each. In what group should a Department of Clinical Engineering reside? Explain your answer.
7. Based on its definition, what attributes should a *clinical engineer* have?
8. List at least seven (7) specific activities of clinical engineers.
9. Provide modern examples (i.e., names of individuals and their activities) of the three major roles played by biomedical engineers: (a) the problem solver; (b) the technological entrepreneur; and (c) the engineer scientist.
10. Do the following groups fit the definition of a *profession*? Discuss how they do or do not. (a) registered nurse; (b) biomedical technician; (c) respiratory therapist; (d) hospital administrator.
11. List the areas of knowledge necessary to practice biomedical engineering. Identify where in the normal educational process one can acquire knowledge. How best can administrative skills be acquired?
12. Prosthetic limbs are often created for specialized activities, such as mountain biking or driving. Create a design for an upper- or lower-extremity prosthetic for a particular specialty activity.
13. What steps must be taken to become a licensed prosthetician?
14. What are the two means of powering a neural prosthetic?
15. What is the difference between an adult stem cell and an embryonic stem cell? Where does each come from?

16. While stem cell research has recently been granted federal funding, limitations on the types of research allowed are decided by the state. Research your home state's policies on stem cell research, and provide a summary paragraph.
17. Provide a copy of the home page for a biomedical engineering professional society and a list of its major activities for the coming year.
18. Research a biomedical engineering professional society and provide three benefits of joining the society.
19. What is your view regarding the role biomedical engineers will play in the health care system of tomorrow?
20. Discuss the trade-offs in health care that occur as a result of limited financial resources.
21. Discuss whether medical technology is an economic cost factor, a benefit, or both.

Suggested Readings

C. Aston, Biological Warfare Canaries, IEEE Spectrum (2001) 35–40.
I.N. Bankman, Handbook of Medical Imaging, CRC Press, Boca Raton, FL, 2000.
J.D. Bronzino, Biomedical Engineering Handbook, first and second ed., CRC Press, Boca Raton, FL, 1995; 2000; 2005.
J.D. Bronzino, Management of Medical Technology: A Primer for Clinical Engineering, Butterworth, Boston, 1992.
E. Carson, C. Cobelli, Modeling Methodology for Physiology and Medicine, Academic Press, San Diego, CA, 2001.
D.J. DiLorenzo, J.D. Bronzino, Neuroengineering, CRC Press, Boca Raton, FL, 2008.
C.T. Laurenchin, Repair and Restore with Tissue Engineering, EMBS Magazine 22 (5) (2003) 16–17.
F. Nebekar, Golden Accomplishments in Biomedical Engineering, EMBS Magazine 21 (3) (2002) 17–48.
A. Pacela, Bioengineering Education Directory, Quest Publishing Co., Brea, CA, 1990.
B.O. Palsson, S.N. Bhatia, Tissue Engineering, Prentice Hall, Upper Saddle River, NJ, 2004.
J.B. Park, J.D. Bronzino, Biomaterials: Principles and Applications, CRC Press, Boca Raton, FL, 2003.
D. Serlin, Replaceable You, University of Chicago Press, Chicago, 2004.
The IEEE/EMBS magazine published by the Institute of Electrical and Electronic Engineers, in: J. Enderle (Ed.), especially "Writing the Book on BME," vol. 21, No. 3, 2002.
M.L. Yarmush, M. Toner, Biotechnology for Biomedical Engineers, CRC Press, Boca Raton, FL, 2003.

This page intentionally left blank

CHAPTER

2

Moral and Ethical Issues

Joseph D. Bronzino, PhD, PE

OUTLINE

AFTER COMPLETING THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Define and distinguish between the terms *morals* and *ethics.*
- Present the rationale underlying two major philosophical schools of thought: *utilitarianism* and *nonconsequentialism.*
- Present the codes of ethics for the medical profession, nursing profession, and biomedical engineering.
- Identify the modern moral dilemmas, including redefining death, deciding how

to care for the terminally ill, and human experimentation, which arise from the two moral norms: *beneficence* (the provision of benefits) and *nonmaleficence* (the avoidance of harm).

- Discuss the moral judgments associated with present policies regarding the regulation of the development and use of new medical devices.

The tremendous infusion of technology into the practice of medicine has created a new medical era. Advances in material science have led to the production of artificial limbs, heart valves, and blood vessels, thereby permitting "spare parts" surgery. Numerous patient disorders are now routinely diagnosed using a wide range of highly sophisticated imaging devices, and the lives of many patients are being extended through significant improvements in resuscitative and supportive devices, such as respirators, pacemakers, and artificial kidneys.

These technological advances, however, have not been benign. They have had significant moral consequences. Provided with the ability to develop cardiovascular assist devices, perform organ transplants, and maintain the breathing and heartbeat of terminally ill patients, society has been forced to reexamine the meaning of such terms as *death, quality of life, heroic efforts,* and *acts of mercy*, and to consider such moral issues as the right of patients to refuse treatment (living wills) and to participate in experiments (informed consent). As a result, these technological advances have made the moral dimensions of health care more complex and have posed new and troubling moral dilemmas for medical professionals, the biomedical engineer, and society at large.

The purpose of this chapter is to examine some of the moral questions related to the use of new medical technologies. The objective, however, is not to provide solutions or recommendations for these questions. Rather, the intent is to demonstrate that each technological advance has consequences that affect the very core of human values.

Technology and ethics are not foreigners; they are neighbors in the world of human accomplishment. Technology is a human achievement of extraordinary ingenuity and utility and is quite distant from the human accomplishment of ethical values. They face each other rather than interface. The personal face of ethics looks at the impersonal face of technology in order to comprehend technology's potential and its limits. The face of technology looks to ethics to be directed to human purposes and benefits.

In the process of making technology and ethics face each other, it is our hope that individuals engaged in the development of new medical devices, as well as those responsible for the care of patients, will be stimulated to examine and evaluate critically "accepted" views and to reach their own conclusions. This chapter, therefore, begins with some definitions related to morality and ethics, followed by a more detailed discussion of some of the moral issues of special importance to biomedical engineers.

2.1 MORALITY AND ETHICS: A DEFINITION OF TERMS

From the very beginning, individuals have raised concerns about the nature of life and its significance. Many of these concerns have been incorporated into the four fundamental questions posed by the German philosopher Immanuel Kant (1724–1804): *What can I know?*

What ought I to do? What can I hope? What is man? Evidence that early societies raised these questions can be found in the generation of rather complex codes of conduct embedded in the customs of the earliest human social organization: the tribe. By 600 BC, the Greeks were successful in reducing many primitive speculations, attitudes, and views on these questions to some type of order or system and integrating them into the general body of wisdom called *philosophy*. Being seafarers and colonizers, the Greeks had close contact with many different peoples and cultures. In the process, struck by the variety of customs, laws, and institutions that prevailed in the societies that surrounded them, they began to examine and compare all human conduct in these societies. This part of philosophy they called *ethics.*

The term *ethics* comes from the Greek *ethos,* meaning "custom." On the other hand, the Latin word for custom is *mos,* and its plural, *mores,* is the equivalent of the Greek *ethos* and the root of the words *moral* and *morality.* Although both terms (*ethics* and *morality*) are often used interchangeably, there is a distinction between them that should be made.

Philosophers define ethics as a particular kind of study and use morality to refer to its subject matter. For example, customs that result from some abiding principal human interaction are called *morals.* Some examples of morals in our present society are telling the truth, paying one's debts, honoring one's parents, and respecting the rights and property of others. Most members of society usually consider such conduct not only customary but also correct or right. Thus, morality encompasses what people believe to be right and good and the reasons they give for it.

Most of us follow these rules of conduct and adjust our lifestyles in accordance with the principles they represent. Many even sacrifice life itself rather than diverge from them, applying them not only to their own conduct but also to the behavior of others. Individuals who disregard these accepted codes of conduct are considered deviants and, in many cases, are punished for engaging in an activity that society as a whole considers unacceptable. For example, individuals committing "criminal acts" (defined by society) are often "outlawed" and, in many cases, severely punished. These judgments regarding codes of conduct, however, are not inflexible; they must continually be modified to fit changing conditions and thereby avoid the trauma of revolution as the vehicle for change.

While morality represents the codes of conduct of a society, ethics is the study of right and wrong, of good and evil in human conduct. Ethics is not concerned with providing any judgments or specific rules for human behavior, but rather with providing an objective analysis about what individuals "ought to do." Defined in this way, it represents the philosophical view of morals, and, therefore, is often referred to as *moral philosophy.*

Consider the following three questions: "Should badly deformed infants be kept alive?"; "Should treatment be stopped to allow a terminally ill patient to die?"; "Should humans be used in experiments?" Are these questions of morality or ethics? In terms of the definitions just provided, all three of these inquiries are questions of moral judgment.

Philosophers argue that all moral judgments are considered to be "normative judgments"—that is, they can be recognized simply by their characteristic evaluative terms such as good, bad, right, wrong, and so on. Typical normative judgments include the following:

- Stealing is wrong.
- Everyone ought to have access to an education.
- Voluntary euthanasia should not be legalized.

Each of these judgments expresses an evaluation; that is, it conveys a negative or positive attitude toward some state of affairs. Each, therefore, is intended to play an action-guiding function.

Arriving at moral judgments, however, requires knowledge of valid moral standards in our society. Nevertheless, how is such knowledge obtained? The efforts to answer this question lie in two competing schools of thought that currently dominate normative ethical theory: *utilitarianism,* a form of consequentialism, and *Kantianism,* a form of nonconsequentialism. Consequentialism holds that the morally right action is always the one among the available options that has the best consequences. An important implication of consequentialism is that no specific actions or courses of conduct are automatically ruled out as immoral or ruled in as morally obligatory. The rightness or wrongness of an action is wholly contingent upon its effects.

According to utilitarianism, there are two steps to determining what ought to be done in any situation. First, determine which courses of action are open. Second, determine the consequences of each alternative. When this has been accomplished, the morally right course of action is the one that maximizes pleasure, minimizes pain, or both—the one that does the "greatest good for the greatest number." Because the central motivation driving the design, development, and use of medical devices is improvement of medicine's capacity to protect and restore health, an obvious virtue of utilitarianism is that it assesses medical technology in terms of what many believe makes health valuable: the attainment of well-being and the avoidance of pain.

Utilitarianism, therefore, advocates that the end justifies the means. As long as any form of treatment maximizes good consequences, it should be used. Many people, though, believe that the end does not always justify the means and that individuals have rights that are not to be violated no matter how good the consequences might be.

In opposition to utilitarianism stands the school of normative ethical thought known as *nonconsequentialism*. Proponents of this school deny that moral evaluation is simply and wholly a matter of determining the consequences of human conduct. They agree that other considerations are relevant to moral assessment and so reject the view that morally right conduct is whatever has the best consequences. Based largely on the views of Immanuel Kant, this ethical school of thought insists that there is something uniquely precious about human beings from the moral point of view. According to Kant's theory, humans have certain "rights" that do not apply to any other animal. For example, the moral judgments that we should not kill and eat one another for food or hunt one another for sport or experiment on one another for medical science are all based on this view of human rights. Humans are, in short, owed a special kind of respect simply because they are people.

These two philosophies may be extended to apply to animal testing in scientific research as well. On the utilitarianism side of the argument for animal experimentation, the health care advancements for humans made possible through animal research far outweigh the majority of arguments against the practice. In contrast, nonconsequentialism would state that maltreatment of innocent and unprotected living beings is morally unjust and as such is an inappropriate means to the ends of better health care for people. Ultimately researchers must decide for themselves, based on their own beliefs and reasoning, which philosophy wins out.

In terms of human experimentation, to better understand the Kantian perspective, it may be helpful to recognize that Kant's views are an attempt to capture in secular form a basic

tenet of Christian morality. What makes human beings morally special entities deserving a unique type of respect? Christianity answers in terms of the doctrine of ensoulment. This doctrine holds that only human beings are divinely endowed with an eternal soul. According to Christian ethics, the soul makes humans the only beings with intrinsic value. Kant's secular version of the doctrine of ensoulment asserts that human beings are morally unique and deserve special respect because of their autonomy. Autonomy is taken by Kant to be the capacity to make choices based on rational deliberation. The central task of ethics then is to specify what human conduct is required to respect the unique dignity of human beings. For most Kantians, this means determining what limits human beings must observe in the way they treat one another, and this, in turn, is taken to be a matter of specifying each individual's fundamental moral rights.

These two ethical schools of thought, therefore, provide some rationale for moral judgments. However, when there is no clear moral judgment, one is faced with a dilemma. In medicine, moral dilemmas arise in those situations that raise fundamental questions about right and wrong in the treatment of sickness and the promotion of health in patients. In many of these situations, the health professional usually faces two alternative choices, neither of which seems to be a satisfactory solution to the problem. For example, is it more important to preserve life or to prevent pain? Is it right to withhold treatment when doing so may lead to a shortening of life? Does an individual have the right to refuse treatment when refusing it may lead to death? All these situations seem to have no clear-cut imperative based on our present set of convictions about right and wrong. That is the dilemma raised by Kant: *What ought I do?*

CASE STUDY: STEM CELL RESEARCH

At the moment of conception—that is, when a sperm penetrates an egg—the process of fertilization occurs. The formation of an embryo is initiated. Once the sperm enters the egg, there is an immediate opening of ion channels, which depolarizes the plasma membrane of the cell and prevents other sperm from fusing with it. DNA replication then begins, and the first cell division occurs approximately 36 hours later. As the process continues, the cell begins to experience cleavage, where the cells repeatedly divide, cycling between the S (DNA synthesis) and M (mitosis) phases of cell division, essentially skipping the G_1 and G_2 phases, when most cell growth normally occurs. Thus, there is no net growth of the cells, merely subdivision into smaller cells, individually called blastomeres.

Five days after fertilization, the number of cells composing the embryo is in the hundreds, and the cells form tight junctions characteristic of a compact epithelium, which is arranged around a central cavity. This is the embryonic stage known as the blastocyst. Within the cavity exists a mass of cells, which protrude inward. These cells are known as the inner cell mass and become the embryo. The exterior cells are the trophoblast and eventually form the placenta. It is the cells from the inner cell mass of the blastocyst, however, that, when isolated and grown in a culture, are identified as embryonic stem cells.

It is important to note that if cell division continues, determination and differentiation happen. Differentiation occurs when a cell begins to exhibit the specific attributes of a predestined specialized cellular role. Determination is related to differentiation but is somewhat dissimilar.

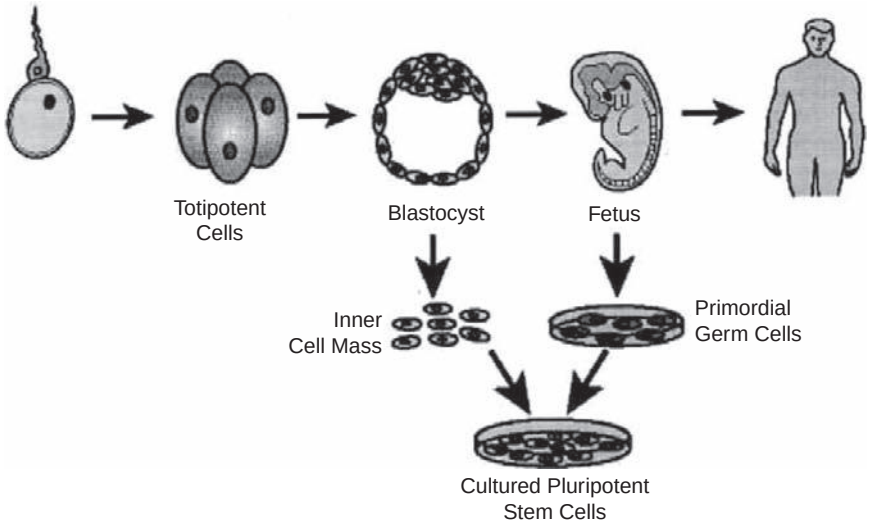


FIGURE 2.1 Using the inner cell mass to form pluripotent stem cells. *Courtesy of http://www.nih.gov/news/stemcell/primer.htm.*

When a cell group that has been determined is transplanted, it will not assimilate with the other cells but will grow into cells that comprised the original organ it was destined to become.

Since the process of obtaining embryonic stem cells (Figure 2.1) destroys the embryo, the following questions arise:

1. Is the embryo a living human being, entitled to all of the same rights that a human at any other age would be granted? Discuss the answer to this question from a Utilitarian and Kantian point of view.
2. Should any research that is potentially beneficial to the well-being of mankind be pursued? In 2009, President Obama passed groundbreaking legislation entitled "Executive Order 13505—Removing Barriers to Responsible Scientific Research Involving Human Stem Cells." The order calls for a review of NIH (National Institute of Health) guidelines for stem cell research and, more importantly, removes the requirement of President Action to approve NIH-funded stem cell investigations.
3. Should the federal government support (i.e., use tax dollars to fund) such research? Or, in contrast, should the government be allowed to interfere?

In the practice of medicine, moral dilemmas are certainly not new. They have been present throughout medical history. As a result, over the years there have been efforts to provide a set of guidelines for those responsible for patient care. These efforts have resulted in the development of specific codes of professional conduct. Let us examine some of these codes or guidelines.

For the medical profession, the World Medical Association adopted a version of the Hippocratic Oath entitled the *Geneva Convention Code of Medical Ethics* in 1949. This declaration contains the following statements:

I solemnly pledge myself to consecrate my life to the services of humanity;
I will give to my teachers the respect and gratitude which is their due;
I will practice my profession with conscience and dignity;
The health of my patient will be my first consideration;
I will respect the secrets which are confided in me;
I will maintain by all the means in my power, the honour and the noble traditions of the medical profession;
My colleagues will be my brothers;
I will not permit considerations of religion, nationality, race, party politics, or social standing to intervene between my duty and my patient;
I will maintain the utmost respect for human life from the time of conception, even under threat;
I will not use my medical knowledge contrary to the laws of humanity;
I make these promises solemnly, freely, and upon my honour.

In the United States, the American Medical Association (AMA) adopted a set of Principles of Medical Ethics in 1980 and revised them in June 2001. A comparison of the two sets of principles is provided following.

Revised Principles
Version adopted by the AMA House of Delegates, June 17, 2001

The medical profession has long subscribed to a body of ethical statements developed primarily for the benefit of the patient. As a member of this profession, a physician must recognize responsibility to patients first and foremost, as well as to society, to other health professionals, and to self. The following Principles adopted by the American Medical Association are not laws but standards of conduct that define the essentials of honorable behavior for the physician.

I. A physician shall be dedicated to providing competent medical care, with compassion and respect for human dignity and rights.
II. A physician shall uphold the standards of professionalism, be honest in all professional interactions, and strive to report physicians deficient in character or

Previous Principles
As adopted by the AMA's House of Delegates, 1980

The medical profession has long subscribed to a body of ethical statements developed primarily for the benefit of the patient. As a member of this profession, a physician must recognize responsibility not only to patients, but also to society, to other health professionals, and to self. The following Principles adopted by the American Medical Association are not laws but standards of conduct that define the essentials of honorable behavior for the physician.

I. A physician shall be dedicated to providing competent medical service with compassion and respect for human dignity.
II. A physician shall deal honestly with patients and colleagues, and strive to expose those physicians deficient in character or competence, or who engage in fraud or deception.

competence, or engaging in fraud or deception, to appropriate entities.

III. A physician shall respect the law and also recognize a responsibility to seek changes in those requirements which are contrary to the best interests of the patient.

IV. A physician shall respect the rights of patients, colleagues, and other health professionals, and shall safeguard patient confidences within the constraints of the law.

V. A physician shall continue to study, apply, and advance scientific knowledge; make relevant information available to patients, colleagues, and the public; obtain consultation; and use the talents of other health professionals when indicated.

VI. A physician shall, in the provision of appropriate patient care, except in emergencies, be free to choose whom to serve, with whom to associate, and the environment in which to provide medical care.

VII. A physician shall recognize a responsibility to participate in activities contributing to the improvement of the community and the betterment of public health.

VIII. A physician shall, while caring for a patient, regard responsibility to the patient as paramount.

IX. A physician shall support access to medical care for all people.

III. A physician shall respect the law and also recognize a responsibility to seek changes in those requirements which are contrary to the best interests of the patient.

IV. A physician shall respect the rights of patients, of colleagues, and of other health professionals, and shall safeguard patient confidences within the constraints of the law.

V. A physician shall continue to study, apply and advance scientific knowledge; make relevant information available to patients, colleagues, and the public; obtain consultation; and use the talents of other health professionals when indicated.

VI. A physician shall, in the provision of appropriate patient care, except in emergencies, be free to choose whom to serve, with whom to associate, and the environment in which to provide medical services.

VII. A physician shall recognize a responsibility to participate in activities contributing to an improved community.

For the nursing profession, the American Nurses Association formally adopted in 1976 the *Code For Nurses,* whose statements and interpretations provide guidance for conduct and relationships in carrying out nursing responsibilities:

PREAMBLE: The *Code for Nurses* is based on belief about the nature of individuals, nursing, health, and society. Recipients and providers of nursing services are viewed as individuals and groups who possess basic rights and responsibilities, and whose values and circumstances

command respect at all times. Nursing encompasses the promotion and restoration of health, the prevention of illness, and the alleviation of suffering. The statements of the Code and their interpretation provide guidance for conduct and relationships in carrying out nursing responsibilities consistent with the ethical obligations of the profession and quality in nursing care.

> *The nurse provides services with respect for human dignity and the uniqueness of the client, unrestricted by considerations of social or economic status, personal attributes, or the nature of health problems.*
> *The nurse safeguards the client's right to privacy by judiciously protecting information of a confidential nature.*
> *The nurse acts to safeguard the client and the public when health care and safety are affected by the incompetent, unethical, or illegal practice of any person.*
> *The nurse assumes responsibility and accountability for individual nursing judgments and actions.*
> *The nurse maintains competence in nursing.*
> *The nurse exercises informed judgment and uses individual competence and qualifications as criteria in seeking consultation, accepting responsibilities, and delegating nursing activities to others.*
> *The nurse participates in activities that contribute to the ongoing development of the profession's body of knowledge.*
> *The nurse participates in the profession's efforts to implement and improve standards of nursing.*
> *The nurse participates in the profession's efforts to establish and maintain conditions of employment conducive to high-quality nursing care.*
> *The nurse participates in the profession's effort to protect the public from misinformation and misrepresentation and to maintain the integrity of nursing.*
> *The nurse collaborates with members of the health professions and other citizens in promoting community and national efforts to meet the health needs of the public.*

These codes take as their guiding principle the concepts of service to humankind and respect for human life. When reading these codes of conduct, it is difficult to imagine that anyone could improve on them as summary statements of the primary goals of individuals responsible for the care of patients. However, some believe that such codes fail to provide answers to many of the difficult moral dilemmas confronting health professionals today. For example, in many situations, all the fundamental responsibilities of the nurse cannot be met at the same time. When a patient suffering from a massive insult to the brain is kept alive by artificial means and this equipment is needed elsewhere, it is not clear from these guidelines how "nursing competence is to be maintained to conserve life and promote health." Although it may be argued that the decision to treat or not to treat is a medical and not a nursing decision, both professions are so intimately involved in the care of patients that they are both concerned with the ultimate implications of any such decision.

For biomedical engineers, an increased awareness of the ethical significance of their professional activities has also resulted in the development of codes of professional ethics. Typically consisting of a short list of general rules, these codes express both the minimal standards to which all members of a profession are expected to conform and the ideals for which all members are expected to strive. Such codes provide a practical guide for the ethical conduct of the profession's practitioners. Consider, for example, the code of ethics endorsed by the American College of Clinical Engineers:

As a member of the American College of Clinical Engineering, I subscribe to the established Code of Ethics in that I will:

- *Accurately represent my level of responsibility, authority, experience, knowledge, and education.*
- *Strive to prevent a person from being placed at risk due to the use of technology.*
- *Reveal conflicts of interest that may affect information provided or received.*
- *Respect the confidentiality of information.*
- *Work toward improving the delivery of health care.*
- *Work toward the containment of costs by the better management and utilization of technology.*
- *Promote the profession of clinical engineering.*

Although these codes can be useful in promoting ethical conduct, such rules obviously cannot provide ethical guidance in every situation. A profession that aims to maximize the ethical conduct of its members must not limit the ethical consciousness of its members to knowledge of their professional code alone. It must also provide them with resources that will enable them to determine what the code requires in a particular concrete situation and thereby enable them to arrive at ethically sound judgments in situations in which the directives of the code are ambiguous or simply do not apply.

2.2 TWO MORAL NORMS: BENEFICENCE AND NONMALEFICENCE

Two moral norms have remained relatively constant across the various moral codes and oaths that have been formulated for health care providers since the beginnings of Western medicine in classical Greek civilization. They are *beneficence,* which is the provision of benefits, and *nonmaleficence,* which is the avoidance of doing harm. These norms are traced back to a body of writings from classical antiquity known as the *Hippocratic Corpus.* Although these writings are associated with the name of Hippocrates, the acknowledged founder of Western medicine, medical historians remain uncertain whether any of them, including the *Hippocratic Oath,* were actually his work. Although portions of the *Corpus* are believed to have been authored during the sixth century BC, other portions are believed to have been written as late as the beginning of the Christian era. Medical historians agree that many of the specific moral directives of the *Corpus* represent neither the actual practices nor the moral ideals of the majority of physicians of Ancient Greece and Rome.

Nonetheless, the general injunction *"As to disease, make a habit of two things: (1) to help or, (2) at least, to do no harm"* was accepted as a fundamental medical ethical norm by at least some ancient physicians. With the decline of Hellenistic civilization and the rise of Christianity, beneficence and nonmaleficence became increasingly accepted as the fundamental principles of morally sound medical practice. Although beneficence and nonmaleficence were regarded merely as concomitant to the craft of medicine in classical Greece and Rome, the emphasis upon compassion and the brotherhood of humankind, central to Christianity, increasingly made these norms the only acceptable motives for medical practice. Even today, the provision of benefits and the avoidance of doing harm are stressed just as much in virtually all contemporary Western codes of conduct for health professionals as they were in the oaths and codes that guided the health-care providers of past centuries.

Traditionally, the ethics of medical care has given greater prominence to nonmaleficence than to beneficence. This priority was grounded in the fact that, historically, medicine's capacity to do harm far exceeded its capacity to protect and restore health. Providers of health care possessed many treatments that posed clear and genuine risks to patients and that offered little prospect of benefit. Truly effective therapies were all too rare. In this context, it is surely rational to give substantially higher priority to avoiding harm than to providing benefits.

The advent of modern science changed matters dramatically. Knowledge acquired in laboratories, tested in clinics, and verified by statistical methods has increasingly dictated the practice of medicine. This ongoing alliance between medicine and science became a critical source of the plethora of technologies that now pervade medical care. The impressive increases in therapeutic, preventive, and rehabilitative capabilities that these technologies have provided have pushed beneficence to the forefront of medical morality. Some have even gone so far as to hold that the old medical ethic of *"Above all, do no harm"* should be superseded by the new ethic *"The patient deserves the best."* However, the rapid advances in medical technology capabilities have also produced great uncertainty as to what is most beneficial or least harmful for the patient. In other words, along with increases in ability to be beneficent, medicine's technology has generated much debate about what actually counts as beneficent or nonmaleficent treatment. Having reviewed some of the fundamental concepts of ethics and morality, let us now turn to several specific moral issues posed by the use of medical technology.

2.3 REDEFINING DEATH

Although medicine has long been involved in the observation and certification of death, many of its practitioners have not always expressed philosophical concerns regarding the beginning of life and the onset of death. Since medicine is a clinical and empirical science, it would seem that health professionals had no medical need to consider the concept of death: the fact of death was sufficient. The distinction between life and death was viewed as the comparison of two extreme conditions separated by an infinite chasm. With the advent of technological advances in medicine to assist health professionals to prolong life, this view has changed.

There is no doubt that the use of medical technology has in many instances warded off the coming of the grim reaper. One need only look at the trends in average life expectancy for confirmation. For example, in the United States today, the average life expectancy for males is 74.3 years and for females 76 years, whereas in 1900 the average life expectancy for both sexes was only 47 years. Infant mortality has been significantly reduced in developed nations where technology is an integral part of the culture. Premature births no longer constitute a threat to life because of the artificial environment that medical technology can provide. Today, technology has not only helped individuals avoid early death but has also been effective in delaying the inevitable. Pacemakers, artificial kidneys, and a variety of other medical devices have enabled individuals to add many more productive years to their lives. Technology has been so successful that health professionals responsible for the care of critically ill patients have been able to maintain their "vital signs of life" for extensive periods of time. In the process, however, serious philosophical questions concerning the quality of the life provided these patients have arisen.

Consider the case of the patient who sustains a serious head injury in an automobile accident. To the attendants in the ambulance who reached the scene of the accident, the patient was unconscious but still alive with a beating heart. After the victim was rushed to the hospital and into the emergency room, the resident in charge verified the stability of the vital signs of heartbeat and respiration during examination and ordered a computerized tomography (CT) scan to indicate the extent of the head injury. The results of this procedure clearly showed extensive brain damage. When the EEG was obtained from the scalp electrodes placed about the head, it was noted to be significantly abnormal. In this situation, then, the obvious questions arise: What is the status of the patient? Is the patient alive?

Alternatively, consider the events encountered during one open-heart surgery. During this procedure, the patient was placed on the heart bypass machine while the surgeon attempted to correct a malfunctioning valve. As the complex and long operation continued, the EEG monitors that had indicated a normal pattern of electrical activity at the onset of the operation suddenly displayed a relatively straight line indicative of feeble electrical activity. However, since the heart-lung bypass was maintaining the patient's so-called vital signs, what should the surgeon do? Should the medical staff continue on the basis that the patient is alive, or is the patient dead?

The increasing occurrence of these situations has stimulated health professionals to reexamine the definition of "death." In essence, advances in medical technology that delay death actually hastened its redefinition. This should not be so surprising because the definition of death has always been closely related to the extent of medical knowledge and available technology. For many centuries, death was defined solely as the absence of breathing. Since it was felt that the spirit of the human being resided in the spiritus (breath), its absence became indicative of death. With the continuing proliferation of scientific information regarding human physiology and the development of techniques to revive a nonbreathing person, attention turned to the pulsating heart as the focal point in determination of death. However, this view was to change through additional medical and technological advances in supportive therapy, resuscitation, cardiovascular assist devices, and organ transplantation.

As understanding of the human organism increased, it became obvious that one of the primary constituents of the blood is oxygen and that any organ deprived of oxygen for a specified period of time will cease to function and die. The higher functions of the brain are particularly vulnerable to this type of insult, since the removal of oxygen from the blood supply even for a short period of time (three minutes) produces irreversible damage to the brain tissues. Consequently, the evidence of "death" began to shift from the pulsating heart to the vital, functioning brain. Once medicine was provided with the means to monitor the brain's activity (i.e., the EEG), another factor was introduced in the definition of death. Advocates of the concept of brain death argued that the human brain is truly essential to life. When the brain is irreversibly damaged, so are the functions that are identified with self and our own humanness: memory, feeling, thinking, knowledge, and so on.

As a result, it became widely accepted that the meaning of clinical death implies that the spontaneous activity of the lungs, heart, and brain is no longer present. The irreversible cessation of functioning of all three major organs—the heart, lungs, and brain—was required before anyone was pronounced dead. Although damage to any other organ system such as the liver or kidney may ultimately cause the death of the individual through a fatal effect on the essential functions of the heart, lungs, or brain, this aspect was not included in the definition of clinical death.

With the development of modern respirators, however, the medical profession encountered an increasing number of situations in which a patient with irreversible brain damage could be maintained almost indefinitely. Once again, a new technological advance created the need to reexamine the definition of death.

The movement toward redefining death received considerable impetus with the publication of a report sponsored by the Ad Hoc Committee of the Harvard Medical School in 1968, in which the committee offered an alternative definition of death based on the functioning of the brain. The report of this committee was considered a landmark attempt to deal with death in light of technology.

In summary, the criteria for death established by this committee included the following: (1) the patient must be unreceptive and unresponsive—that is, in a state of irreversible coma; (2) the patient must have no movements of breathing when the mechanical respirator is turned off; (3) the patient must not demonstrate any reflexes; and (4) the patient must have a flat EEG for at least 24 hours, indicating no electrical brain activity. When these criteria are satisfied, then death may be declared.

At the time, the committee also strongly recommended that the decision to declare the person dead and then to turn off the respirator should not be made by physicians involved in any later efforts to transplant organs or tissues from the deceased individual. In this way, a prospective donor's death would not be hastened merely for the purpose of transplantation. Thus, complete separation of authority and responsibility for the care of the recipient from the physician or group of physicians who are responsible for the care of the prospective donor is essential.

The shift to a brain-oriented concept involved deciding that much more than just biological life is necessary to be a human person. The brain death concept was essentially a statement that mere vegetative human life is not personal human life. In other words, an otherwise intact and alive but brain-dead person is not a human person. Many of us have taken for granted the assertion that being truly alive in this world requires an "intact functioning brain." Yet, precisely this issue was at stake in the gradual movement from using heartbeat and respiration as indices of life to using brain-oriented indices instead.

Indeed, total and irreparable loss of brain function, referred to as "brainstem death," "whole brain death," or, simply, "brain death," has been widely accepted as the legal standard for death. By this standard, an individual in a state of brain death is legally indistinguishable from a corpse and may be legally treated as one even though respiratory and circulatory functions may be sustained through the intervention of technology. Many take this legal standard to be the morally appropriate one, noting that once destruction of the brainstem has occurred, the brain cannot function at all, and the body's regulatory mechanisms will fail unless artificially sustained. Thus mechanical sustenance of an individual in a state of brain death is merely postponement of the inevitable and sustains nothing of the personality, character, or consciousness of the individual. It is simply the mechanical intervention that differentiates such an individual from a corpse, and a mechanically ventilated corpse is a corpse nonetheless.

Even with a consensus that brainstem death is death, and thus that an individual in such a state is indeed a corpse, difficult cases remain. Consider the case of an individual in a persistent vegetative state, the condition known as "neocortical death." Although severe brain injury has been suffered, enough brain function remains to make mechanical sustenance of respiration and circulation unnecessary. In a persistent vegetative state, an individual exhibits no purposeful response to external stimuli and no evidence of self-awareness. The eyes

may open periodically, and the individual may exhibit sleep-wake cycles. Some patients even yawn, make chewing motions, or swallow spontaneously. Unlike the complete unresponsiveness of individuals in a state of brainstem death, a variety of simple and complex responses can be elicited from an individual in a persistent vegetative state. Nonetheless, the chances that such an individual will regain consciousness are remote. Artificial feeding, kidney dialysis, and the like make it possible to sustain an individual in a state of neocortical death for decades.

If brainstem death is death, is neocortical death also death? Again, the issue is not a straightforward factual matter. For it, too, is a matter of specifying which features of living individuals distinguish them from corpses and so make treatment of them as corpses morally impermissible. Irreparable cessation of respiration and circulation, the classical criterion for death, would entail that an individual in a persistent vegetative state is not a corpse and so, morally speaking, must not be treated as one. The brainstem death criterion for death would also entail that a person in a state of neocortical death is not yet a corpse. On this criterion, what is crucial is that brain damage be severe enough to cause failure of the regulatory mechanisms of the body.

Is an individual in a state of neocortical death any less in possession of the characteristics that distinguish the living from cadavers than one whose respiration and circulation are mechanically maintained? It is a matter that society must decide. And until society decides, it is not clear what counts as beneficent or nonmaleficent treatment of an individual in a state of neocortical death.

CASE STUDY: TERRI SCHIAVO AND THE BRAIN DEATH DEBATE

In February 1990, an otherwise healthy 27-year-old Terri Schiavo suffered heart failure in her home and fell into a coma. While Schiavo ultimately woke and initially proved responsive, after a year of multiple rehabilitation facilities and nursing homes, the by then 28-year-old was diagnosed as in an irreversible persistent vegetative state (PVS). In 1998, Schiavo's husband, Michael Schiavo, made a petition to the Florida courts to remove his wife from life support, a petition fought vehemently by the woman's parents.

In 2001, after a doctor confirmed brain death with a report of significant brainstem damage and 80 percent loss of upper brain function, Schiavo's feeding tube was removed, but was replaced days later, following a Court Appeal by her parents. Ultimately, the feeding tube was ordered to be removed on three separate occasions, each time her legal guardian and husband fighting to allow his wife to "die in peace," while her parents insisted that their daughter maintained cognitive function and requested more tests.

Finally in 2005, 15 years after her injury, and under constant national media coverage, Schiavo died from dehydration two weeks after her tube had been removed for the final time and while her case was still pending with the highest word in the nation: the Supreme Court.

1. Without a Living Will, who is responsible for deciding the would-be intentions of a victim of brain death? Who is responsible for mediation when loved ones disagree?
2. Who is responsible for the years of health care costs of a potentially brain dead individual? In 2006, Rom Houben, a man presumed brain dead for 23 years, was discovered to have full brain function after a series of advanced brain scan imaging tests. Houben was paralyzed in an accident

in 1983 at the age of 20. A formal martial arts enthusiast, Houben was assumed to be in a PVS for over two decades. After therapy he is now able to communicate via typing, and he reads books while lying down, using an assistive device. "I want to read," says Houben, via keyboard, "talk with my friends via the computer, and enjoy my life now that people know I am not dead."

3. With diagnosis technologies constantly in development, should PVS victims ever be "allowed to die"?

2.4 THE TERMINALLY ILL PATIENT AND EUTHANASIA

Terminally ill patients today often find themselves in a strange world of institutions and technology devoted to assisting them in their fight against death. However, at the same time, this modern technologically oriented medical system may cause patients and their families considerable economic, psychological, and physical pain. In enabling medical science to prolong life, modern technology has in many cases made dying slower and more undignified. As a result of this situation, there is a moral dilemma in medicine. Is it right or wrong for medical professionals to stop treatment or administer a lethal dose to terminally ill patients?

This problem has become a major issue for our society to consider. Although death is all around us in the form of accidents, drug overdoses, alcoholism, murders, and suicides, for most of us, the end lies in growing older and succumbing to some form of chronic illness. As the aged approach the end of life's journey, they may eventually wish for the day when all troubles can be brought to an end. Such a desire, frequently shared by a compassionate family, is often shattered by therapies provided with only one concern: to prolong life regardless of the situation. As a result, many claim a dignified death is often not compatible with today's standard medical view.

Consider the following hypothetical version of the kind of case that often confronts contemporary patients, their families, health care workers, and society as a whole. Suppose a middle-aged man suffers a brain hemorrhage and loses consciousness as a result of a ruptured aneurysm. Suppose that he never regains consciousness and is hospitalized in a state of neocortical death, a chronic vegetative state. His life is maintained by a surgically implanted gastronomy tube that drips liquid nourishment from a plastic bag directly into his stomach. The care of this individual takes seven and one-half hours of nursing time daily and includes shaving, oral hygiene, grooming, attending to his bowels and bladder, and so forth. Suppose further that his wife undertakes legal action to force his caregivers to end all medical treatment, including nutrition and hydration so complete bodily death of her husband will occur. She presents a preponderance of evidence to the court to show that her husband would have wanted just this result in these circumstances.

The central moral issue raised by this sort of case is whether the quality of the individual's life is sufficiently compromised to make intentional termination of that life morally permissible. While alive, he made it clear to both family and friends that he would prefer to be allowed to die rather than be mechanically maintained in a condition of irretrievable loss of consciousness. Deciding whether his judgment in such a case should be allowed requires deciding which capacities and qualities make life worth living, which qualities are sufficient to endow it with value worth sustaining, and whether their absence justifies deliberate termination of a life, at least when this would be the wish of the individual in question. Without this decision, the traditional norms of medical ethics, beneficence and

nonmaleficence, provide no guidance. Without this decision, it cannot be determined whether termination of life support is a benefit or harm to the patient.

For many individuals, the fight for life is a correct professional view. They believe that the forces of medicine should always be committed to using innovative ways of prolonging life for the individual. However, this cannot be the only approach to caring for the terminally ill. Certain moral questions regarding the extent to which physicians engaged in heroic efforts to prolong life must be addressed if the individual's rights are to be preserved. The goal of those responsible for patient care should not solely be to prolong life as long as possible by the extensive use of drugs, operations, respirators, hemodialyzers, pacemakers, and the like, but rather to provide a reasonable "quality of life" for each patient. It is out of this new concern that euthanasia has once again become a controversial issue in the practice of medicine.

The term *euthanasia* is derived from two Greek words meaning "good" and "death." Euthanasia was practiced in many primitive societies in varying degrees. For example, on the island of Cos, the ancient Greeks assembled elderly and sick people at an annual banquet to consume a poisonous potion. Even Aristotle advocated euthanasia for gravely deformed children. Other cultures acted in a similar manner toward their aged by abandoning them when they felt these individuals no longer served any useful purpose. However, with the spread of Christianity in the Western world, a new attitude developed toward euthanasia. Because of the Judeo-Christian belief in the biblical statements "Thou shalt not kill" (Exodus 20: 13) and "He who kills a man should be put to death" (Leviticus 24: 17), the practice of euthanasia decreased. As a result of these moral judgments, killing was considered a sin, and the decision about whether someone should live or die was viewed solely as God's responsibility, not humans'.

In today's society, euthanasia implies to many "death with dignity," a practice to be followed when life is merely being prolonged by machines and no longer seems to have value. In many instances, it has come to mean a contract for the termination of life in order to avoid unnecessary suffering at the end of a fatal illness and, therefore, has the connotation of relief from pain.

Discussions of the morality of euthanasia often distinguish active from passive euthanasia, a distinction that rests upon the difference between an act of commission and an act of omission. When failure to take steps that could effectively forestall death results in an individual's demise, the resultant death is an act of omission and a case of letting a person die. When a death is the result of doing something to hasten the end of a person's life (for example, giving a lethal injection), that death is caused by an act of commission and is a case of killing a person. The important difference between active and passive euthanasia is that in passive euthanasia, the physician does not do anything to bring about the patient's death. The physician does nothing, and death results due to whatever illness already afflicts the patient. In active euthanasia, however, the physician does something to bring about the patient's death. The physician who gives the patient with cancer a lethal injection has caused the patient's death, whereas if the physician merely ceases treatment, the cancer is the cause of death.

In active euthanasia, someone must do something to bring about the patient's death, and in passive euthanasia, the patient's death is caused by illness rather than by anyone's conduct. Is this notion correct? Suppose a physician deliberately decides not to treat a patient who is terminally ill, and the patient dies. Suppose further that the physician were to attempt to exonerate himself by saying, "I did nothing. The patient's death was the result

of illness. I was not the cause of death." Under current legal and moral norms, such a response would have no credibility. The physician would be blameworthy for the patient's death as surely as if he or she had actively killed the patient. Thus, the actions taken by a physician to continue treatment to the very end are understood.

Euthanasia may also be classified as *involuntary* or *voluntary*. An act of euthanasia is involuntary if it hastens the individual's death for his or her own good, but against their wishes. Involuntary euthanasia, therefore, is no different in any morally relevant way from unjustifiable homicide. However, what happens when the individual is incapable of agreeing or disagreeing? Suppose that a terminally ill person is unconscious and cannot make his or her wishes known. Would hastening their death be permissible? It would be if there was substantial evidence that the individual has given prior consent. The individual may have told friends and relatives that, under certain circumstances, efforts to prolong their life should not be undertaken or continued and might even have recorded their wishes in the form of a living will or an audio- or videotape. When this level of substantial evidence of prior consent exists, the decision to hasten death would be morally justified. A case of this sort would be a case of voluntary euthanasia.

For a living will to be valid, the person signing it must be of sound mind at the time the will is made and shown not to have altered their opinion in the interim between its signing and the onset of the illness. In addition, the witnesses must not be able to benefit from the individual's death. As the living will itself states, it is not a legally binding document. It is essentially a passive request and depends on moral persuasion. Proponents of the will, however, believe that it is valuable in relieving the burden of guilt often carried by health professionals and the family in making the decision to allow a patient to die.

Those who favor euthanasia point out the importance of individual rights and freedom of choice and look on euthanasia as a kindness ending the misery endured by the patient. The thought of a dignified death is much more attractive than the process of continuous suffering and gradual decay into nothingness. Viewing each person as a rational being possessing a unique mind and personality, proponents argue that terminally ill patients should have the right to control the ending of their own life.

On the other hand, those opposed to euthanasia demand to know who has the right to end the life of another. Some use religious arguments, emphasizing that euthanasia is in direct conflict with the belief that only God has the power to decide when a human life ends. Their view is that anyone who practices euthanasia is essentially acting in the place of God and that no human should ever be considered omnipotent.

Others turn to the established codes, reminding those responsible for the care of patients that they must do whatever is in their power to save a life. Their argument is that health professionals cannot honor their pledge and still believe that euthanasia is justified. If terminally ill patients are kept alive, there is at least a chance of finding a cure that might be useful to them. Opponents of euthanasia feel that legalizing it would destroy the bonds of trust between doctor and patient. How would sick individuals feel if they could not be sure that their physician and nurse would try everything possible to cure them, knowing that if their condition worsened, they would just lose faith and decide that death would be better? Opponents of euthanasia also question whether it will be truly beneficial to the suffering person or will only be a means to relieve the agony of the family. They believe that destroying life (no matter how minimal) merely to ease the emotional suffering of others is indeed unjust.

Many fear that if euthanasia is legalized, it will be difficult to define and develop clear-cut guidelines that will serve as the basis for euthanasia to be carried out. Furthermore, once any form of euthanasia is accepted by society, its detractors fear that many other problems will arise. Even the acceptance of passive euthanasia could, if carried to its logical conclusion, be applied in state hospitals and institutions for the mentally handicapped and the elderly. Such places currently house thousands of people who have neither hope nor any prospect of a life that even approaches normality. Legalization of passive euthanasia could prompt an increased number of suits by parents seeking to end the agony of incurably afflicted children or by children seeking to shorten the suffering of aged and terminally ill parents. In Nazi Germany, for example, mercy killing was initially practiced to end the suffering of the terminally ill. Eventually, however, the practice spread so even persons with the slightest deviation from the norm (e.g., the mentally ill, minority groups such as Jews and others) were terminated. Clearly, the situation is delicate and thought provoking.

2.5 TAKING CONTROL

Medical care decisions can be tremendously difficult. They often involve unpleasant topics and arise when we are emotionally and physically most vulnerable. Almost always these choices involve new medical information that feels alien and can seem overwhelming. In an attempt to assist individuals to make these decisions, it is often helpful to follow these steps:

1. Obtain all the facts—that is, clarify the medical facts of the situation.
2. Understand all options and their consequences.
3. Place a value on each of the options based upon your own set of personal values.

A LIVING WILL

TO MY FAMILY, MY PHYSICIAN, MY CLERGYMAN, MY LAWYER:

If the time comes when I can no longer take part in decisions about my own future, let this statement stand as testament of my wishes: If there is no reasonable expectation of my recovery from physical or mental disability, I request that I be allowed to die and not be kept alive by artificial means or heroic measures. Death is as much a reality as birth, growth, maturity, and old age—it is the one certainty. I do not fear death as much as I fear the indignity of deterioration, dependence, and hopeless pain. I ask that drugs be mercifully administered to me for the terminal suffering even if they hasten the moment of death. This request is made after careful consideration. Although this document is not legally binding, you who care for me will, I hope, feel morally bound to follow its mandate. I recognize that it places a heavy burden of responsibility upon you, and it is with the intention of sharing that responsibility and of mitigating any feelings of guilt that this statement is made.

Signed _____

Date _____

Witnessed by _____

The three-step facts/options/values path concerns the "how" of decisions, but equally important is the "who." Someone must make every single medical decision. Ideally it will be made by the person most intimately involved: the patient. Very often, however, it is made by someone else—spouse, family, physician—or a group of those people acting on behalf of the patient. It is, therefore, important to recognize the four concentric circles of consent:

- The first, and primary, circle is the patient.
- The second circle is the use of advance directives—that is, choosing in advance through the use of such documents as the living will.
- The third circle is others deciding for the patient—that is, the move from personal control to surrogate control.
- The fourth and final circle is the courts and bureaucrats. It is the arena of last resort where our society has decreed that we go when the patient may be incapacitated, where there is no clear advance directive, and where it is not clear who should make the decision.

These three steps and four circles are simply attempts to impose some order on the chaos that is medical decision making. They can help individuals take control.

2.6 HUMAN EXPERIMENTATION

Medical research has long held an exalted position in our modern society. It has been acclaimed for its significant achievements that range from the development of the Salk and Sabin vaccines for polio to the development of artificial organs. In order to determine their effectiveness and value, however, these new drugs and medical devices eventually are used on humans. The issue is, therefore, not only whether humans should be involved in clinical studies designed to benefit themselves or their fellow humans but also clarifying or defining more precisely the conditions under which such studies are to be permitted.

For example, consider the case of a 50-year-old female patient suffering from severe coronary artery disease. What guidelines should be followed in the process of experimenting with new drugs or devices that may or may not help her? Should only those procedures viewed as potentially beneficial to her be tried? Should experimental diagnostic procedures or equipment be tested on this patient to evaluate their effectiveness when compared to more accepted techniques, even though they will not be directly beneficial to the patient?

On the other hand, consider the situation of conducting research on the human fetus. This type of research is possible as a result of the legalization of abortion in the United States, as well as the technological advances that have made fetal studies more practical than in the past. Under what conditions should medical research be conducted on these subjects? Should potentially hazardous drugs be given to women planning to have abortions to determine the effect of these drugs on the fetus? Should the fetus, once aborted, be used in any experimental studies? Although these questions are difficult to answer, clinical researchers responsible for the well-being of their patients must face the moral issues involved in testing new equipment and procedures and at the same time safeguard the individual rights of their patients.

CASE STUDY: NEONATAL INTENSIVE CARE UNIT (NICU)

Throughout time, low birth weight, oftentimes arising from premature birth, has been a major factor affecting infant survival. Underweight infants, who are typically classified as either low birth weight (LBW) (less than 1,500 g) or very low birth weight (VLBW) (less than 1,000 g), must be treated with the utmost caution and care in order to maximize their chances of survival. Advances in premature-infant medical care, such as improved thermoregulation and ventilation techniques, have greatly decreased the mortality rate among LBW and VLBW infants. Included in these advances was the creation of the NICU (Figure 2.2), where all the necessary equipment needed to sustain the life of the child could be kept conveniently in close proximity to one another.

One of the most important devices used in the NICU is the incubator. This device, typically molded of see-through plastic, is used to stabilize the body temperature of the infant. In essence, the incubator allows the medical staff to keep the newborn warm without having to wrap it in

FIGURE 2.2 A Neonatal Intensive Care Unit. *Courtesy of http://www.pediatrics.ucsd.edu/Divisions/Neonatology/Pictures/Image%20Library/NICU%20Bed.jpg.*

blankets. The incubator also aids in preventing infection, as well as in stabilizing the humidity of the child's environment. By keeping the temperature and humidity levels of the newborn's environment static, the baby remains well hydrated and water loss is kept to a minimum.

A complication that many preterm infants suffer from is the inability to breathe normally on their own. The child may be completely unable to breathe for himself, or he may suffer from a condition known as apnea, where the breathing pattern is either aperiodic or irregular.

In these cases, children susceptible to an apneic event are closely monitored so if they stop breathing, nurses can rush to the bedside and wake them up. However, it is often minutes before the nurse can arrive at the scene. To facilitate the process of waking the infant experiencing an apneic event, biomedical engineers developed a tactile vibrator that when triggered by such an event vibrates against the infant's foot and wakes her. In order to prove that the device is effective and safe, a human experiment must be initiated. In this case, the following questions need to be resolved:

1. Who is responsible for proposing the conduction of this study?
2. What should the process of approval of such a study include?
3. What should the policy be related to informed consent?
4. Should changes that were made in the device during the course of the study, which would alter the nature of the initially proposed device, be allowed?

2.7 DEFINITION AND PURPOSE OF EXPERIMENTATION

One may ask, what exactly constitutes a human experiment? Although experimental protocols may vary, it is generally accepted that human experimentation occurs whenever the clinical situation of the individual is consciously manipulated to gather information regarding the capability of drugs and devices. In the past, experiments involving human subjects have been classified as either therapeutic or nontherapeutic. A *therapeutic experiment* is one that may have direct benefit for the patient, while the goal of *nontherapeutic research* is to provide additional knowledge without direct benefit to the person. The central difference is a matter of intent or aim rather than results.

Throughout medical history, there have been numerous examples of therapeutic research projects. The use of nonconventional radiotherapy to inhibit the progress of a malignant cancer, of pacemakers to provide the necessary electrical stimulation for proper heart function, or of artificial kidneys to mimic nature's function and remove poisons from the blood were all, at one time, considered novel approaches that might have some value for the patient. In the process, they were tried and found not only to be beneficial for the individual patient but also for humankind.

Nontherapeutic research has been another important vehicle for medical progress. Experiments designed to study the impact of infection from the hepatitis virus or the malarial parasite or the procedures involved in cardiac catheterization have had significant impacts on the advancement of medical science and the ultimate development of appropriate medical procedures for the benefit of all humans.

In the mid-1970s, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research offered the terms *practice* and *research* to replace the

conventional therapeutic and nontherapeutic distinction just mentioned. Quoting the commission, Alexander Capron in 1986 wrote the following:

> The term *practice* refers to interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success. In the medical sphere, practices usually involve diagnosis, preventive treatment, or therapy; in the social sphere, practices include governmental programs such as transfer payments, education, and the like.
>
> By contrast, the term *research* designates an activity designed to test a hypothesis, to permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge (expressed, for example, in theories, principles, or statements of relationships). In the polar cases, then, practice uses a proven technique in an attempt to benefit one or more individuals, while research studies a technique in an attempt to increase knowledge.

Although the practice/research dichotomy has the advantage of not implying that therapeutic activities are the only clinical procedures intended to benefit patients, it is also based on intent rather than outcome. Interventions are "practices" when they are proven techniques intended to benefit the patient, while interventions aimed at increasing generalizable knowledge constitute research. What about those interventions that do not fit into either category?

CASE STUDY: THE ARTIFICIAL HEART

In the early 1980s, a screening committee had been set up to pick the first candidate for the "Jarvik 7," a new (at the time) artificial heart (Figure 2.3). It was decided that the first recipient had to be someone so sick that death was imminent. It was thought unethical to pick someone who might have another year to live when the artificial heart might well kill the patient immediately.

FIGURE 2.3 The Jarvik-7 artificial heart, 1985. *Courtesy of http://www.smithsonianlegacies.si.edu/objectdescription.cfm?ID=172.*

1. Is this an example of nonvalidated practice?
2. Is informed consent still required?

   A week after the operation, Barney Clark began having seizures from head to toe. Suffering a seizure, Clark's unconscious body quivered for several hours. The seizures and spells of mental confusion continued throughout the next months. As a result, Clark expressed a desire to die. Although he did issue a positive statement during a videotaped interview, Clark was not a happy man, tethered to a huge machine, barely conscious, and in some pain. In March 1983, Barney Clark died of multiple organ collapse.

3. Discuss in detail the notions of "criteria for success" and quality of life in this case.
4. Barney Clark suffered a great deal. In response to this, who should be the responsible party in deciding what is right for the patient? When both sides hope for positive results, is it possible to make an unbiased decision based on what's best for the patient?

One such intervention is "nonvalidated practice," which may encompass prevention as well as diagnosed therapy. The primary purpose of the use of a nonvalidated practice is to benefit the patient while emphasizing that it has not been shown to be safe and efficacious. For humans to be subjected to nonvalidated practice, they must be properly informed and give their consent.

2.8 INFORMED CONSENT

Informed consent has long been considered by many to be the most important moral issue in human experimentation. It is the principal condition that must be satisfied in order for human experimentation to be considered both lawful and ethical. All adults have the legal capacity to give medical consent (unless specifically denied through some legal process). As a result, issues concerning legal capability are usually limited to minors. Many states, if not all, have some exceptions that allow minors to give consent.

Informed consent is an attempt to preserve the rights of individuals by giving them the opportunity for self-determination—that is, to determine for themselves whether they wish to participate in any experimental effort. In 1964, the World Medical Association (WMA) in Finland endorsed a code of ethics for human experimentation as an attempt to provide some guidelines in this area. In October 2000, the 52nd WMA General Assembly in Edinburgh, Scotland, revised these guidelines.

Because it is often essential to use the results obtained in human experiments to further scientific knowledge, the World Medical Association prepared the following recommendations to serve as a guide to physicians all over the world. However, it is important to point out that these guidelines do not relieve physicians, scientists, and engineers from criminal, civil, and ethical responsibilities dictated by the laws of their own countries.

2.8.1 Basic Principles

- Biomedical research involving human subjects must conform to generally accepted scientific principles and should be based on adequately performed laboratory and animal experimentation and on a thorough knowledge of the scientific literature.

- The design and performance of each experimental procedure involving human subjects should be clearly formulated in an experimental protocol, which should be transmitted to a specially appointed independent committee for consideration, comment, and guidance.
- Biomedical research involving human subjects should be conducted only by scientifically qualified persons and under the supervision of a clinically competent medical person. The responsibility for the human subject must always rest with a medically qualified person and never rest on the subject of the research, even though the subject has given his or her consent.
- Biomedical research involving human subjects cannot legitimately be carried out unless the importance of the objective is in proportion to the inherent risk to the subject.
- Every biomedical research project involving human subjects should be preceded by careful assessment of predictable risks in comparison with foreseeable benefits to the subject or to others. Concern for the interests of the subject must always prevail over the interests of science and society.
- The right of the research subject to safeguard his or her integrity must always be respected. Every precaution should be taken to respect the privacy of the subject and to minimize the impact of the study on the subject's physical and mental integrity and on the personality of the subject.
- Doctors should abstain from engaging in research projects involving human subjects unless they are satisfied that the hazards involved are believed to be predictable. Doctors should cease any investigation if the hazards are found to outweigh the potential benefits.
- In publication of the results of his or her research, the doctor is obliged to preserve the accuracy of the results. Reports of experimentation not in accordance with the principles laid down in this Declaration should not be accepted for publication.
- In any research on human beings, each potential subject must be adequately informed of the aims, methods, anticipated benefits and potential hazards of the study and the discomfort it may entail. He or she should be informed that he or she is at liberty to abstain from participation in the study and that he or she is free to withdraw his or her consent to participation at any time. The doctor should then obtain the subject's freely-given informed consent, preferably in writing.
- When obtaining informed consent for the research project, the doctor should be particularly cautious if the subject is in a dependent relationship to him or her or may consent under duress. In that case, the informed consent should be obtained by a doctor who is not engaged in the investigation and who is completely independent of this official relationship.
- In the case of legal incompetence, informed consent should be obtained from the legal guardian in accordance with national legislation. Where physical or mental incapacity makes it impossible to obtain informed consent, or when the subject is a minor, permission from the responsible relative replaces that of the subject in accordance with national legislation.
- The research protocol should always contain a statement of the ethical considerations involved and should indicate that the principles enunciated in the present Declaration are complied with.

2.8.2 Medical Research Combined with Professional Care

- In the treatment of the sick person, the doctor must be free to use a new diagnostic and therapeutic measure if in his or her judgment it offers hope of saving life, reestablishing health, or alleviating suffering.
- The potential benefits, hazards, and discomfort of a new method should be weighed against the advantages of the best current diagnostic and therapeutic methods.
- In any medical study, every patient—including those of a control group, if any— should be assured of the best-proven diagnostic and therapeutic method.
- The refusal of the patient to participate in a study must never interfere with the doctor-patient relationship.
- If the doctor considers it essential not to obtain informed consent, the specific reasons for this proposal should be stated in the experimental protocol for transmission to the independent committee.
- The doctor can combine medical research with professional care, the objective being the acquisition of new medical knowledge, only to the extent that medical research is justified by its potential diagnostic or therapeutic value for the patient.

2.8.3 Nontherapeutic Biomedical Research Involving Human Subjects

- In the purely scientific application of medical research carried out on a human being, it is the duty of the doctor to remain the protector of the life and health of that person on whom biomedical research is being carried out.
- The subjects should be volunteers—that is, either healthy persons or patients for whom the experimental design is not related to the patient's illness.
- The investigator or the investigating team should discontinue the research if in his/her or their judgment it may, if continued, be harmful to the individual.
- In research on humans, the interest of science and society should never take precedence over considerations related to the well-being of the subject.

These guidelines generally converge on six basic requirements for ethically sound human experimentation. First, research on humans must be based upon prior laboratory research and research on animals, as well as upon established scientific fact, so the point under inquiry is well focused and has been advanced as far as possible by nonhuman means. Second, research on humans should use tests and means of observation that are reasonably believed to be able to provide the information being sought by the research. Methods that are not suited for providing the knowledge sought are pointless and rob the research of its scientific value. Third, research should be conducted only by persons with the relevant scientific expertise. Fourth, all foreseeable risks and reasonably probable benefits, to the subject of the investigation and to science, or more broadly to society, must be carefully assessed, and the comparison of those projected risks and benefits must indicate that the latter clearly outweighs the former. Moreover, the probable benefits must not be obtainable through other less risky means. Fifth, participation in research should be based on informed and voluntary consent. Sixth, participation by a subject in an experiment should be halted immediately if the subject finds continued participation undesirable or a prudent investigator has cause to believe that the experiment is likely to result in

injury, disability, or death to the subject. Conforming to conditions of this sort probably does limit the pace and extent of medical progress, but society's insistence on these conditions is its way of saying that the only medical progress truly worth having must be consistent with a high level of respect for human dignity. Of these conditions, the requirement to obtain informed and voluntary consent from research subjects is widely regarded as one of the most important protections.

A strict interpretation of the criteria mentioned above for subjects automatically rules out whole classes of individuals from participating in medical research projects. Children, the mentally retarded, and any patient whose capacity to think is affected by illness are excluded on the grounds of their inability to comprehend exactly what is involved in the experiment. In addition, those individuals having a dependent relationship to the clinical investigator, such as the investigator's patients and students, would be eliminated based on this constraint. Since mental capacity also includes the ability of subjects to appreciate the seriousness of the consequences of the proposed procedure, this means that even though some minors have the right to give consent for certain types of treatments, they must be able to understand all the risks involved.

Any research study must clearly define the risks involved. The patient must receive a total disclosure of all known information. In the past, the evaluation of risk and benefit in many situations belonged to the medical professional alone. Once made, it was assumed that this decision would be accepted at face value by the patient. Today, this assumption is not valid. Although the medical staff must still weigh the risks and benefits involved in any procedure they suggest, it is the patient who has the right to make the final determination. The patient cannot, of course, decide whether the procedure is medically correct, since that requires more medical expertise than the average individual possesses. However, once the procedure is recommended, the patient then must have enough information to decide whether the hoped-for benefits are sufficient to risk the hazards. Only when this is accomplished can a valid consent be given.

Once informed and voluntary consent has been obtained and recorded, the following protections are in place:

- It represents legal authorization to proceed. The subject cannot later claim assault and battery.
- It usually gives legal authorization to use the data obtained for professional or research purposes. Invasion of privacy cannot later be claimed.
- It eliminates any claims in the event that the subject fails to benefit from the procedure.
- It is defense against any claim of an injury when the risk of the procedure is understood and consented to.
- It protects the investigator against any claim of an injury resulting from the subject's failure to follow safety instructions if the orders were well explained and reasonable.

CASE STUDY: CONFIDENTIALITY, PRIVACY, AND CONSENT

Integral to the change currently taking place in the United States health care industry is the application of computer technology to the development of a health care information system. Most major hospitals in the United States have now updated their systems to entirely electronic databases. Patient medications are scheduled and followed by a nurse on a computer module present

FIGURE 2.4 MRI scans are just one of the PPI elements available on electronic databases. *Courtesy of http://images.medicinenet.com/images/SlideShow/dementia_s21_mri_doctor.jpg.*

in every hospital room. MRI scans (Figure 2.4) are no longer printed in film but are uploaded to the patient's file, where physicians with proper approval can access the images. Entire medical histories are stored on patient databases.

1. Discuss the benefits of the electronic system and the potential risks.
2. Discuss in detail where and how the issue of consent to access should be handled. While access to the patient information database is limited to accredited physicians and employees, the issue of illegal access from the inside is a prominent one. Hospital employees with access, be they physicians or researchers, may be capable of accessing family accounts or those of friends. With a paper system, protected patient information (PPI) had the potential to be leaked as well via lost files and irresponsible handling. With the electronic system, however, more intentional breach of privacy may be possible to those with access to the system.
3. How can a hospital employee with a medical record on the system be guaranteed privacy from colleagues?
4. Should patients be allowed to decide personally whether their information is stored electronically? Would an integrated system function efficiently?

Nevertheless, can the aims of research ever be reconciled with the traditional moral obligations of physicians? Is the researcher/physician in an untenable position? Informed and voluntary consent once again is the key only if subjects of an experiment agree to participate in the research. What happens to them during and because of the experiment is then a product of their own decision. It is not something that is imposed on them but rather, in a very real sense, something they elected to have done to themselves. Because their autonomy is thus respected, they are not made a mere resource for the benefit of others. Although they may suffer harm for the benefit of others, they do so of their own volition as a result of the exercise of their own autonomy, rather than as a result of having their autonomy limited or diminished.

For consent to be genuine, it must be truly voluntary and not the product of coercion. Not all sources of coercion are as obvious and easy to recognize as physical violence.

A subject may be coerced by fear that there is no other recourse for treatment, by the fear that nonconsent will alienate the physician on whom the subject depends for treatment, or even by the fear of disapproval of others. This sort of coercion, if it truly ranks as such, is often difficult to detect and, in turn, to remedy.

Finally, individuals must understand what they are consenting to do. Therefore, they must be given information sufficient to arrive at an intelligent decision concerning whether to participate in the research. Although a subject need not be given all the information a researcher has, it is important to determine how much should be provided and what can be omitted without compromising the validity of the subject's consent. Another difficulty lies in knowing whether the subject is competent to understand the information given and to render an intelligent opinion based upon it. In any case, efforts must be made to ensure that sufficient relevant information is given and that the subject is sufficiently competent to process it. These are matters of judgment that probably cannot be made with absolute precision and certainty, but rigorous efforts must be made in good faith to prevent research on humans from involving gross violations of human dignity.

2.9 REGULATION OF MEDICAL DEVICE INNOVATION

The Food and Drug Administration (FDA) is the sole federal agency charged by Congress with regulating medical devices to ensure their safety and effectiveness. Unlike food and drugs, which have been regulated by the FDA since 1906, medical devices first became subject to FDA regulation in 1938. At that time, the FDA's major concern was to ensure that legitimate medical devices were in the marketplace and were truthfully labeled, not misbranded. Over time, the scope of FDA review of medical devices has evolved, as has the technology employed by medical devices). The first substantive legislative attempt to address the premarket review of all medical devices occurred with the Medical Device Amendment of 1976 (Pub. L. No. 94-295, 90 Stat. 539). This statute requires approval from the FDA before new devices are marketed and imposes requirements for the clinical investigation of new medical devices on human subjects. For details related to the FDA process, visit http://www.fda.gov/.

The FDA is organized into five major program centers: the Center for Biologics Evaluation and Research, the Center for Drug Evaluation and Research, the Center for Food Safety and Applied Nutrition, the Center for Veterinary Medicine, and the Center for Devices and Radiological Health (CDRH). Each FDA program center has primary jurisdiction over a different subject area. According to the FDA, the CDRH is responsible for ensuring the safety and effectiveness of medical devices and eliminating unnecessary human exposure to man-made radiation from medical, occupational, and consumer products.

The CDRH has six distinct offices: the Office of Systems and Management, the Office of Compliance, the Office of Science and Technology, the Office of Health and Industry Programs, the Office of Surveillance and Biometrics, and the Office of Device Evaluation (ODE). The ODE has several principal functions, including the following:

- Advising the CDRH director on all premarket notification 510(k) submissions, premarket approvals (PMAs), device classifications, and investigational device exemptions (IDEs).

- Planning, conducting, and coordinating CDRH actions regarding approval, denial, and withdrawals of 510(k)s, PMAs, and IDEs.
- Ongoing review, surveillance, and medical evaluation of the labeling, clinical experience, and required reports submitted by sponsors of approval applications.
- Developing and interpreting regulations and guidelines regarding the classification of devices, 510(k)'s, PMAs, and IDEs.
- Participating in the development of national and international consensus standards.

Everyone who develops or markets a medical device will likely have multiple interactions with ODE before, during, and after the development of a medical device.

In principle, if a manufacturer makes medical claims about a product, it is considered a device, and may be subject to FDA pre- and postmarket regulatory controls (Figure 2.5). The device definition distinguishes a medical device from other FDA-regulated products, such as drugs. According to the FDA, a medical device is:

> An instrument, apparatus, machine, contrivance, implant, in vitro reagent, or other similar or related article intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease in man or other animals OR intended to affect the structure or any function of the body of man or other animals, and which does not achieve any of its primary intended purposes through chemical action or is not dependent upon being metabolized.

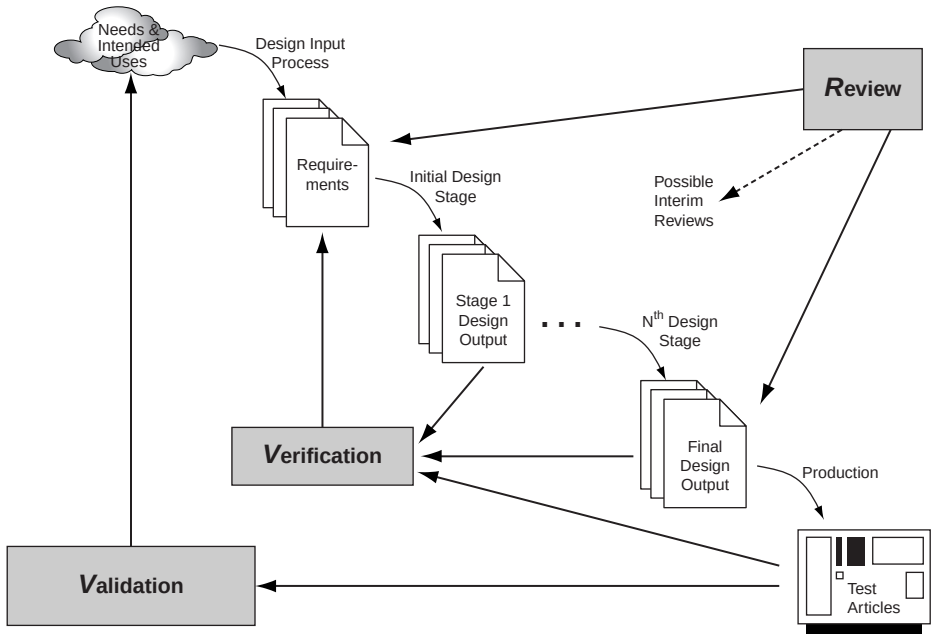


FIGURE 2.5 The purpose of the regulatory process is to conduct product review to ensure (1) device safety and effectiveness, (2) quality of design, and (3) surveillance to monitor device quality. Therefore, the review process results in verification and validation of the medical device.

2.10 MARKETING MEDICAL DEVICES

The four principal routes to marketing a medical device in the United States are as follows.

Premarket Approval (PMA)

A marketing approach for high-risk (Class III) medical devices must be accomplished through a PMA unless the device can be marketed through the 510(k) process (see following). The PMA hinges on the FDA determining that the medical device is safe and effective. The PMA process can be quite costly. The collection of the data required for a PMA may costs hundreds of thousands, if not several million, dollars. Moreover, the timeline for a PMA applicant to collect the requisite data could take several years. However, an approved PMA is akin to a private license granted to the applicant to market a particular medical device, because other firms seeking to market the same type of device for the same use must also have an approved PMA.

Investigational Device Exemption (IDE)

The IDE is an approved regulatory mechanism that permits manufacturers to receive an exemption for those devices solely intended for investigational use on human subjects (clinical evaluation). Because an IDE is specifically for clinical testing and not commercial distribution, the FDCA authorizes the FDA to exempt these devices from certain requirements that apply to devices in commercial distribution. The clinical evaluation of all devices may not be cleared for marketing, unless otherwise exempt by resolution, requires an IDE. An IDE may be obtained either by an institutional review board (IRB), or an IRB and the FDA.

Product Development Protocol (PDP)

An alternative to the IDE and PMA processes for Class III devices subject to premarket approval, the PDP is a mechanism allowing a sponsor to come to early agreement with the FDA as to what steps are necessary to demonstrate the safety and effectiveness of a new device. In the years immediately subsequent to the enactment of the Medical Device Amendment, the FDA did not focus its energies on the PDP but worked to effectively implement the major provisions of the Amendment, including device classification systems, and the 510(k) and PMA processes.

510(k) Notification

Unless specifically exempted by federal regulation, all manufacturers are required to give the FDA 90 days' notice before they intend to introduce a device to the U.S. market by submitting a 510(k). During that 90-day period, the FDA is charged with determining whether the device is or is not substantially equivalent to a pre-Amendment device. The premarket notification is referred to in the industry as a 510(k) because 510(k) is the relevant section number of the FDCA. The 510(k) is used to demonstrate that the medical device is or is not substantially equivalent to a legally marketed device.

With respect to clinical research on humans, the FDA distinguishes devices into two categories: devices that pose significant risk and those that involve insignificant risk. Examples of the former included orthopedic implants, artificial hearts, and infusion pumps.

Examples of the latter include various dental devices and contact lenses. Clinical research involving a significant risk device cannot begin until an institutional review board (IRB) has approved both the protocol and the informed consent form and the FDA itself has given permission. This requirement to submit an IDE application to the FDA is waived in the case of clinical research where the risk posed is insignificant. In this case, the FDA requires only that approval from an IRB be obtained certifying that the device in question poses only insignificant risk. In deciding whether to approve a proposed clinical investigation of a new device, the IRB and the FDA must determine the following:

1. That risk to subjects is minimized.
2. That risks to subjects are reasonable in relation to anticipated benefit and knowledge to be gained.
3. That subject selection is equitable.
4. That informed consent materials and procedures are adequate.
5. That provisions for monitoring the study and protecting patient information are acceptable.

The FDA allows unapproved medical devices to be used without an IDE in three types of situations: feasibility studies, emergency use, and treatment use.

2.11 ETHICAL ISSUES IN FEASIBILITY STUDIES

In a feasibility study, or "limited investigation," human research involving the use of a new device would take place at a single institution and involve no more than ten human subjects. The sponsor of a limited investigation is required to submit to the FDA a "Notice of Limited Investigation," which includes a description of the device, a summary of the purpose of the investigation, the protocol, a sample of the informed consent form, and a certification of approval by the responsible medical board. In certain circumstances, the FDA could require additional information or require the submission of a full IDE application or suspend the investigation.

Investigations of this kind are limited to (1) investigations of new uses for existing devices, (2) investigations involving temporary or permanent implants during the early developmental stages, and (3) investigations involving modification of an existing device.

To comprehend adequately the ethical issues posed by clinical use of unapproved medical devices outside the context of an IDE, it is necessary to use the distinctions among practice, nonvalidated practice, and research elaborated upon in the previous pages. How do those definitions apply to feasibility studies?

Clearly, the goal of the feasibility study, which is a generalizable knowledge, makes it an instance of research rather than practice. Manufacturers seek to determine the performance of a device with respect to a particular patient population in an effort to gain information about its efficacy and safety. Such information is important in order to determine whether further studies (animal or human) need to be conducted, whether the device needs modification before further use, and the like. The main difference between using an unapproved device in a feasibility study and using it under the terms of an IDE is that the former would be subject to significantly less intensive FDA review than the latter. This, in turn, means

that the responsibility for ensuring that the use of the device is ethically sound would fall primarily to the IRB of the institution conducting the study.

The ethical concerns posed here can be best comprehended only with a clear understanding of what justifies research in the first place. Ultimately, no matter how much basic research and animal experimentation has been conducted on a given device, the risks and benefits it poses for humans cannot be adequately determined until it is actually used on humans. The benefit of research on humans lies primarily in the generalizable information that is provided. This information is crucial to medical science's ability to generate new modes of medical treatment that are both efficacious and safe. Therefore, one condition for experimentation to be ethically sound is that it must be scientifically sound.

Although scientific soundness is a necessary condition of ethically sound research on humans, it is not of and by itself sufficient. The human subjects of such research are at risk of being mere research resources—that is, having value only for the ends of the research. Human beings are not valuable wholly or solely for the uses to which they can be put. They are valuable simply by being the kinds of entities they are. To treat them as such is to respect them as people. Treating individuals as people means respecting their autonomy. This requirement is met by ensuring that no competent person is subjected to any clinical intervention without first giving voluntary and informed consent. Furthermore, respect for people means that the physician will not subject a human to unnecessary risks and will minimize the risks to patients in required procedures.

Much of the scrutiny that the FDA imposes upon use of unapproved medical devices in the context of an IDE addresses two conditions of ethically sound research: Is the experiment scientifically sound? and Does it respect the rights of the human subjects involved? Medical ethicists argue that decreased FDA scrutiny will increase the likelihood that either or both of these conditions will not be met. This possibility exists because many manufacturers of medical devices are, after all, commercial enterprises, companies that are motivated to generate profit and thus to get their devices to market as soon as possible with as little delay and cost as possible. These self-interest motives are likely, at times, to conflict with the requirements of ethically sound research and thus to induce manufacturers to fail to meet these requirements. Profit is not the only motive that might induce manufacturers to contravene the requirements of ethically sound research on humans. A manufacturer may sincerely believe that its product offers great benefit to many people and be prompted to take shortcuts that compromise the quality of the research. Whether the consequences being sought by the research are desired for reasons of self-interest, altruism, or both, the ethical issue is the same. Research subjects may be placed at risk of being treated as mere objects rather than as people.

What about the circumstances under which feasibility studies would take place? Are these not sufficiently different from the "normal" circumstances of research to warrant reduced FDA scrutiny? As just noted, manufacturers seek to engage in feasibility studies in order to investigate new uses of existing devices, to investigate temporary or permanent implants during the early developmental stages, and to investigate modifications to an existing device. As also noted, a feasibility study would take place at only one institution and would involve no more than ten human subjects. Given these circumstances, is the sort of research that is likely to occur in a feasibility study less likely to be scientifically sound or to fail to respect people than normal research upon humans in "normal" circumstances?

Research in feasibility studies would be done on a very small subject pool, and the harm of any ethical lapses would likely affect fewer people than if such lapses occurred under more usual research circumstances. Yet even if the harm done is limited to ten or fewer subjects in a single feasibility study, the harm is still ethically wrong. To wrong ten or fewer people is not as bad as to wrong in the same way more than ten people, but it is to engage in wrongdoing nonetheless.

Are ethical lapses more likely to occur in feasibility studies than in studies that take place within the requirements of an IDE? Although nothing in the preceding discussion provides a definitive answer to this question, it is a question to which the FDA should give high priority. The answer to this question might be quite different when the device at issue is a temporary or permanent implant than when it is an already approved device being put to new uses or modified in some way. Whatever the contemplated use under the feasibility studies mechanism, the FDA would be ethically advised not to allow this kind of exception to IDE use of an unapproved device without a reasonably high level of certainty that research subjects would not be placed in greater jeopardy than in "normal" research circumstances.

2.12 ETHICAL ISSUES IN EMERGENCY USE

What about the mechanism for avoiding the rigors of an IDE for emergency use? The FDA has authorized emergency use in instances where an unapproved device offers the only alternative for saving the life of a dying patient. However, what if an IDE has not yet been approved for the device or its use, or an IDE has been approved but the physician who wishes to use the device is not an investigator under the IDE?

The purpose of emergency use of an unapproved device is to attempt to save a dying patient's life under circumstances where no other alternative is available. This sort of use constitutes practice rather than research. Its aim is primary benefit to the patient rather than provision of new and generalizable information. Because this sort of use occurs before the completion of clinical investigation of the device, it constitutes a nonvalidated practice. What does this mean?

First, it means that while the aim of the use is to save the life of the patient, the nature and likelihood of the potential benefits and risks engendered by use of the device are far more speculative than in the sort of clinical intervention that constitutes validated practice. In validated practice, thorough investigation of a device, including preclinical studies, animal studies, and studies on human subjects, has established its efficacy and safety. The clinician thus has a well-founded basis upon which to judge the benefits and risks such an intervention poses for the patient.

It is precisely this basis that is lacking in the case of a nonvalidated practice. Does this mean that emergency use of an unapproved device should be regarded as immoral? This conclusion would follow only if there were no basis upon which to make an assessment of the risks and benefits of the use of the device. The FDA requires that a physician who engages in emergency use of an unapproved device must have substantial reason to believe that benefits will exist. This means that there should be a body of preclinical and animal tests allowing a prediction of the benefit to a human patient.

CASE STUDY: MEDICAL EXPERT SYSTEMS

Expert systems have been developed in various disciplines, including clinical decision making. These systems have been designed to simulate the decision-making skills of physicians. Their adaptability, however, depends on the presence of an accepted body of knowledge regarding the "prescribed path" physicians would take given specific input data. These systems have been viewed as "advisory systems" providing the clinician with suggested/recommended courses of action. The ultimate decision remains with the physician.

Consider one such system designed to monitor drug treatment in a psychiatric clinic. This system, designed and implemented by biomedical engineers working with clinicians, begins by the entry of a specific diagnosis and immediately recommends the appropriate drugs to be considered for the treatment of someone who has that mental disorder (Figure 2.6). The physician selects

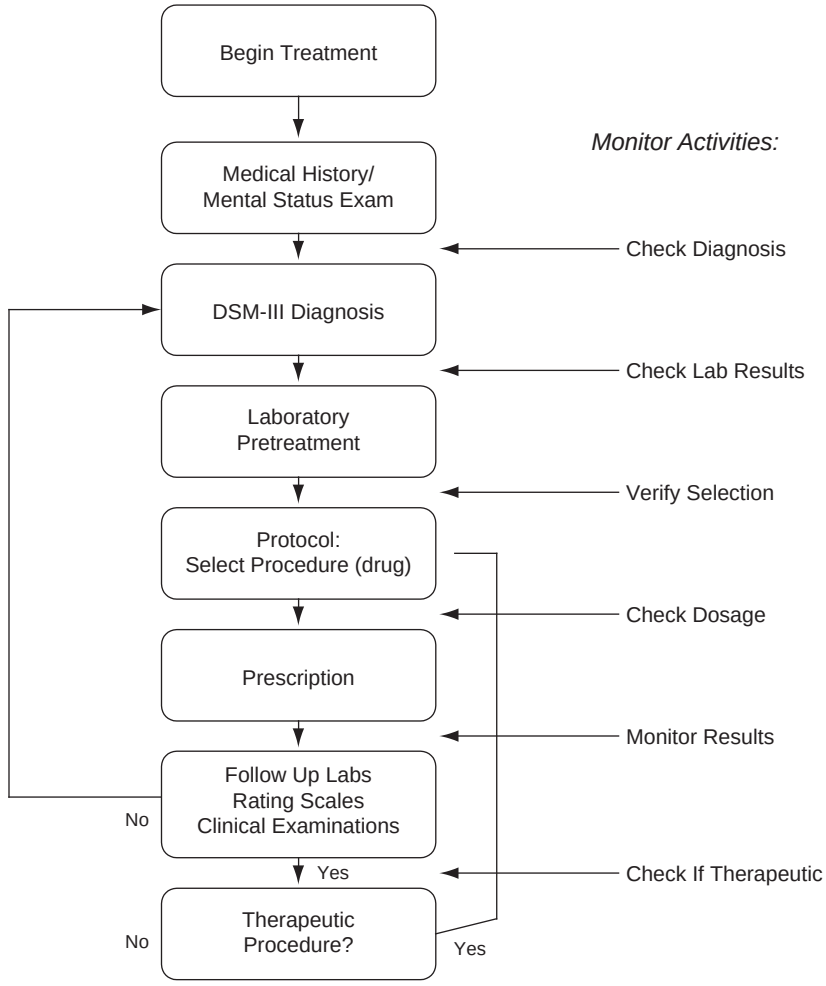


FIGURE 2.6 The drug treatment process followed by clinicians.

one of the recommended drugs and conducts a dose regimen to determine the effectiveness of the drug for the particular patient. During the treatment, blood tests are conducted to ascertain the presence of drug toxicity, and other psychiatric measures obtained to determine if the drug is having the desired effect.

As these data elements are entered, they are compared with standard expected outcomes, and if the outcomes are outside the expected limits, an alert is sent to the physician indicating further action needs to be taken.

In this situation:

1. Who is liable for mistreatment—the clinician, the programmer, or the systems administrator?
2. What constitutes mistreatment?
3. What is the role of the designers of such a system; in other words, what constitutes a successful design?
4. How does the clinic evaluate the performance of a physician using the system, as well as the system itself?

Thus, although the benefits and risks posed by use of the device are highly speculative, they are not entirely speculative. Although the only way to validate a new technology is to engage in research on humans at some point, not all nonvalidated technologies are equal. Some will be largely uninvestigated, and assessment of their risks and benefits will be wholly or almost wholly speculative. Others will at least have the support of preclinical and animal tests. Although this is not sufficient support for incorporating use of a device into regular clinical practice, it may, however, represent sufficient support to justify use in the desperate circumstances at issue in emergency situations. Desperate circumstances can justify desperate actions, but desperate actions are not the same as reckless actions, hence the ethical soundness of the FDA's requirement that emergency use be supported by solid results from preclinical and animal tests of the unapproved device.

A second requirement that the FDA imposes on the emergency use of unapproved devices is the expectation that physicians "exercise reasonable foresight with respect to potential emergencies and . . . make appropriate arrangements under the IDE procedures." Thus, a physician should not "create" an emergency in order to circumvent IRB review and avoid requesting the sponsor's authorization of the unapproved use of a "device." From a Kantian point of view, which is concerned with protecting the dignity of people, this is a particularly important requirement. To create an emergency in order to avoid FDA regulations is to treat the patient as a mere resource whose value is reducible to service to the clinician's goals. Hence, the FDA is quite correct to insist that emergencies are circumstances that reasonable foresight would not anticipate.

Also especially important here is the nature of the patient's consent. Individuals facing death are especially vulnerable to exploitation and deserve greater measures for their protection than might otherwise be necessary. One such measure would be to ensure that the patient, or his legitimate proxy, knows the highly speculative nature of the intervention being offered—that is, to ensure that it is clearly understood that the clinician's estimation of the intervention's risks and benefits is far less solidly grounded than in the case of validated practices. The patient's consent must be based on an awareness that the device whose use is contemplated has not undergone complete and rigorous testing on humans and that

estimations of its potential are based wholly on preclinical and animal studies. Above all, the patient must not be led to believe that the risks and benefits of the intervention are not better understood than they in fact are. Another important point is to ensure that the patient understands all of the options—not simply life or death, but also a life with severely impaired quality. Although desperate circumstances may legitimate desperate actions, the decision to take such actions must rest upon the informed and voluntary consent of the patient, certainly for an especially vulnerable patient.

It is important here for a clinician involved in emergency use of an unapproved device to recognize that these activities constitute a form of practice, albeit nonvalidated, and not research. Hence, the primary obligation is to the well-being of the patient. The patient enters into the relationship with the clinician with the same trust that accompanies any normal clinical situation. Treating this sort of intervention as if it were an instance of research and, thus, justified by its benefits to science and society would be an abuse of this trust.

2.13 ETHICAL ISSUES IN TREATMENT USE

The FDA has adopted regulations authorizing the use of investigational new drugs in certain circumstances where a patient has not responded to approved therapies. This "treatment use" of unapproved new drugs is not limited to life-threatening emergency situations but is also available to treat "serious" diseases or conditions. The FDA has not approved treatment use of unapproved medical devices, but it is possible that a manufacturer could obtain such approval by establishing a specific protocol for this kind of use within the context of an IDE.

The criteria for treatment use of unapproved medical devices would be similar to criteria for treatment use of investigational drugs: (1) the device is intended to treat a serious or life-threatening disease or condition; (2) there is no comparable or satisfactory alternative product available to treat that condition; (3) the device is under an IDE or has received an IDE exemption, or all clinical trials have been completed and the device is awaiting approval; and (4) the sponsor is actively pursuing marketing approval of the investigational device. The treatment use protocol would be submitted as part of the IDE and would describe the intended use of the device, the rationale for use of the device, the available alternatives and why the investigational product is preferable, the criteria for patient selection, the measures to monitor the use of the device and to minimize risk, and technical information that is relevant to the safety and effectiveness of the device for the intended treatment purpose.

Were the FDA to approve treatment use of unapproved medical devices, what ethical issues would be posed? First, because such use is premised on the failure of validated interventions to improve the patient's condition adequately, it is a form of practice rather than research. Second, since the device involved in an instance of treatment use is unapproved, such use would constitute nonvalidated practice. As such, like emergency use, it should be subject to the FDA's requirement that prior preclinical tests and animal studies have been conducted that provide substantial reason to believe that patient benefit will result. As with emergency use, although this does not prevent assessment of the intervention's benefits and risks from being highly speculative, it does prevent assessment from being totally speculative. Here, too, although desperate circumstances can justify desperate action, they do not

justify reckless action. Unlike emergency use, the circumstances of treatment use involve serious impairment of health rather than the threat of premature death. Hence, an issue that must be considered is how serious such impairment must be to justify resorting to an intervention whose risks and benefits have not been solidly established.

In cases of emergency use, the FDA requires that physicians not create an exception to an IDE to avoid requirements that would otherwise be in place. As with emergency use of unapproved devices, the patients involved in treatment uses would be particularly vulnerable patients. Although they are not dying, they are facing serious medical conditions and are thereby likely to be less able to avoid exploitation than patients under less desperate circumstances. Consequently, here too it is especially important that patients be informed of the speculative nature of the intervention and of the possibility that treatment may result in little to no benefit to them.

2.14 THE ROLE OF THE BIOMEDICAL ENGINEER IN THE FDA PROCESS

On November 28, 1991, the *Safe Medical Devices Act of 1990* (Public Law 101-629) went into effect. This regulation requires a wide range of health care institutions, including hospitals, ambulatory-surgical facilities, nursing homes, and outpatient treatment facilities, to report information that "reasonably suggests" the likelihood that the death, serious injury, or serious illness of a patient at that facility was caused or contributed to by a medical device. When a death is device-related, a report must be made directly to the FDA and to the manufacturer of the device. When a serious illness or injury is device-related, a report must be made to the manufacturer or to the FDA in cases where the manufacturer is not known. In addition, summaries of previously submitted reports must be submitted to the FDA on a semiannual basis. Prior to this regulation, such reporting was wholly voluntary. This new regulation was designed to enhance the FDA's ability to learn quickly about problems related to medical devices and supplements the medical device reporting (MDR) regulations promulgated in 1984. MDR regulations require that manufacturers and importers submit reports of device-related deaths and serious injuries to the FDA. The new law extends this requirement to users of medical devices along with manufacturers and importers. This act gives the FDA authority over device-user facilities.

The FDA regulations are ethically significant because by attempting to increase the FDA's awareness of medical device-related problems, it attempts to increase that agency's ability to protect the welfare of patients. The main controversy over the FDA's regulation policies is essentially utilitarian in nature. Skeptics of the law are dubious about its ability to provide the FDA with much useful information. They worry that much of the information generated by this new law will simply duplicate information already provided under MDR regulations. If this were the case, little or no benefit to patients would accrue from compliance with the regulation. Furthermore, these regulations, according to the skeptics, are likely to increase lawsuits filed against hospitals and manufacturers and will require device-user facilities to implement formal systems for reporting device-related problems and to provide personnel to operate those systems. This would, of course, add to the costs of health care and thereby exacerbate the problem of access to care, a situation that many

believe to be of crisis proportions already. In short, the controversy over FDA policy centers upon the worry that its benefits to patients will be marginal and significantly outweighed by its costs.

Biomedical engineers need to be aware of FDA regulations and the process for FDA approval of the use of medical devices and systems. These regulatory policies are, in effect, society's mechanism for controlling the improper use of these devices.

2.15 EXERCISES

1. Explain the distinction between the terms *ethics* and *morality*. Provide examples that illustrate this distinction in the medical arena.
2. Explain the distinction between the terms *beneficence* and *nonmaleficence*, and provide a real-world example of each. Which has been favored by medicine in the ethical sense?
3. Provide three examples of medical moral judgments.
4. What do advocates of the utilitarian school of thought believe?
5. What does Kantianism expect in terms of the patient's rights and wishes?
6. Discuss how the code of ethics for clinical engineers provides guidance to practitioners in the field.
7. Discuss what is meant by brainstem death. How is this distinguished from neocortical death?
8. In response to the Schiavo and Houben case studies, what steps, if any, can be taken to guarantee brain death? Should there be set procedures for determination?
9. Distinguish between active and passive euthanasia, as well as voluntary and involuntary euthanasia. In your view, which, if any, are permissible? Provide your reasoning and any conditions that must be satisfied to meet your approval.
10. Should the federal government be able to require an individual to sign a living will in case of an accident?
11. If the family of a patient in the intensive care unit submits the individual's "living will," should it be honored immediately, or should there be a discussion between physicians and the family? Who should make the decision? Why?
12. What constitutes a human experiment? Under what conditions are they permitted? What safeguards should hospitals have in place?
13. Should animal experimentation be required prior to human experimentation? How does animal research play into the philosophies of nonconsequentialism and utilitarianism?
14. Discuss the relationship between cost (or risk) and benefit in the decision for a patient to participate in a human experiment.
15. In the event of unfavorable and potentially painful results in consented human experimentation, who should be held liable? Why?
16. A biomedical engineer has designed a new sleep apnea monitor. Discuss the steps that should be taken before it is used in a clinical setting.
17. Discuss the distinctions among practice, research, and nonvalidated practice. Provide examples of each in the medical arena.
18. What are the two major conditions for ethically sound research?

19. Informed consent is one of the essential factors in permitting humans to participate in medical experiments. What ethical principles are satisfied by informed consent? What should be done to ensure it is truly voluntary? What information should be given to human subjects?

20. What are the distinctions between feasibility studies and emergency use?

21. In the practice of medicine, health care professionals use medical devices to diagnose and treat patients. Therefore, the clinical staff must not only become knowledgeable and skilled in their understanding of human physiology, but they must also be competent in using the medical tools at their disposal. This requirement often results in litigation when a device fails. The obvious question is, "Who is to blame?"

Consider the case of a woman undergoing a surgical procedure that requires the use of a ground plate—an 8 × 11-inch pad that serves as a return path for any electrical current that comes from electrosurgical devices used during the procedure. As a result of the procedure, this woman received a major burn that seriously destroyed tissue at the site of the ground plate.

(a) Discuss the possible individuals and/or organizations that may have been responsible for this injury.

(b) Outside of seeking the appropriate responsible party, are there specific ethical issues here?

Suggested Readings

N. Abrams, M.D. Buckner (Eds.), Medical Ethics, MIT Press, Cambridge, MA, 1983.
J.D. Bronzino, V.H. Smith, M.L. Wade, Medical Technology and Society, MIT Press, Cambridge, MA, 1990.
J.D. Bronzino, Management of Medical Technology, Butterworth, Boston, 1992.
A.R. Chapman, Health Care and Information Ethics: Protecting Fundamental Human Rights, Sheed and Ward, Kansas City, KS, 1997.
N. Dubler, D. Nimmons, Ethics on Call, Harmony Books, New York, 1992.
A.R. Jonsen, The New Medicine and The Old Ethics, Harvard University Press, Cambridge, MA, 1990.
J.C. Moskop, L. Kopelman (Eds.), Ethics and Critical Care Medicine, D. Reidel Publishing Co., Boston, 1985.
G.E. Pence, Classic Cases in Medical Ethics, McGraw-Hill, New York, 1990.
J. Rachels, Ethics at the End of Life: Euthanasia and Morality, Oxford University Press, Oxford, 1986.
J. Reiss, Bringing Your Medical Device to Market, FDLI Publishers, Washington, DC, 2001.
E.G. Seebauer, R.L. Barry, Fundamentals of Ethics for Scientists and Engineers, Oxford Press, NY, 2001.

This page intentionally left blank

CHAPTER

3

Anatomy and Physiology

Susan Blanchard, PhD, and Joseph D. Bronzino, PhD, PE

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Define *anatomy* and *physiology* and explain why they are important to biomedical engineering.
- Define important anatomical terms.
- Describe the cell theory.
- List the major types of organic compounds and other elements found in cells.
- Explain how the plasma membrane maintains the volume and internal concentrations of a cell.
- Calculate the internal osmolarity and ionic concentrations of a model cell at equilibrium.
- List and describe the functions of the major organelles found within mammalian cells.
- Describe the similarities, differences, and purposes of replication, transcription, and translation.
- List and describe the major components and functions of five organ systems: cardiovascular, respiratory, nervous, skeletal, and muscular.
- Define *homeostasis* and describe how feedback mechanisms help maintain it.

3.1 INTRODUCTION

Since biomedical engineering is an interdisciplinary field based in both engineering and the life sciences, it is important for biomedical engineers to have knowledge about and be able to communicate in both areas. Biomedical engineers must understand the basic components of the body and how they function well enough to exchange ideas and information with physicians and life scientists. Two of the most basic terms and areas of study in the life sciences are anatomy and physiology. *Anatomy* refers to the internal and external structures of the body and their physical relationships, whereas *physiology* refers to the study of the functions of those structures.

Figure 3.1a shows a male body in anatomical position. In this position, the body is erect and facing forward, with the arms hanging at the sides and the palms facing outward. This particular view shows the anterior (ventral) side of the body, whereas Figure 3.1c illustrates the posterior (dorsal) view of another male body that is also in anatomical position, and Figure 3.1b presents the lateral view of the female body. In clinical practice, directional terms are used to describe the relative positions of various parts of the body. Proximal parts are nearer to the trunk of the body or to the attached end of a limb than are distal parts (Figure 3.1a). Parts of the body that are located closer to the head than other parts when the body is in anatomical position are said to be *superior* (Figure 3.1b), whereas those located closer to the feet than other parts are termed *inferior*. *Medial* implies that a part is toward the midline of the body, whereas *lateral* means away from the midline (Figure 3.1c). Parts of the body that lie in the direction of the head are said to be in the cranial direction,

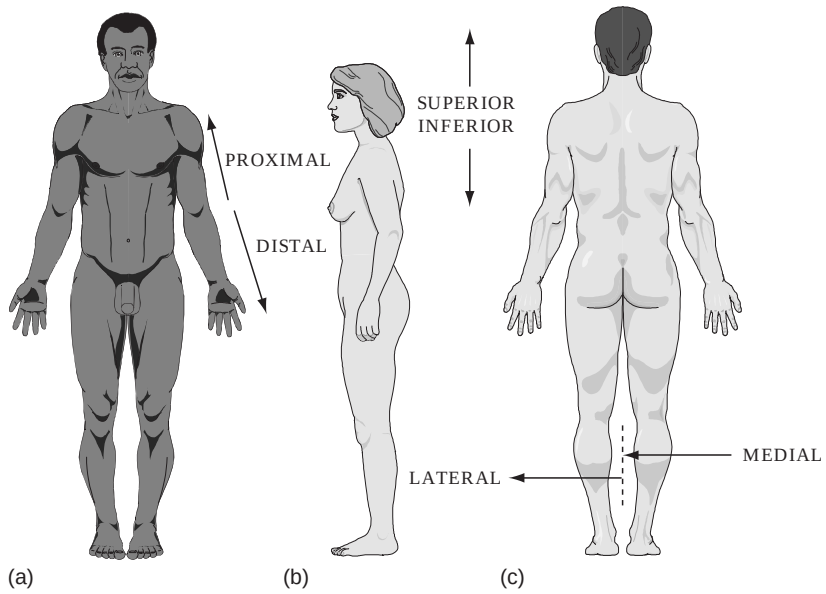


FIGURE 3.1 (a) Anterior view of male body in anatomical position. (b) Lateral view of female body. (c) Posterior view of male body in anatomical position. Relative directions (proximal and distal, superior and inferior, and medial and lateral) are also shown.

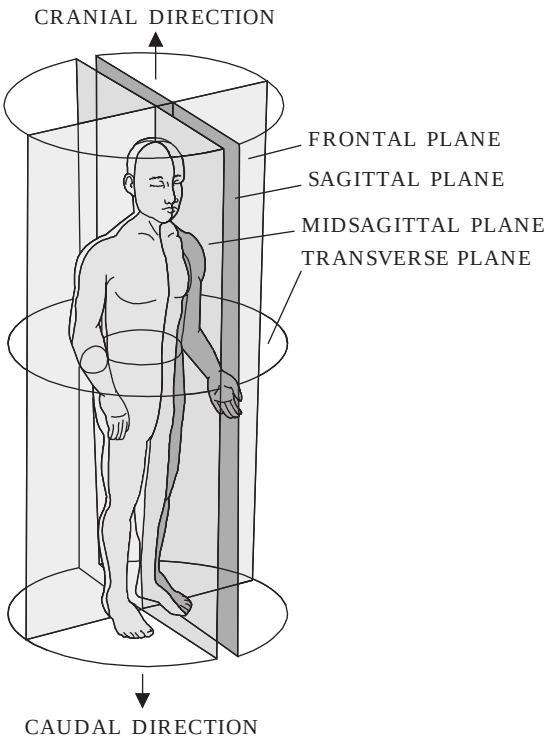


FIGURE 3.2 The body can be divided into sections by the frontal, sagittal, and transverse planes. The midsagittal plane goes through the midline of the body.

whereas those parts that lie in the direction of the feet are said to be in the caudal direction (Figure 3.2).

Anatomical locations can also be described in terms of planes. The plane that divides the body into two symmetric halves along its midline is called the midsaggital plane (Figure 3.2). Planes that are parallel to the midsaggital plane but do not divide the body into symmetric halves are called sagittal planes. The frontal plane is perpendicular to the midsaggital plane and divides the body into asymmetric anterior and posterior portions. Planes that cut across the body and are perpendicular to the midsaggital and frontal planes are called transverse planes.

Human bodies are divided into two main regions: axial and appendicular. The axial part consists of the head, neck, thorax (chest), abdomen, and pelvis, while the appendicular part consists of the upper and lower extremities. The upper extremities, or limbs, include the shoulders, upper arms, forearms, wrists, and hands, while the lower extremities include the hips, thighs, lower legs, ankles, and feet. The abdominal region can be further divided into nine regions or four quadrants.

The cavities of the body hold the internal organs. The major cavities are the dorsal and ventral body cavities, while smaller ones include the nasal, oral, orbital (eye), tympanic (middle ear), and synovial (movable joint) cavities. The dorsal body cavity includes the

cranial cavity that holds the brain and the spinal cavity that contains the spinal cord. The ventral body cavity contains the thoracic and abdominopelvic cavities that are separated by the diaphragm. The thoracic cavity contains the lungs and the mediastinum, which contains the heart and its attached blood vessels, the trachea, the esophagus, and all other organs in this region except for the lungs. The abdominopelvic cavity is divided by an imaginary line into the abdominal and pelvic cavities. The former is the largest cavity in the body and holds the stomach, small and large intestines, liver, spleen, pancreas, kidneys, and gallbladder. The latter contains the urinary bladder, the rectum, and the internal portions of the reproductive system.

The anatomical terms described previously are used by physicians, life scientists, and biomedical engineers when discussing the whole human body or its major parts. Correct use of these terms is vital for biomedical engineers to communicate with health care professionals and to understand the medical problem of concern or interest. While it is important to be able to use the general terms that describe the human body, it is also important for biomedical engineers to have a basic understanding of some of the more detailed aspects of human anatomy and physiology.

3.2 CELLULAR ORGANIZATION

Although there are many smaller units such as enzymes and organelles that perform physiological tasks or have definable structures, the smallest anatomical and physiological unit in the human body that can, under appropriate conditions, live and reproduce on its own is the cell. Cells were first discovered more than 300 years ago shortly after Antony van Leeuwenhoek, a Dutch optician, invented the microscope. With his microscope, van Leeuwenhoek was able to observe "many very small animalcules, the motions of which were very pleasing to behold" in tartar scrapings from his teeth. Following the efforts of van Leeuwenhoek, Robert Hooke, a Curator of Instruments for the Royal Society of England, in the late 1600s further described cells when he used one of the earliest microscopes to look at the plant cell walls that remain in cork. These observations and others led to the cell theory developed by Theodor Schwann and Matthias Jakob Schleiden and formalized by Rudolf Virchow in the mid-1800s. The cell theory states that (1) all organisms are composed of one or more cells, (2) the cell is the smallest unit of life, and (3) all cells come from previously existing cells. Thus, cells are the basic building blocks of life.

Cells are composed mostly of organic compounds and water, with more than 60 percent of the weight in a human body coming from water. The organic compounds—carbohydrates, lipids, proteins, and nucleic acids—that cells synthesize are the molecules that are fundamental to sustaining life. These molecules function as energy packets, storehouses of energy and hereditary information, structural materials, and metabolic workers. The most common elements found in humans (in descending order based on percent of body weight) are oxygen, carbon, hydrogen, nitrogen, calcium, phosphorus, potassium, sodium, chlorine, magnesium, sulfur, iron, and iodine. Carbon, hydrogen, oxygen, and nitrogen contribute more than 99 percent of all the atoms in the body. Most of these elements are incorporated into organic compounds, but some exist in other forms, such as phosphate groups and ions.

Carbohydrates are used by cells not only as structural materials but also to transport and store energy. The three classes of carbohydrates are monosaccharides (e.g., glucose),

oligosaccharides (e.g., lactose, sucrose, maltose), and polysaccharides (e.g., glycogen). Lipids are greasy or oily compounds that will dissolve in each other but not in water. They form structural materials in cells and are the main reservoirs of stored energy. Proteins are the most diverse form of biological molecules. Specialized proteins, called enzymes, make metabolic reactions proceed at a faster rate than would occur if the enzymes were not available and enable cells to produce the organic compounds of life. Other proteins provide structural elements in the body, act as transport channels across plasma membranes, function as signals for changing activities, and provide chemical weapons against disease-carrying bacteria. These diverse proteins are built from a small number (20) of essential amino acids.

Nucleotides and nucleic acids make up the last category of important biological molecules. Nucleotides are small organic compounds that contain a five-carbon sugar (ribose or deoxyribose), a phosphate group, and a nitrogen-containing base that has a single or double carbon ring structure. Adenosine triphosphate (ATP) is the energy currency of the cell and plays a central role in metabolism. Other nucleotides are subunits of coenzymes that are enzyme helpers. The two nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA (Figure 3.3) is a unique, helical molecule that contains chains of paired nucleotides that run in opposite directions. Each nucleotide contains either a pyrimidine base—thymine (T) or cytosine (C)—with a single ring structure or a purine base—adenine (A) or guanine (G)—with a double ring. In the double helix of DNA, thymine always pairs with adenine (T-A) and cytosine always pairs with guanine (C-G). RNA is similar to DNA except that it consists of a single helical strand, contains ribose instead of deoxyribose, and has uracil (U) instead of thymine.

All cells are surrounded by a plasma membrane that separates, but does not isolate, the cell's interior from its environment. Animal cells, such as those found in humans, are eukaryotic cells. A generalized animal cell is shown in Figure 3.4. In addition to the plasma membrane, eukaryotic cells contain membrane-bound organelles and a membrane-bound nucleus. Prokaryotic cells, such as bacteria, lack membrane-bound structures other than the plasma membrane. In addition to a plasma membrane, all cells have a region that contains DNA (which carries the hereditary instructions for the cell) and cytoplasm (which is a semifluid substance that includes everything inside the plasma membrane except for the DNA).

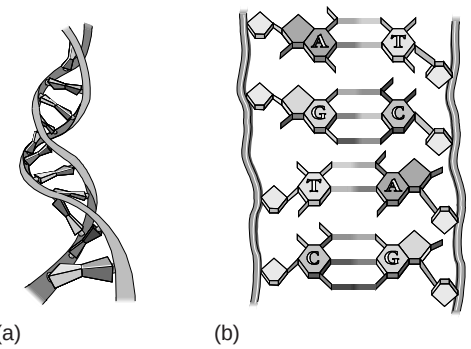


FIGURE 3.3 (a) DNA consists of two chains of paired nucleotides that run in opposite directions and form a helical structure. (b) Thymine pairs with adenine (T-A) and cytosine pairs with guanine (C-G) due to hydrogen bonding between the bases.

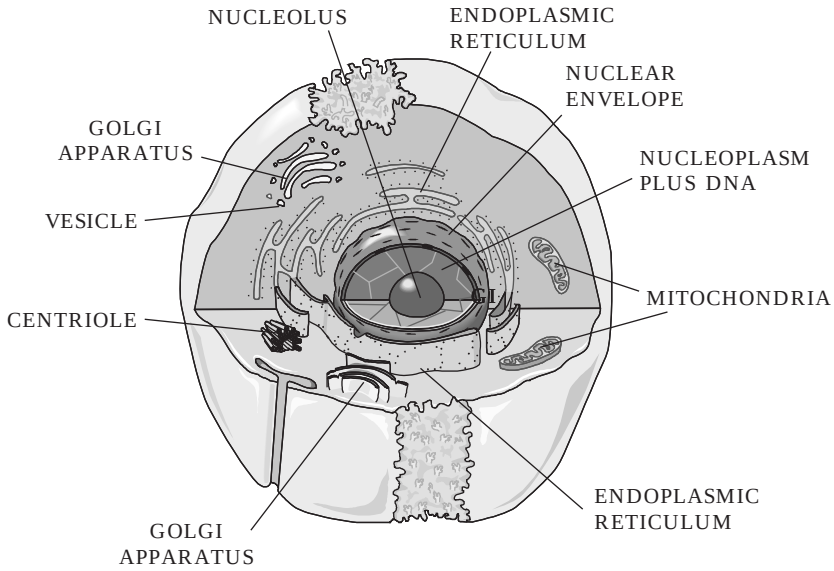


FIGURE 3.4 Animal cells are surrounded by a plasma membrane. They contain a membrane-bound region, the nucleus, which contains DNA. The cytoplasm lies outside of the nucleus and contains several types of organelles that perform specialized functions.

3.2.1 Plasma Membrane

The plasma membrane performs several functions for the cell. It gives mechanical strength, provides structure, helps with movement, and controls the cell's volume and its activities by regulating the movement of chemicals in and out of the cell. The plasma membrane is composed of two layers of phospholipids interspersed with proteins and cholesterol (Figure 3.5). The proteins in the plasma membranes of mammalian cells provide

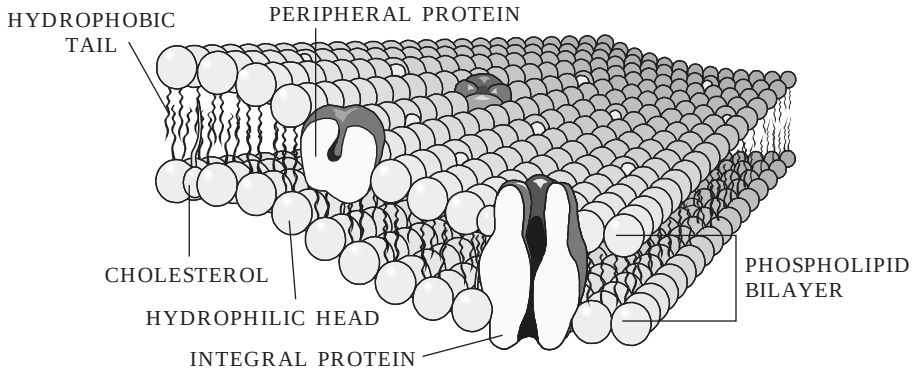


FIGURE 3.5 The plasma membrane surrounds all cells. It consists of a double layer of phospholipids interspersed with proteins and cholesterol.

binding sites for hormones, recognition markers for identifying cells as one type or another, adhesive mechanisms for binding adjacent cells to each other, and channels for transporting materials across the plasma membrane. The phospholipids are arranged with their "water loving" (hydrophilic) heads pointing outward and their "water fearing" (hydrophobic) tails pointing inward. This double-layer arrangement of phospholipids interspersed with protein channels helps maintain the internal environment of a cell by controlling the substances that move across the membrane, whereas the cholesterol molecules act as stabilizers to prevent extensive lateral movement of the lipid molecules.

Some molecules, such as oxygen, carbon dioxide, and water, can easily cross the plasma membrane, whereas other substances, such as large molecules and ions, must move through the protein channels. *Osmosis* is the process by which substances move across a selectively permeable membrane such as a cell's plasma membrane, whereas *diffusion* refers to the movement of molecules from an area of relatively high concentration to an area of relatively low concentration. Substances that can easily cross the plasma membrane achieve diffusion equilibrium when there is no net movement of these substances across the membrane; that is, the concentration of the substance inside the cell equals the concentration of the substance outside of the cell. Active transport, which requires an input of energy usually in the form of ATP, can be used to move ions and molecules across the plasma membrane and is often used to move them from areas of low concentration to areas of high concentration. This mechanism helps maintain concentrations of ions and molecules inside a cell that are different from the concentrations outside the cell. A typical mammalian cell has internal sodium ion (Na^+) concentrations of 12 mM (12 moles of Na^+ per 1,000 liters of solution) and extracellular Na^+ concentrations of 120 mM, whereas intracellular and extracellular potassium ion (K^+) concentrations are on the order of 125 mM and 5 mM, respectively. In addition to positively charged ions (cations), cells also contain negatively charged ions (anions). A typical mammalian cell has intracellular and extracellular chloride ion (Cl^-) concentrations of 5 mM and 125 mM and internal anion (e.g., proteins, charged amino acids, sulfate ions, and phosphate ions) concentrations of 108 mM. These transmembrane ion gradients are used to make ATP, to drive various transport processes, and to generate electrical signals.

EXAMPLE PROBLEM 3.1

How many molecules of sodium and potassium ions would a cell that has a volume of 2 nl contain?

Solution

Assuming that the intracellular concentrations of Na^+ and K^+ are 12 mM and 125 mM, respectively, the number of molecules for each can be determined by using the volume of the cell and Avogadro's number.

$$Na^+: \ 12\frac{moles}{1,000\ liters} \times 6.023 \times 10^{23}\frac{molecules}{mole} \times 2 \times 10^{-9}\ liters = 1.45 \times 10^{13}\ molecules$$

$$K^+: \ 125\frac{moles}{1,000\ liters} \times 6.023 \times 10^{23}\frac{molecules}{mole} \times 2 \times 10^{-9}\ liters = 1.51 \times 10^{14}\ molecules$$

The plasma membrane plays an important role in regulating cell volume by controlling the internal osmolarity of the cell. Osmolarity is defined in terms of concentration of dissolved substances. A 1 osmolar (1 Osm) solution contains 1 mole of dissolved particles per liter of solution, while a 1 milliosmolar (1 mOsm) solution has 1 mole of dissolved particles per 1,000 liters of solution. Thus, solutions with high osmolarity have low concentrations of water or other solvents. For biological purposes, solutions with 0.1 Osm glucose and 0.1 Osm urea have essentially the same concentrations of water. It is important to note that a 0.1 M solution of sodium chloride (NaCl) will form a 0.2 Osm solution, since NaCl dissociates into Na^+ and Cl^- ions and thus has twice as many dissolved particles as a solution of a substance—for example, glucose—that does not dissociate into smaller units. Two solutions are isotonic if they have the same osmolarity. One solution is hypotonic to another if it has a lower osmolarity and hypertonic to another if it has a higher osmolarity. It is important to note that tonicity (isotonic, hypotonic, or hypertonic) is only determined by those molecules that cannot cross the plasma membrane, since molecules that can freely cross will eventually reach equilibrium with the same concentration inside and outside of the cell.

Consider a simple model cell that consists of a plasma membrane and cytoplasm. The cytoplasm in this model cell contains proteins that cannot cross the plasma membrane and water that can. At equilibrium, the total osmolarity inside the cell must equal the total osmolarity outside the cell. If the osmolarity inside and the osmolarity outside of the cell are out of balance, there will be a net movement of water from the side of the plasma membrane where it is more highly concentrated to the other until equilibrium is achieved. For example, assume that a model cell (Figure 3.6) contains 0.2 M protein and is placed in a hypotonic solution that contains 0.1 M sucrose. The plasma membrane of this model cell is impermeable to proteins and sucrose but freely permeable to water. The volume of the

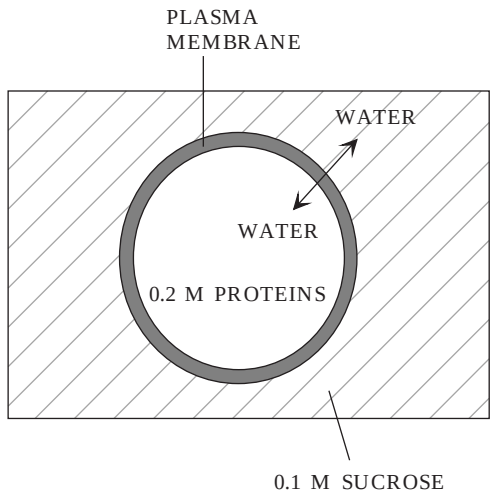


FIGURE 3.6 A simple model cell that consists of cytoplasm, containing 0.2 M proteins, and a plasma membrane is placed in a solution of 0.1 M sucrose. The plasma membrane is insoluble to proteins and sucrose but allows water to pass freely in either direction. The full extent of the extracellular volume is not shown and is much larger than the cell's volume of 1 nl.

cell, 1 nl, is very small relative to the volume of the solution. In other words, changes in the cell's volume have no measurable effect on the volume of the external solution. What will happen to the volume of the cell as it achieves equilibrium?

At equilibrium, the osmolarity inside the cell must equal the osmolarity outside the cell. The initial osmolarity inside the cell is 0.2 Osm, since the proteins do not dissociate into smaller units. The osmolarity outside the cell is 0.1 Osm due to the sucrose solution. A 0.2 Osm solution has 0.2 moles of dissolved particles per liter of solution, while a 0.1 Osm solution has half as many moles of dissolved particles per liter. The osmolarity inside the cell must decrease by a factor of 2 in order to achieve equilibrium. Since the plasma membrane will not allow any of the protein molecules to leave the cell, this can only be achieved by doubling the cell's volume. Thus, there will be a net movement of water across the plasma membrane until the cell's volume increases to 2 nl and the cell's internal osmolarity is reduced to 0.1 Osm—the same as the osmolarity of the external solution. The water moves down its concentration gradient by diffusing from where it is more highly concentrated in the 0.1 M sucrose solution to where it is less concentrated in the 0.2 M protein solution in the cell.

EXAMPLE PROBLEM 3.2

What would happen to the model cell in Figure 3.6 if it were placed in pure water?

Solution

Water can pass through the plasma membrane and would flow down its concentration gradient from where it is more concentrated (outside of the cell) to where it is less concentrated (inside of the cell). Eventually, enough water would move into the cell to rupture the plasma membrane, since the concentration of water outside the cell would be higher than the concentration of water inside the cell as long as there were proteins trapped within the cell.

EXAMPLE PROBLEM 3.3

Assume that the model cell in Figure 3.6 has an initial volume of 2 nl and contains 0.2 M protein. The cell is placed in a large volume of 0.2 M NaCl. In this model, neither Na^+ nor Cl^- can cross the plasma membrane and enter the cell. Is the 0.2 M NaCl solution hypotonic, isotonic, or hypertonic relative to the osmolarity inside the cell? Describe what happens to the cell as it achieves equilibrium in this new environment. What will be the final osmolarity of the cell? What will be its final volume?

Solution

The osmolarity inside the cell is 0.2 Osm. The osmolarity of the 0.2 M NaCl solution is 0.4 Osm (0.2 Osm Na^+ + 0.2 Osm Cl^-). Thus, the NaCl solution is hypertonic relative to the osmolarity inside the cell ($\text{osmolarity}_{\text{outside}} > \text{osmolarity}_{\text{inside}}$). Since none of the particles (protein, Na^+, and Cl^-) can cross the membrane, water will move out of the cell until the osmolarity inside

Continued

the cell is 0.4 Osm. This will be achieved when the volume inside the cell has been reduced from 2 nl to 1 nl.

$$C_1V_1 = C_2V_2$$
$$\frac{0.2\ Osm}{0.4\ Osm} \times 2nl = V_2$$
$$1nl = V_2$$

Real cells are much more complex than the simple model just described. In addition to achieving osmotic balance at equilibrium, real cells must also achieve electrical balance with regard to the ions that are present in the cytoplasm. The principle of electrical neutrality requires that the overall concentration of cations in a biological compartment—for example, a cell—must equal the overall concentration of anions in that compartment. Consider another model cell (Figure 3.7) with internal and external cation and anion concentrations similar to those of a typical mammalian cell. Is the cell at equilibrium if the plasma membrane is freely permeable to K^+ and Cl^- but impermeable to Na^+ and the internal anions? The total osmolarity inside the cell is 250 mOsm (12 mM Na^+, 125 mM K^+, 5 mM Cl^-, 108 mM anions), while the total osmolarity outside the cell is also 250 mOsm (120 mM Na^+, 5 mM K^+, 125 mM Cl^-), so the cell is in osmotic balance—that is, there will be no net movement of water across the plasma membrane. If the average charge per molecule of the anions inside the cell is considered to be –1.2, then the cell is also approximately in electrical equilibrium (12 + 125 positive charges for Na^+ and K^+; 5 + 1.2 * 108 negative charges for Cl^- and the other anions). Real cells, however, cannot maintain this equilibrium without expending energy, since real cells are slightly permeable to Na^+. In order to maintain equilibrium and keep Na^+ from accumulating intracellularly, mammalian cells must actively pump Na^+ out of the cell against its diffusion and electrical gradients. Since Na^+

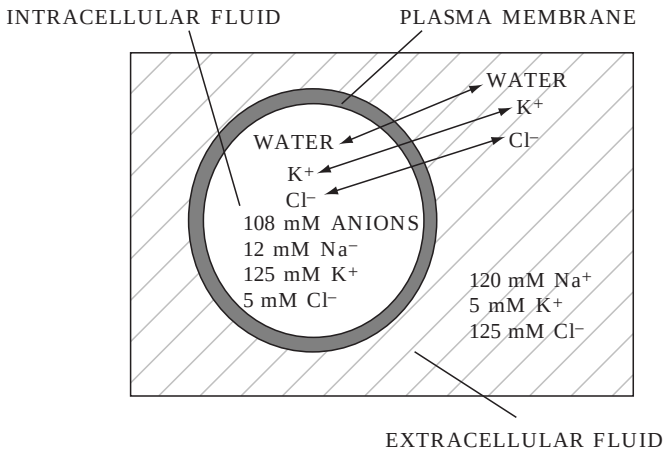


FIGURE 3.7 A model cell with internal and external concentrations similar to those of a typical mammalian cell. The full extent of the extracellular volume is not shown and is much larger than the cell's volume.

is pumped out through specialized protein channels at a rate equivalent to the rate at which it leaks in through other channels, it behaves osmotically as if it cannot cross the plasma membrane. Thus, mammalian cells exist in a steady state, rather than at equilibrium, since energy in the form of ATP must be used to prevent a net movement of ions across the plasma membrane.

EXAMPLE PROBLEM 3.4

Consider a simple model cell, such as the one in Figure 3.7, that has the following ion concentrations. Is the cell at equilibrium? Explain your answer.

Ion	Intracellular Concentration (mM)	Extracellular Concentration (mM)
K^+	158	4
Na^+	20	163
Cl^-	52	167
A^-	104	—

Solution

Yes. The cell is both electrically and osmotically at equilibrium because the charges within the inside and outside compartments are equal, and the osmolarity inside the cell equals the osmolarity outside of the cell.

	Inside	Outside
Positive	$158 + 20 = 178$ mM	$4 + 163 = 167$ mM
Negative	$52 + 1.2 * 104 = 177$ mM	167 mM
	$178\ mM_{pos} \approx 177\ mM_{neg}$	$167\ mM_{pos} = 167\ mM_{neg}$
Osmolarity	$158 + 20 + 52 + 104 = 334$ mM	$4 + 163 + 167 = 334$ mM
	$334\ mM_{inside} = 334\ mM_{outside}$	

One of the consequences of the distribution of charged particles in the intracellular and extracellular fluids is that an electrical potential exists across the plasma membrane. The value of this electrical potential depends on the intracellular and extracellular concentrations of ions that can cross the membrane and will be described more fully in Chapter 11.

In addition to controlling the cell's volume, the plasma membrane also provides a route for moving large molecules and other materials into and out of the cell. Substances can be moved into the cell by means of endocytosis (Figure 3.8a) and out of the cell by means of exocytosis (Figure 3.8b). In endocytosis, material—for example, a bacterium—outside of the cell is engulfed by a portion of the plasma membrane that encircles it to form a vesicle. The vesicle then pinches off from the plasma membrane and moves its contents to the inside of the cell. In exocytosis, material within the cell is surrounded by a membrane to form a vesicle. The vesicle then moves to the edge of the cell, where its membrane fuses with the plasma membrane and its contents are released to the exterior of the cell.

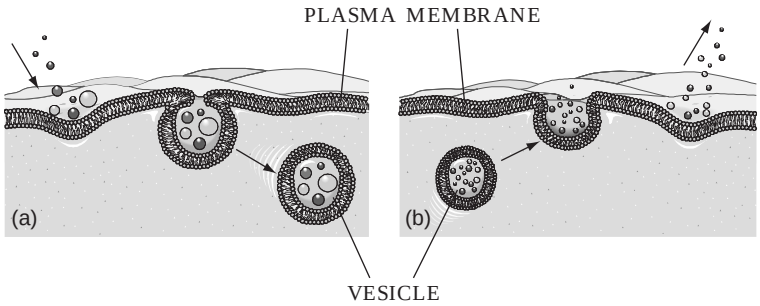


FIGURE 3.8 Substances that are too large to pass through the integral proteins in the plasma membrane can be moved into the cell by means of endocytosis (a) and out of the cell by means of exocytosis (b).

3.2.2 Cytoplasm and Organelles

The cytoplasm contains fluid (cytosol) and organelles. Ions (such as Na^+, K^+, and Cl^-) and molecules (such as glucose) are distributed through the cytosol via diffusion. Membrane-bound organelles include the nucleus, rough and smooth endoplasmic reticulum, the Golgi apparatus, lysosomes, and mitochondria. Nonmembranous organelles include nucleoli, ribosomes, centrioles, microvilli, cilia, flagella, and the microtubules, intermediate filaments, and microfilaments of the cytoskeleton.

The nucleus (see Figure 3.4) consists of the nuclear envelope (a double membrane) and the nucleoplasm (a fluid that contains ions, enzymes, nucleotides, proteins, DNA, and small amounts of RNA). Within its DNA, the nucleus contains the instructions for life's processes. Nuclear pores are protein channels that act as connections for ions and RNA, but not proteins or DNA, to leave the nucleus and enter the cytoplasm and for some proteins to enter the nucleoplasm. Most nuclei contain one or more nucleoli. Each nucleolus contains DNA, RNA, and proteins and synthesizes the components of the ribosomes that cells use to make proteins.

The smooth and rough endoplasmic reticulum (ER), Golgi apparatus, and assorted vesicles (Figures 3.4, 3.9a, and 3.9b) make up the cytomembrane system, which delivers proteins and lipids for manufacturing membranes and accumulates and stores proteins and lipids for specific uses. The ER also acts as a storage site for calcium ions. The rough ER differs from the smooth ER in that it has ribosomes attached to its exterior surface. Ribosomes provide the platforms for synthesizing proteins. Those that are synthesized on the rough ER are passed into its interior, where nonproteinaceous side chains are attached to them. These modified proteins move to the smooth ER, where they are packaged in vesicles. The smooth ER also manufactures and packages lipids into vesicles and is responsible for releasing stored calcium ions. The vesicles leave the smooth ER and become attached to the Golgi apparatus, where their contents are released, modified, and repackaged into new vesicles. Some of these vesicles, called lysosomes, contain digestive enzymes that are used to break down materials that move into the cells via endocytosis. Other vesicles contain proteins, such as hormones and neurotransmitters, that are secreted from the cells by means of exocytosis.

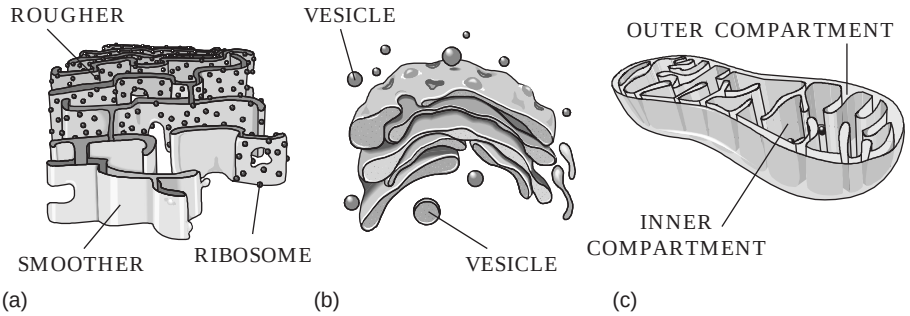


FIGURE 3.9 Subcellular organelles. The endoplasmic reticulum (a), the Golgi apparatus (b), and vesicles (b) make up the cytomembrane system in the cell. The small circles on the endoplasmic reticulum (ER) represent ribosomes. The area containing ribosomes is called the rough ER, while the area that lacks ribosomes is called the smooth ER. The mitochondria (c) have a double membrane system that divides the interior into two compartments that contain different concentrations of enzymes, substrates, and hydrogen ions (H^+). Electrical and chemical gradients between the inner and outer compartments provide the energy needed to generate ATP.

The mitochondria (Figures 3.9c and 3.10) contain two membranes: an outer membrane that surrounds the organelle and an inner membrane that divides the organelle's interior into two compartments. Approximately 95 percent of the ATP required by the cell is produced in the mitochondria in a series of oxygen-requiring reactions that produce carbon dioxide as a by-product. Mitochondria are different from most other organelles in that they contain their own DNA. The majority of the mitochondria in sexually reproducing organisms, such

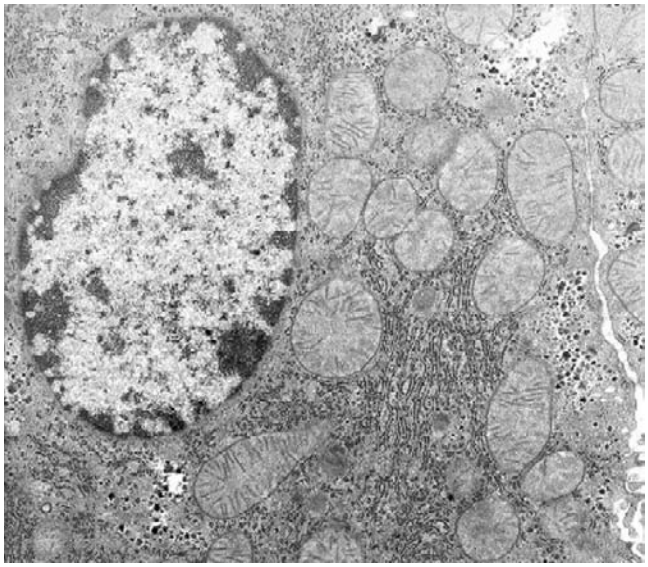


FIGURE 3.10 Scanning electron micrograph of a normal mouse liver at 8,000X magnification. The large round organelle on the left is the nucleus. The smaller round and oblong organelles are mitochondria that have been sliced at different angles. The narrow membranes in parallel rows are endoplasmic reticula. The small black dots on the ERs are ribosomes. *Photo courtesy of Valerie Knowlton, Center for Electron Microscopy, North Carolina State University.*

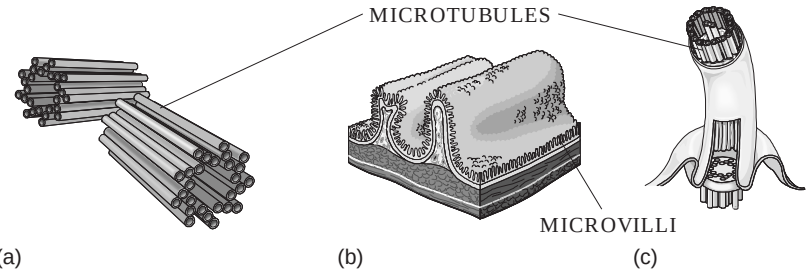


FIGURE 3.11 Centrioles (a) contain microtubules and are located at right angles to each other in the cell's centrosome. These organelles play an important part in cell division by anchoring the microtubules that are used to divide the cell's genetic material. Microvilli (b), which are extensions of the plasma membrane, line the villi, tiny fingerlike protrusions in the mucosa of the small intestine, and help increase the area available for the absorption of nutrients. Cilia (c) line the respiratory tract. The beating of these organelles helps move bacteria and particles trapped in mucus out of the lungs.

as humans, come from the mother's egg cell, since the father's sperm contributes little more than the DNA in a haploid (half) set of chromosomes to the developing offspring.

Microtubules, intermediate filaments, and microfilaments provide structural support and assist with movement. Microtubules are long, hollow, cylindrical structures that radiate from microtubule organizing centers and, during cell division, from centrosomes, a specialized region of the cytoplasm that is located near the nucleus and contains two centrioles (Figures 3.4 and 3.11a) oriented at right angles to each other. Microtubules consist of spiraling subunits of a protein called tubulin, whereas centrioles consist of nine triplet microtubules that radiate from their centers like the spokes of a wheel. Intermediate filaments are hollow and provide structure to the plasma membrane and nuclear envelope. They also aid in cell-to-cell junctions and in maintaining the spatial organization of organelles. Myofilaments are found in most cells and are composed of strings of protein molecules. Cell movement can occur when actin and myosin, protein subunits of myofilaments, interact. Microvilli (Figure 3.11b) are extensions of the plasma membrane that contain microfilaments. They increase the surface area of a cell to facilitate absorption of extracellular materials.

Cilia (Figure 3.11c) and flagella are parts of the cytoskeleton that have shafts composed of nine pairs of outer microtubules and two single microtubules in the center. Both types of shafts are anchored by a basal body that has the same structure as a centriole. Flagella function as whiplike tails that propel cells such as sperm. Cilia are generally shorter and more profuse than flagella and can be found on specialized cells such as those that line the respiratory tract. The beating of the cilia helps move mucus-trapped bacteria and particles out of the lungs.

3.2.3 DNA and Gene Expression

DNA (see Figure 3.3) is found in the nucleus and mitochondria of eukaryotic cells. In organisms that reproduce sexually, the DNA in the nucleus contains information from both parents, while that in the mitochondria comes from the organism's mother. In the nucleus, the DNA is wrapped around protein spools, called nucleosomes, and is organized into pairs of chromosomes. Humans have 22 pairs of autosomal chromosomes and two sex

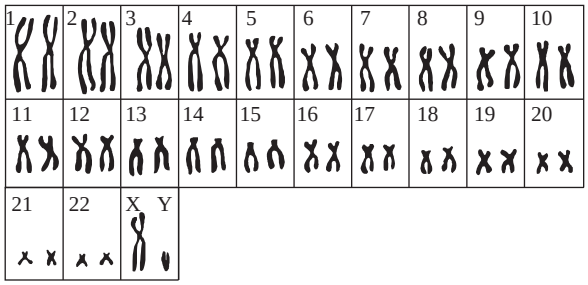


FIGURE 3.12 This karyotype of a normal human male shows the 22 pairs of autosomal chromosomes in descending order based on size, as well as the X and Y sex chromosomes.

chromosomes, XX for females and XY for males (Figure 3.12). If the DNA from all 46 chromosomes in a human somatic cell—that is, any cell that does not become an egg or sperm cell—was stretched out end to end, it would be about 2 nm wide and 2 m long. Each chromosome contains thousands of individual genes that are the units of information about heritable traits. Each gene has a particular location in a specific chromosome and contains the code for producing one of the three forms of RNA (ribosomal RNA, messenger RNA, and transfer RNA). The Human Genome Project was begun in 1990 and had as its goal to first identify the location of at least 3,000 specific human genes and then to determine the sequence of nucleotides (about 3 billion!) in a complete set of haploid human chromosomes (one chromosome from each of the 23 pairs). See Chapter 13 for more information about the Human Genome Project.

DNA replication occurs during cell division (Figure 3.13). During this semiconservative process, enzymes unzip the double helix, deliver complementary bases to the nucleotides, and bind the delivered nucleotides into the developing complementary strands. Following replication, each strand of DNA is duplicated so two double helices now exist, each consisting of one strand of the original DNA and one new strand. In this way, each daughter cell gets the same hereditary information that was contained in the original dividing cell. During replication, some enzymes check for accuracy, while others repair pairing mistakes so the error rate is reduced to approximately one per billion.

Since DNA remains in the nucleus, where it is protected from the action of the cell's enzymes, and proteins are made on ribosomes outside of the nucleus, a method (transcription) exists for transferring information from the DNA to the cytoplasm. During transcription (Figure 3.14), the sequence of nucleotides in a gene that codes for a protein is transferred to messenger RNA (mRNA) through complementary base pairing of the nucleotide sequence in the gene. For example, a DNA sequence of TACGCTCCGATA would become AUGCGAGGCUAU in the mRNA. The process is somewhat more complicated, since the transcript produced directly from the DNA contains sequences of nucleotides, called introns, that are removed before the final mRNA is produced. The mRNA also has a tail, called a poly-A tail, of about 100–200 adenine nucleotides attached to one end. A cap with a nucleotide that has a methyl group and phosphate groups bonded to it is attached at the other end of the mRNA. Transcription differs from replication in that (1) only a certain stretch of DNA acts as the template and not the whole strand, (2) different enzymes are used, and (3) only a single strand is produced.

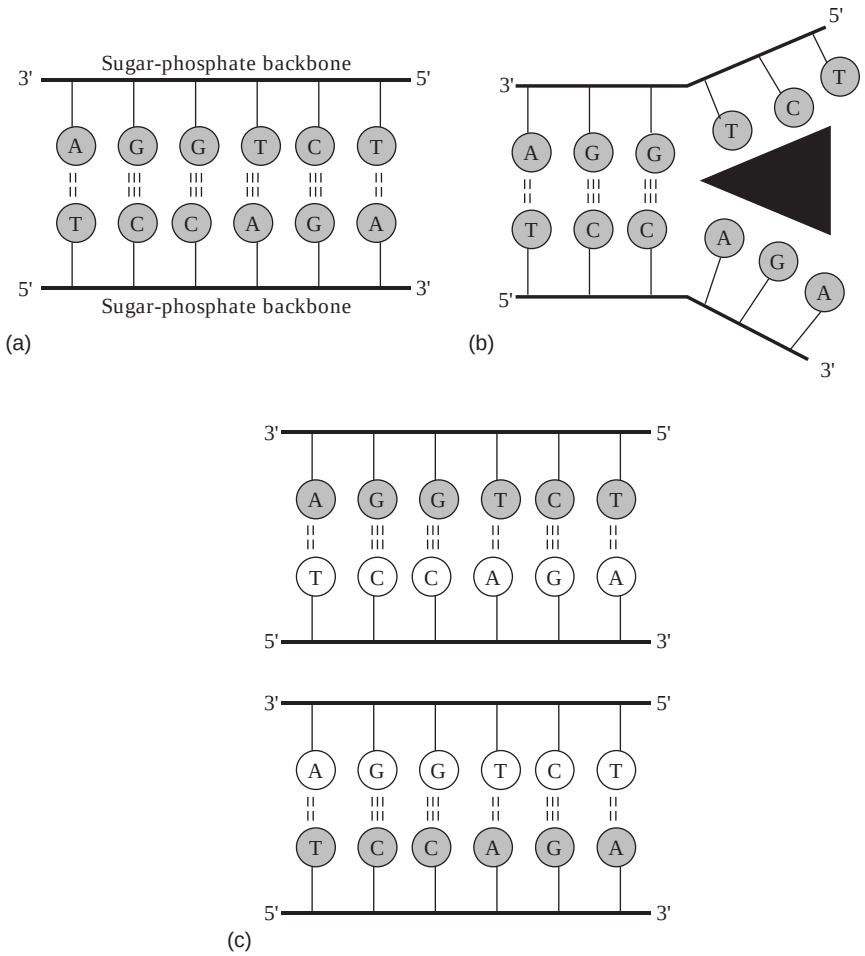


FIGURE 3.13 During replication, DNA helicase shown as a black wedge in (b) unzips the double helix (a). Another enzyme, DNA polymerase, then copies each side of the unzipped chain in the 5′ to 3′ direction. One side of the chain (5′ to 3′) can be copied continuously, while the opposite side (3′ to 5′) is copied in small chunks in the 5′ to 3′ direction that are bound together by another enzyme, DNA ligase. Two identical double strands of DNA are produced as a result of replication.

After being transcribed, the mRNA moves out into the cytoplasm through the nuclear pores and binds to specific sites on the surface of the two subunits that make up a ribosome (Figure 3.15). In addition to the ribosomes, the cytoplasm contains amino acids and another form of RNA: transfer RNA (tRNA). Each tRNA contains a triplet of bases, called an anticodon, and binds at an area away from the triplet to an amino acid that is specific for that particular anticodon. The mRNA that was produced from the gene in the nucleus also contains bases in sets of three. Each triplet in the mRNA is called a codon. The four

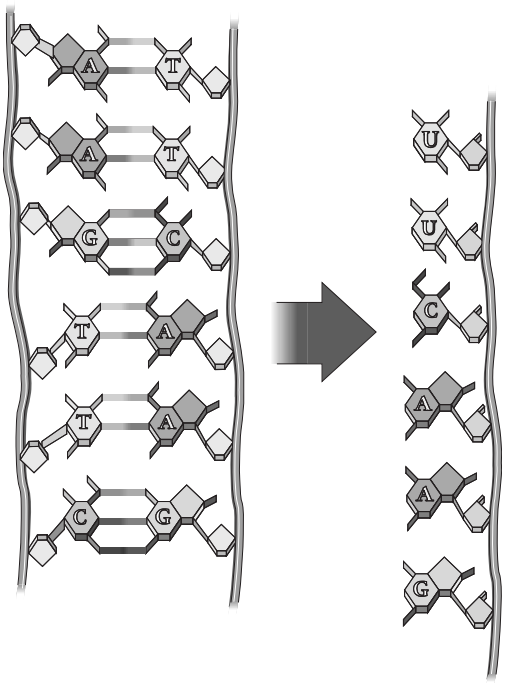


FIGURE 3.14 During transcription, RNA is formed from genes in the cell's DNA by complementary base pairing to one of the strands. RNA contains uracil (U) rather than thymine (T), so the Ts in the first two pairs of the DNA become Us in the single-stranded RNA.

possibilities for nucleotides (A, U, C, G) in each of the three places give rise to 64 (4^3) possible codons. These 64 codons make up the genetic code. Each codon codes for a specific amino acid, but some amino acids are specified by more than one codon (Table 3.1). For example, AUG is the only mRNA codon for methionine (the amino acid that always signals the starting place for translation—the process by which the information from a gene is used to produce a protein), while UUA, UUG, CUU, CUC, CUA, and CUG are all codons for leucine. The anticodon on the tRNA that delivers the methionine to the ribosome is UAC, whereas tRNAs with anticodons of AAU, AAC, GAA, GAG, GAU, and GAC deliver leucine.

During translation, the mRNA binds to a ribosome and tRNA delivers amino acids to the growing polypeptide chain in accordance with the codons specified by the mRNA. Peptide bonds are formed between each newly delivered amino acid and the previously delivered one. When the amino acid is bound to the growing chain, it is released from the tRNA, and the tRNA moves off into the cytoplasm, where it joins with another amino acid that is specified by its anticodon. This process continues until a stop codon (UAA, UAG, or UGA) is reached on the mRNA. The protein is then released into the cytoplasm or into the rough ER for further modifications.

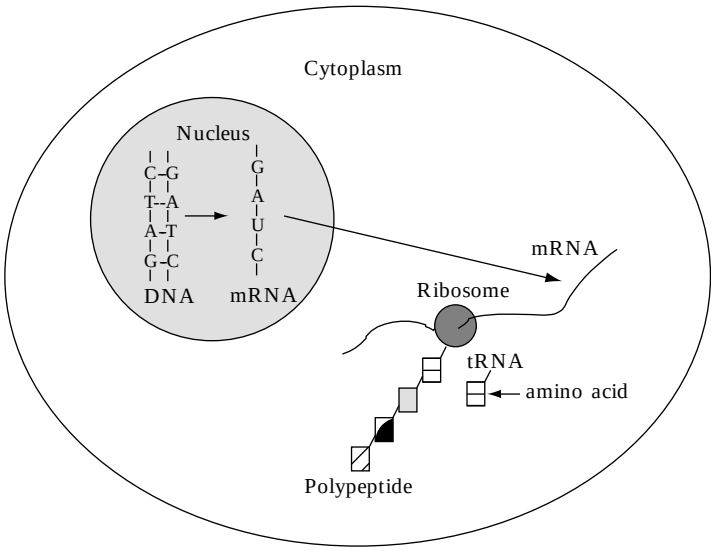


FIGURE 3.15 Following transcription from DNA and processing in the nucleus, mRNA moves from the nucleus to the cytoplasm. In the cytoplasm, the mRNA joins with a ribosome to begin the process of translation. During translation, tRNA delivers amino acids to the growing polypeptide chain. Which amino acid is delivered depends on the three-base codon specified by the mRNA. Each codon is complementary to the anticodon of a specific tRNA. Each tRNA binds to a particular amino acid at a site that is opposite the location of the anticodon. For example, the codon CUG in mRNA is complementary to the anticodon GAC in the tRNA that carries leucine and will result in adding the amino acid leucine to the polypeptide chain.

EXAMPLE PROBLEM 3.5

Consider a protein that contains the amino acids asparagine, phenylalanine, histidine, and serine in sequence. Which nucleotide sequences on DNA (assuming that there were no introns) would result in this series of amino acids? What would be the anticodons for the tRNAs that delivered these amino acids to the ribosomes during translation?

Solution

The genetic code (see Table 3.1) provides the sequence for the mRNA codons that specify these amino acids. The mRNA codons can be used to determine the sequence in the original DNA and the anticodons of the tRNA, since the mRNA bases must pair with the bases in both DNA and tRNA. Note that DNA contains thymine (T) but no uracil (U) and that both mRNA and tRNA contain U and not T. See Figures 3.3 and 3.14 for examples of base pairing.

	Asparagine (Asn)	**Phenylalanine (Phe)**	**Histidine (His)**	**Serine (Ser)**
mRNA codon	AAU or AAC	UUU or UUC	CAU or CAC	UC(A, G, U, or C)
DNA	TTA or TTG	AAA or AAG	GTA or GTG	AG(T, C, A, or G)
tRNA anticodon	UUA or UUG	AAA or AAG	GUA or GUG	AG(U, C, A, or G)

TABLE 3.1 The Genetic Code

First Base	Second Base				Third Base
	A	*U*	*G*	*C*	
A	Lys	Ile	Arg	Thr	**A**
	Lys	**Met - Start**	Arg	Thr	**G**
	Asn	Ile	Ser	Thr	**U**
	Asn	Ile	Ser	Thr	**C**
U	Stop	Leu	Stop	Ser	A
	Stop	Leu	Trp	Ser	G
	Tyr	Phe	Cys	Ser	U
	Tyr	Phe	Cys	Ser	C
G	Glu	Val	Gly	Ala	**A**
	Glu	Val	Gly	Ala	**G**
	Asp	Val	Gly	Ala	**U**
	Asp	Val	Gly	Ala	**C**
C	Gln	Leu	Arg	Pro	**A**
	Gln	Leu	Arg	Pro	**G**
	His	Leu	Arg	Pro	**U**
	His	Leu	Arg	Pro	**C**

Amino acid 3-letter and 1-letter codes: Ala (A) = Alanine; Arg (R) = Arginine; Asn (N) = Asparagine; Asp (D) = Aspartic acid; Cys (C) = Cysteine; Glu (E) = Glutamic acid; Gln (Q) = Glutamine; Gly (G) = Glycine; His (H) = Histidine; Ile (I) = Isoleucine; Leu (L) = Leucine; Lys (K) = Lysine; Met (M) = Methionine; Phe (F) = Phenylalanine; Pro (P) = Proline; Ser (S) = Serine; Thr (T) = Threonine; Trp (W) = Tryptophan; Tyr (Y) = Tyrosine; Val (V) = Valine.

3.3 TISSUES

Groups of cells and surrounding substances that function together to perform one or more specialized activities are called tissues (Figure 3.16). The four primary types of tissue in the human body are epithelial, connective, muscle, and nervous. Epithelial tissues are either composed of cells arranged in sheets that are one or more layers thick or are organized into glands that are adapted for secretion. They are also characterized by having a free surface—for example, the inside surface of the intestines or the outside of the skin—and a basilar membrane. Typical functions of epithelial tissue include absorption (lining of the small intestine), secretion (glands), transport (kidney tubules), excretion (sweat glands), protection (skin, Figure 3.16a), and sensory reception (taste buds). Connective tissues are the most abundant and widely distributed. Connective tissue proper can be loose (loosely woven fibers found around and between organs), irregularly dense (protective capsules around organs), and regularly dense (ligaments and tendons), whereas specialized connective tissue includes blood

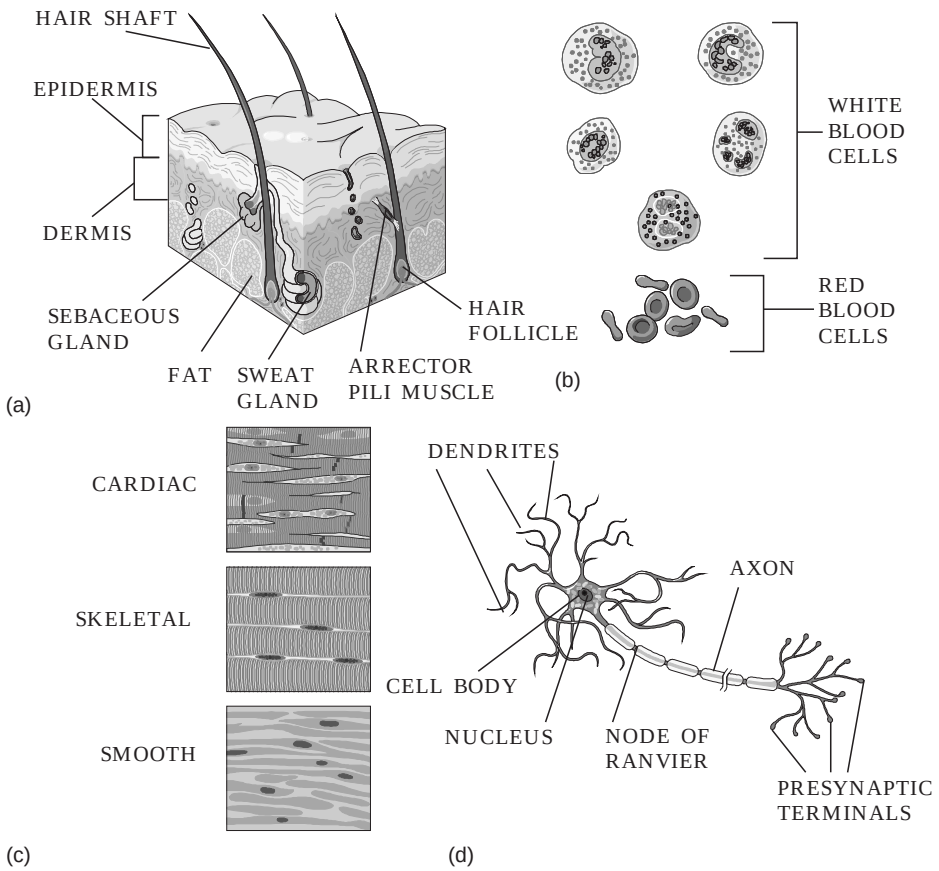


FIGURE 3.16 Four tissue types. Skin (a) is a type of epithelial tissue that helps protect the body. Blood (b) is a specialized connective tissue. The three types of muscle tissue (c) are cardiac, skeletal, and smooth. Motor neurons (d) are a type of nervous tissue that conducts electrical impulses from the central nervous system to effector organs such as muscles.

(Figure 3.16b), bone, cartilage, and adipose tissue. Muscle tissue provides movement for the body through its specialized cells that can shorten in response to stimulation and then return to their uncontracted state. Figure 3.16c shows the three types of muscle tissue: skeletal (attached to bones), smooth (found in the walls of blood vessels), and cardiac (found only in the heart). Nervous tissue consists of neurons (Figure 3.16d) that conduct electrical impulses and glial cells that protect, support, and nourish neurons.

3.4 MAJOR ORGAN SYSTEMS

Combinations of tissues that perform complex tasks are called organs, and organs that function together form organ systems. The human body has 11 major organ systems: integumentary, endocrine, lymphatic, digestive, urinary, reproductive, circulatory, respiratory, nervous, skeletal, and muscular. The integumentary system (skin, hair, nails, and various

glands) provides protection for the body. The endocrine system (ductless glands such as the thyroid and adrenals) secretes hormones that regulate many chemical actions within cells. The lymphatic system (glands, lymph nodes, lymph, lymphatic vessels) returns excess fluid and protein to the blood and helps defend the body against infection and tissue damage. The digestive system (stomach, intestines, and other structures) ingests food and water, breaks food down into small molecules that can be absorbed and used by cells, and removes solid wastes. The urinary system (kidneys, ureters, urinary bladder, and urethra) maintains the fluid volume of the body, eliminates metabolic wastes, and helps regulate blood pressure and acid-base and water-salt balances. The reproductive system (ovaries, testes, reproductive cells, and accessory glands and ducts) produces eggs or sperm and provides a mechanism for the production and nourishment of offspring. The circulatory system (heart, blood, and blood vessels) serves as a distribution system for the body. The respiratory system (airways and lungs) delivers oxygen to the blood from the air and carries away carbon dioxide. The nervous system (brain, spinal cord, peripheral nerves, and sensory organs) regulates most of the body's activities by detecting and responding to internal and external stimuli. The skeletal system (bones and cartilage) provides protection and support as well as sites for muscle attachments, the production of blood cells, and calcium and phosphorus storage. The muscular system (skeletal muscle) moves the body and its internal parts, maintains posture, and produces heat. Although biomedical engineers have made major contributions to understanding, maintaining, and/or replacing components in each of the 11 major organ systems, only the last 5 listed will be examined in greater detail.

3.4.1 Circulatory System

The circulatory system (Figure 3.17) delivers nutrients and hormones throughout the body, removes waste products from tissues, and provides a mechanism for regulating temperature and removing the heat generated by the metabolic activities of the body's internal organs. Every living cell in the body is no more than 10–100 μm from a capillary (small blood vessels with walls only one cell thick that are 8 μm in diameter, approximately the same size as a red blood cell). This close proximity allows oxygen, carbon dioxide, and most other small solutes to diffuse from the cells into the capillary or from the capillary into the cells, with the direction of diffusion determined by concentration and partial pressure gradients.

Accounting for about 8 +/− 1 percent of total body weight, averaging 5,200 ml, blood is a complex, heterogeneous suspension of formed elements—the *blood cells*, or *hematocytes*—suspended in a continuous, straw-colored fluid called *plasma*. Nominally, the composite fluid has a mass density of 1.057 +/− 0.007 g/cm^3, and it is six times as viscous as water. The hematocutes include three basic types of cells: red blood cells (erythrocytes, totaling nearly 95 percent of the formed elements), white blood cells (leukocytes, averaging less than .15 percent of all hematocytes), and platelets (thrombocytes, on the order of 5 percent of all blood cells). Hematocytes are all derived in the active ("red") bone marrow (about 1,500 g) of adults from undifferentiated stem cells called *hemocytoblasts*, and all reach ultimate maturity via a process called *hematocytopoiesis*.

The primary function of erythrocytes is to aid in the transport of blood gases—about 30 to 34 percent (by weight) of each cell consisting of the oxygen- and carbon dioxide–carrying protein hemoglobin (64,000 MW $\leq$ 68,000) and a small portion of the cell containing the

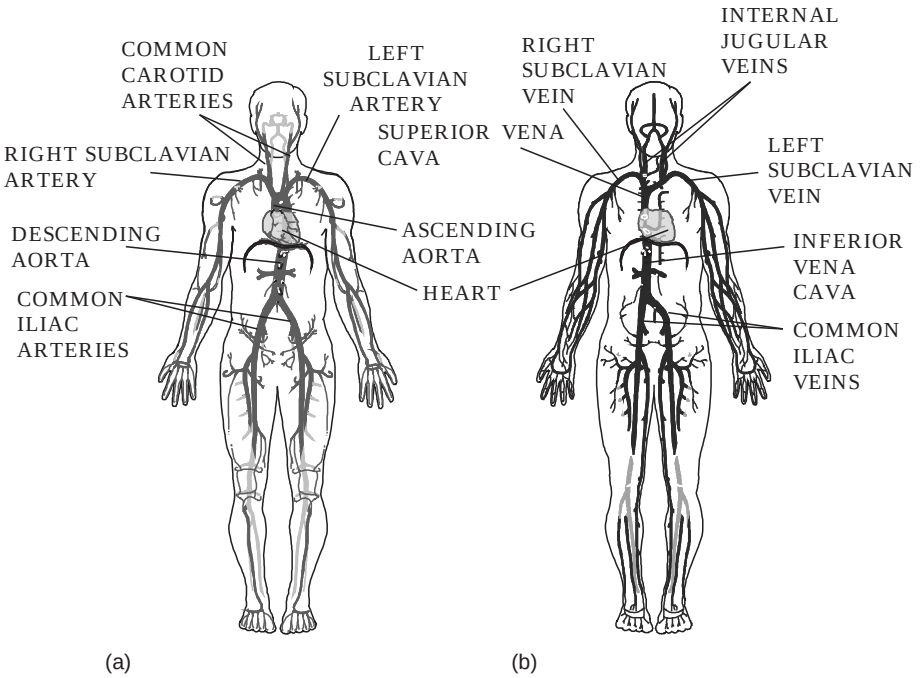


FIGURE 3.17 (a) The distribution of the main arteries in the body that carry blood away from the heart. (b) The distribution of the main veins in the body that return the blood to the heart.

enzyme carbonic anhydrase, which catalyzes the reversible formation of carbonic acid from carbon dioxide and water. The primary function of lukocytes is to endow the human body with the ability to identify and dispose of foreign substances (such as infectious organisms) that do not belong there—agranulocytes (lymphocytes and monocytes) essentially doing the "identifying" and granulocytes (neutrophils, basophils, and esinophils) essentially doing the "disposing." The primary function of platelets is to participate in the blood-clotting process.

Removal of all hematocytes from blood centrifugation or other separating techniques leaves behind the aqueous (91 percent water by weight, 94.8 percent water by volume), saline (0.15N) suspending medium called *plasma*—which has an average mass density of $1.035 +/- 0.005$ g/cm^3 and a viscosity 1½ to 2 times that of water. Some 6.5 to 8 percent by weight of plasma consists of the plasma proteins, of which there are three major types—albumin, the globulins, and fibrinogen—and several of lesser prominence.

The primary functions of albumin are to help maintain the osmotic (oncotic) transmural pressure differential that ensures proper mass exchange between blood and interstitial fluid at the capillary level and to serve as a transport carrier molecule for several hormones and other small biochemical constituents (such as some metal ions). The primary function of the globulin class of proteins is to act as transport carrier molecules (mostly of the α and β) for large biochemical substances, such as fats (lipoproteins) and certain carbohydrates (muco- and glycoproteins) and heavy metals (mineraloproteins), and to work together with leukocytes in the body's immune system. The latter function is primarily the responsibility

of the γ class of immunoglobulins, which have antibody activity. The primary function of fibrinogen is to work with thrombocytes in the formation of a blood clot—a process also aided by one of the most abundant of the lesser proteins, prothrombin.

Of the remaining 2 percent or so (by weight) of plasma, just under half consists of minerals (inorganic ash), trace elements, and electrolytes, mostly the cations sodium, potassium, calcium, and magnesium and the anions chlorine, bicarbonate, phosphate, and sulfate—the latter three helping as buffers to maintain the fluid at a slightly alkaline pH between 7.35 and 7.45 (average 7.4). What is left, about 1,087 mg materials per deciliter of plasma, includes (1) mainly three major types of fat—cholesterol (in a free and esterified form), phospholipids (a major ingredient of cell membranes), and triglyceride—with lesser amounts of the fat-soluble vitamins (A, D, E, and K), free fatty acids, and other lipids, and (2) "extractives" (0.25 percent by weight), of which about two-thirds include glucose and other forms of carbohydrate, the remainder consisting of the water-soluble vitamins (B-complex and C), certain enzymes, nonnitrogenous and nitrogenous waste products of metabolism (including urea, creatine, and creatinine), and many smaller amounts of other biochemical constituents—the list seeming virtuously endless. It is easy to understand why blood is often referred to as the "river of life." This river is made to flow through the vascular piping network by two central pumping stations arranged in series: the left and right sides of the human heart.

The heart (Figure 3.18), the pumping station that moves blood through the blood vessels, consists of two pumps: the right side and the left side. Each side has one chamber (the atrium) that receives blood and another chamber (the ventricle) that pumps the blood away from the heart. The right side moves deoxygenated blood that is loaded with carbon dioxide from the body to the lungs, and the left side receives oxygenated blood that has had most of its carbon dioxide removed from the lungs and pumps it to the body. The vessels that lead to and from the lungs make up the pulmonary circulation, and those that lead to and from the rest of the tissues in the body make up the systemic circulation (Figure 3.19).

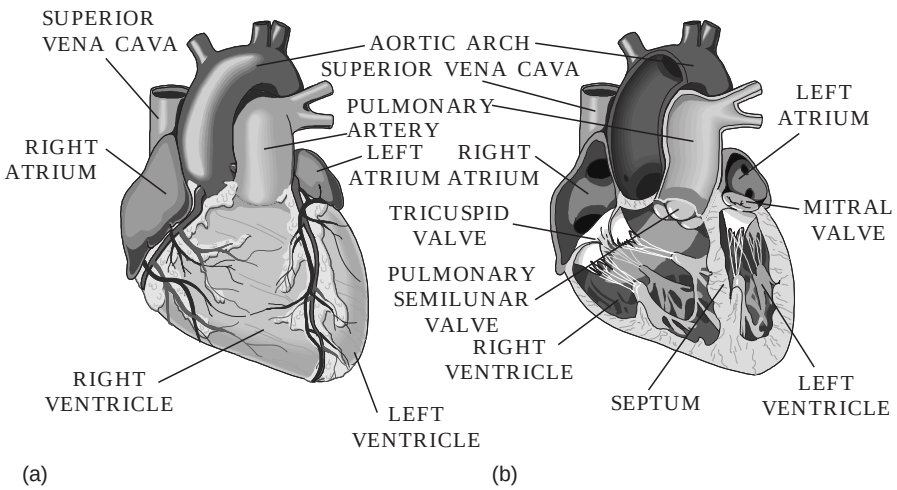


FIGURE 3.18 (a) The outside of the heart as seen from its anterior side. (b) The same view after the exterior surface of the heart has been removed. The four interior chambers—right and left atria and right and left ventricles—as well as several valves are visible.

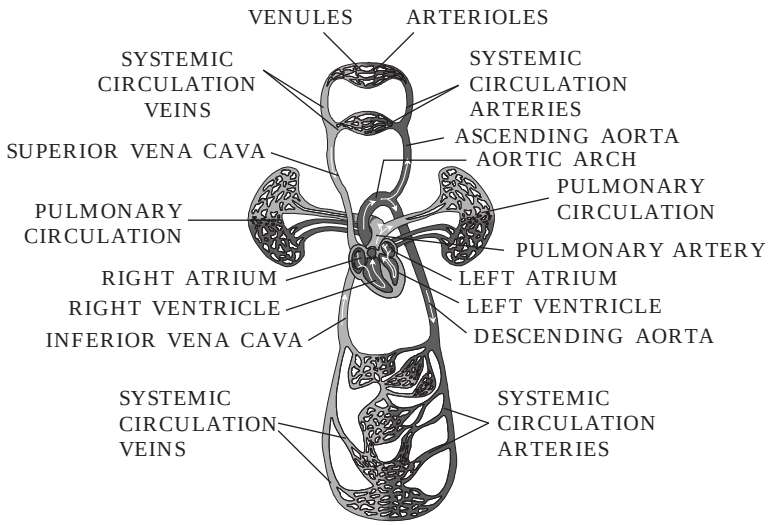


FIGURE 3.19 Oxygenated blood leaves the heart through the aorta. Some of the blood is sent to the head and upper extremities and torso, whereas the remainder goes to the lower torso and extremities. The blood leaves the aorta and moves into other arteries, then into smaller arterioles, and finally into capillary beds, where nutrients, hormones, gases, and waste products are exchanged between the nearby cells and the blood. The blood moves from the capillary beds into venules and then into veins. Blood from the upper part of the body returns to the right atrium of the heart through the superior vena cava, whereas blood from the lower part of the body returns through the inferior vena cava. The blood then moves from the right atrium to the right ventricle and into the pulmonary system through the pulmonary artery. After passing through capillaries in the lungs, the oxygenated blood returns to the left atrium of the heart through the pulmonary vein. It moves from the left atrium to the left ventricle and then out to the systemic circulation through the aorta to begin the same trip over again.

Blood vessels that carry blood away from the heart are called arteries, while those that carry blood toward the heart are called veins. The pulmonary artery is the only artery that carries deoxygenated blood, and the pulmonary vein is the only vein that carries oxygenated blood. The average adult has about 5 L of blood with 80 to 90 percent in the systemic circulation at any one time; 75 percent of the blood is in the systemic circulation in the veins, 20 percent is in the arteries, and 5 percent is in the capillaries.

Because of the anatomic proximity of the heart to the lungs, the right side of the heart does not have to work very hard to drive blood through the pulmonary circulation, so it functions as a low-pressure ($P \leq 40$ mmHg gauge) pump compared with the left side of the heart, which does most of its work at a high pressure (up to 140 mmHg gauge or more) to drive blood through the entire systemic circulation to the furthest extremes of the organism.

In order of size, the somewhat spherically shaped left atrium is the smallest chamber—holding about 45 ml of blood (at rest). The pouch-shaped right atrium is next (63 ml of blood), followed by the conical/cylindrically shaped left ventricle (100 ml of blood) and the crescent-shaped right ventricle (about 130 ml of blood). Altogether, then, the heart chambers collectively have a capacity of some 325 to 350 ml, or about 6.5 percent of the total blood volume in a "typical" individual—but these values are nominal, since the organ alternately fills and expands, contracts, and then empties.

During the 480-ms or so filling phase—diastole—of the average 750-ms cardiac cycle, the inlet valves of the two ventricles (3.8-cm-diameter tricuspid valve from right atrium to right ventricle; 3.1-cm-diameter bicuspid or mitral valve from left atrium to left ventricle) are open, and the outlet valves (2.4-cm-diameter pulmonary valve and 2.25-cm-diameter aortic semilunar valve, respectively) are closed—the heart ultimately expanding to its end-diastolic-volume (EDV), which is on the order of 140 ml of blood for the left ventricle. During the 270-ms emptying phase—systole—electrically induced vigorous contraction of cardiac muscle drives the intraventricular pressure up, forcing the one-way inlet valves closed and the unidirectional outlet valves open as the heart contracts to its end-systolic-volume (ESV), which is typically on the order to 70 ml of blood for the left ventricle. Thus, the ventricles normally empty about half their contained volume with each heartbeat, the remainder being termed the *cardiac reserve volume*. More generally, the difference between the actual EDV and the actual ESV, called the *stroke volume* (SV), is the volume of blood expelled from the heart during each systolic interval, and the ratio of SV to EDV is called the *cardiac ejection faction*, or *ejection ratio* (0.5 to 0.75 is normal, 0.4 to 0.5 signifies mild cardiac damage, 0.25 to 0.40 implies moderate heart damage, and less than 0.25 warns of severe damage to the heart's pumping ability). If the stroke volume is multiplied by the number of systolic intervals per minute, or heart rate (HR), one obtains the total cardiac output (CO):

$$\mathrm{CO} = \mathrm{HR} \times (\mathrm{EDV} - \mathrm{ESV})$$

where EDV-ESV to the stroke volume.

Several investigations have suggested that the cardiac output (in milliliters per minute) is proportional to the weight W (in kilograms) of an individual according to the equation

$$\mathrm{CO} - 224\mathrm{W}^{3/4}$$

and that "normal" heart rate obeys very closely the relation

$$\mathrm{HR} = 229\mathrm{W}^{-1/4}$$

For a "typical" 68.7-kg individual (blood volume = 5,200 ml), these equations yield CO = 5,345 ml/min, HR = 80 beats/min (cardiac cycle period = 754 ms) and SV = CO/HR = $229\mathrm{W}^{-1/4}$/CO-$224\mathrm{W}^{3/4}$ = 0.978W = 67.2 ml/beat, which are very reasonable values. Furthermore, assuming this individual lives to be about 75 years old, his or her heart will have cycled over 3.1536 billion times, pumping a total of 0.2107 billion liters of blood (55.665 million gallons, or 8,134 quarts per day) within their lifetime.

In the normal heart, the cardiac cycle, which refers to the repeating pattern of contraction (systole) and relaxation (diastole) of the chambers of the heart, begins with a self-generating electrical pulse in the pacemaker cells of the sinoatrial node (Figure 3.20). This rapid electrical change in the cells is the result of the movement of ions across their plasma membranes. The permeability of the plasma membrane to Na^+ changes dramatically and allows these ions to rush into the cell. This change in the electrical potential across the plasma membrane from one in which the interior of the cell is more negative than the extracellular fluid (approximately –90 mV) to one in which the interior of the cell is more positive than the extracellular fluid (approximately 20 mV) is called depolarization. After a very short period of time (<0.3 s), changes in the membrane and activation of the sodium-potassium pumps result in repolarization, the restoration of the original ionic balance in the cells. The entire electrical event in which the polarity of the potential across the plasma membrane rapidly

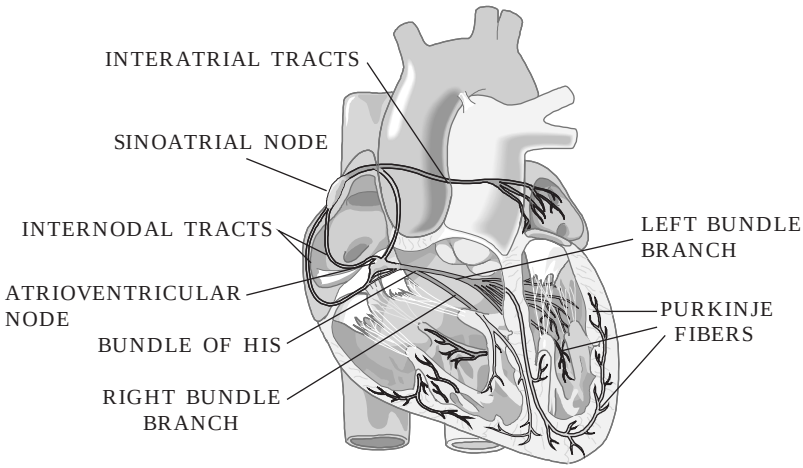


FIGURE 3.20 Pacemaker cells in the sinoatrial node (SA node) depolarize first and send an activation wavefront through the atria. The propagating action potential slows down as it passes through the atrioventricular node (AV node), then moves through the bundle of His and Purkinje system very rapidly until it reaches the cells of the ventricles.

reverses and then becomes reestablished is called an action potential. The cells in the sinoatrial node depolarize on the average of every 0.83 s in a typical adult at rest. This gives a resting heart rate of 72 beats per minute, with about $\frac{5}{8}$ of each beat spent in diastole and $\frac{3}{8}$ in systole.

Cardiac cells are linked and tightly coupled so action potentials spread from one cell to the next. Activation wavefronts move across the atria at a rate of about 1 m/s. When cardiac cells depolarize, they also contract. The contraction process in the atria (atrial systole) moves blood from the right atrium to the right ventricle and from the left atrium to the left ventricle (Figure 3.21). The activation wavefront then moves to the atrioventricular

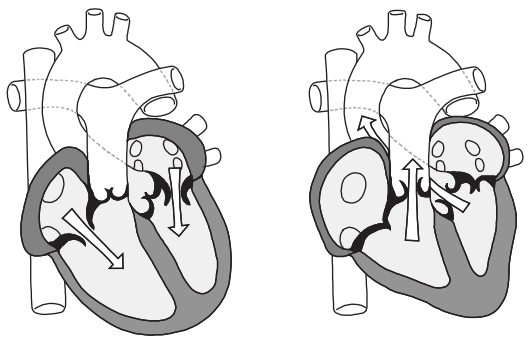


FIGURE 3.21 (a) During the first part of the cardiac cycle, the atria contract (atrial systole) and move blood into the ventricles. (b) During the second part of the cardiac cycle, the atria relax (diastole), and the ventricles contract (ventricular systole) and move blood to the lungs (pulmonary circulation) and to the rest of the body (systemic circulation).

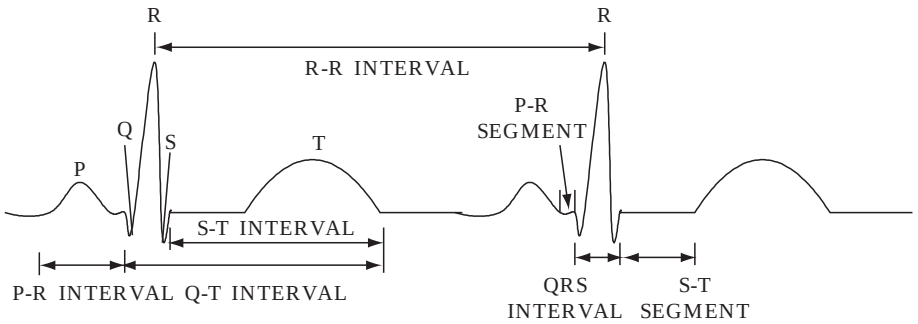


FIGURE 3.22 Typical lead II ECG. This electrocardiogram is typical of one that would be recorded from the body's surface by having a positive electrode on the left leg and a negative electrode on the right arm. The vertical direction represents voltage, and the horizontal direction represents time. The P, R, and T waves are easily identified and are the result of the movement of ions in cells in different parts of the heart. Different intervals and segments have been identified that provide information about the health of the heart and its conduction system. The R-R interval can be used to determine heart rate.

(AV) node, where it slows to a rate of about 0.05 m/s to allow time for the ventricles to completely fill with the blood from the atria. After leaving the AV node, the activation wavefront moves to specialized conduction tissue, the Purkinje system, which spreads the wavefront very rapidly (at about 3 m/s) to many cells in both venricles. The activation wavefront spreads through ventricular tissue at about 0.5 m/s. This results in the simultaneous contraction of both ventricles (ventricular systole) so blood is forced from the heart into the pulmonary artery from the right ventricle and into the aorta from the left ventricle.

The electrocardiogram (ECG; Figure 3.22) is an electrical measure of the sum of these ionic changes within the heart. The P wave represents the depolarization of the atria, while the QRS represents the depolarization of the ventricles. Ventricular repolarization shows up as the T wave, while atrial repolarization is masked by ventricular depolarization. Changes in the amplitude and duration of the different parts of the ECG provide diagnostic information for physicians. Many biomedical engineers have worked on methods for recording and analyzing ECGs.

EXAMPLE PROBLEM 3.6

What would be the heart rate given by an ECG in which 10 R-waves occurred in 6.4 s?

Solution

A sequence of 10 R-waves represents 9 R-R intervals (see Figure 3.22) or beats of the heart.

$$\left(\frac{9\ beats}{6.4\ s}\right)\left(\frac{60\ s}{1\ \text{min}}\right) = 84\ bpm$$

EXAMPLE PROBLEM 3.7

What would be the cardiac output of the heart in Example Problem 3.6 if the stroke volume is 75 ml?

Solution

The cardiac output (given in liters per minute) is the product of the heart rate and the stroke volume.

$$CO = 84\frac{beats}{\text{min}} \times 75\frac{ml}{beat} = 6300\frac{ml}{\text{min}} = 6.3\frac{liters}{\text{min}}$$

During atrial and ventricular systole, special one-way valves (Figure 3.23a) keep the blood moving in the correct direction. When the atria contract, the atrioventricular valves (tricuspid and mitral) open to allow blood to pass into the ventricles. During ventricular systole, the semilunar valves (aortic and pulmonary) open to allow blood to leave the heart, while the atrioventricular valves close and prevent blood from flowing backward from the ventricles to the atria. The aortic and pulmonary valves prevent blood from flowing back from the pulmonary artery and aorta into the right and left ventricles, respectively. If a valve becomes calcified or diseased or is not properly formed during embryonic development, it can be replaced by an artificial valve (Figure 3.23b), a device that has been developed by cooperative work between biomedical engineers and physicians.

Blood pressure can be measured directly or indirectly (noninvasively). Direct blood pressure measurements are made by introducing a catheter or needle that is coupled to a pressure transducer into a vein or artery. Indirect methods include sphygmomanometry,

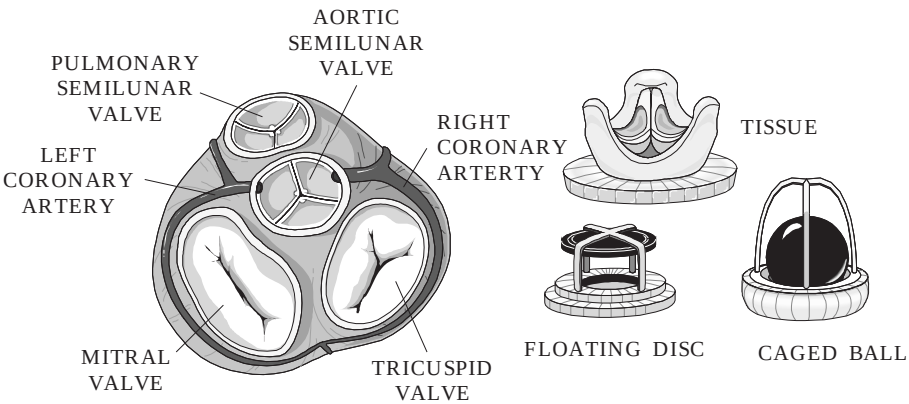


FIGURE 3.23 (a) The atrioventricular (tricuspid and mitral) and semilunar (pulmonary and aortic) valves. (b) Three types of artificial valves—tissue, floating disc, and caged ball—that can be used to replace diseased or malformed human valves.

in which a cuff is used to apply sufficient pressure to an artery, usually in the arm, to prevent the flow of blood through the artery, and a stethoscope is used to listen to the change in sounds as the cuff is slowly deflated. The first Korotkoff sounds occur when the systolic pressure, the highest pressure reached when the ventricles contract and eject blood, first exceeds the pressure in the cuff so blood once again flows through the artery beneath the stethoscope. The Korotkoff sounds become muffled and disappear when the pressure in the cuff drops below the diastolic pressure, the minimum pressure that occurs at the end of ventricular relaxation. Another indirect measurement is the oscillometric method, which uses a microprocessor to periodically inflate and slowly deflate a cuff. When blood breaks through the occlusion caused by the cuff, the walls of the artery begin to vibrate slightly due to the turbulent nature of the blood flow. The onset of these oscillations in pressure correlates with the systolic pressure. The oscillations decrease in amplitude over time with the diastolic pressure event corresponding to the point at which the rate of amplitude decrease suddenly changes slope. A third indirect measurement, the ultrasonic method, depends on the Doppler shift of sound waves that hit red blood cells that are flowing with the blood.

Blood in the systemic circulation leaves the heart through the aorta with an average internal pressure of about 100 mmHg (maximum systolic pressure of about 120 mm Hg, with a diastolic pressure of about 80 mm Hg in a normal adult) and moves to medium-sized arteries (see Figure 3.17a) and arterioles. Arterioles lead to capillaries (average internal pressure of about 30 mm Hg), which are followed by venules. Venules lead to medium-sized veins, then to large veins, and finally to the venae cavae (average internal pressure of about 10 mm Hg), which return blood to the heart at the right atrium. Blood in the pulmonary circulation (see Figure 3.19) leaves the pulmonary artery and moves to arterioles and then the capillary beds within the lungs. It returns to the heart through the left atrium. Blood flow is highest in the large arteries and veins (30–40 cm/s in the aorta; 5 cm/s in the vena cavae) and slowest in the capillary beds (1 mm/s), where the exchange of nutrients, metabolic wastes, gases, and hormones takes place. Pressures in the pulmonary circulation are lower (25 mm Hg/10 mm Hg) than in the systemic circulation due to the decreased pumping power of the smaller right ventricle as compared to the left and to the lower resistance of blood vessels in the lungs.

EXAMPLE PROBLEM 3.8

What would be the pulse pressure and the mean arterial pressure for a person with a blood pressure reading of 118 mmHg/79 mmHg?

Solution

The pulse pressure is defined as the difference between the systolic (118 mmHg) and diastolic (79 mmHg) pressures, which would be 39 mmHg in this case.

Mean arterial pressure is the average blood pressure in the arteries and is estimated as the diastolic pressure plus one-third of the pulse pressure, which would be 92 mmHg in this example.

3.4.2 Respiratory System

The respiratory system (Figure 3.24a) moves air to and from the gas exchange surfaces in the body where diffusion can occur between air and the circulating blood. It includes the conduction zone and the respiratory zone. In the conduction zone (mouth, nose, sinuses, pharynx, trachea, bronchi, and bronchioles), the air that enters the body is warmed, humidified, filtered, and cleaned. Mucus is secreted by cells in the conduction zone and traps small particles (>6 μm) before they can reach the respiratory zone. Epithelial cells that line the trachea and bronchi have cilia that beat in a coordinated fashion to move mucus toward the pharynx, where it can be swallowed or expectorated. The respiratory zone, consisting of respiratory bronchioles with outpouchings of alveoli and terminal clusters of alveolar sacs, is where gas exchange between air and blood occurs (Figure 3.24b). The respiratory zone comprises most of the mass of the lungs.

Conduction of air begins at the larynx, or voice box, at the entrance to the trachea, which is a fibromuscular tube 10 to 12 cm in length and 1.4 to 2.0 cm in diameter. At a location called the *carina*, the trachea terminates and divides in the left and right bronchi. Each bronchus has a discontinuous cartilaginous support in its wall. Muscle fibers capable of controlling airway diameter are incorporated into the walls of the bronchi, as well as in those of air passages closer to the alveoli. Smooth muscle is present throughout the respiratory bronchiolus and alveolar ducts but is absent in the last alveolar duct, which terminates in one to several alveoli. The alveolar walls are shared by other alveoli and are composed of highly pliable and collapsible squamous epithelium cells. The bronchi subdivide into subbronchi, which further subdivide into bronchioli, which further subdivide, and so on, until finally reaching the alveolar level.

Movement of gases in the respiratory airways occurs mainly by bulk flow (convection) throughout the region from the mouth to the nose to the fifteenth generation. Beyond the

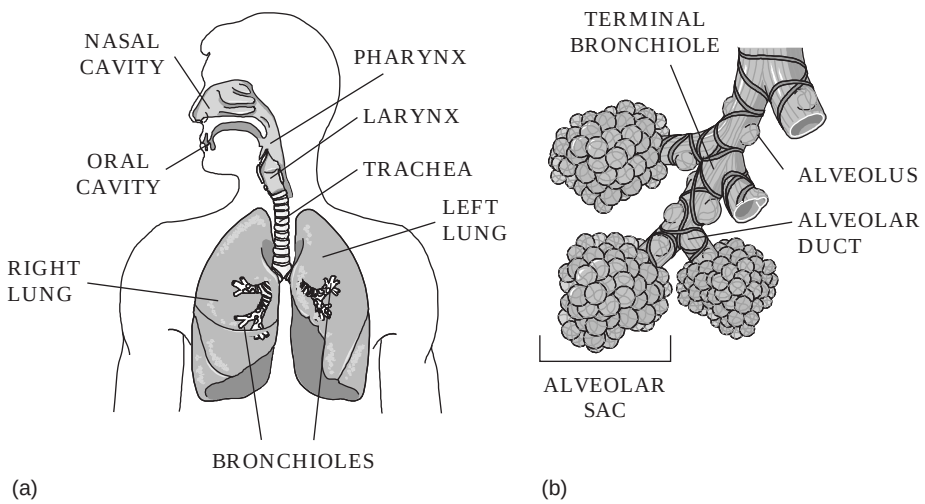


FIGURE 3.24 (a) The respiratory system consists of the passageways that are used to move air into and out of the body and the lungs. (b) The terminal bronchioles and alveolar sacs within the lungs have alveoli where gas exchange occurs between the lungs and the blood in the surrounding capillaries.

fifteenth generation, gas diffusion is relatively more important. With the low gas velocities that occur in diffusion, dimensions of the space over which diffusion occurs (alveolar space) must be small for adequate oxygen delivery into the walls; smaller alveoli are more efficient in transfer of gas than are larger ones. Thus, animals with high levels of oxygen consumption are found to have smaller-diameter alveoli compared with animals with low levels of oxygen consumption.

Alveoli are the structures through which gases diffuse to and from the body. To ensure that gas exchange occurs efficiently, alveolar walls are extremely thin. For example, the total tissue thickness between the inside of the alveolus to pulmonary capillary blood plasma is only about 0.4×10^{-6} m. Consequently, the principal barrier to diffusion occurs at the plasma and red blood cell level, not at the alveolar membrane.

Molecular diffusion within the alveolar volume is responsible for mixing of the enclosed gas. Due to small alveolar dimensions, complete mixing probably occurs in less than 10 ms, fast enough that alveolar mixing time does not limit gaseous diffusion to or from the blood.

Of particular importance to proper alveolar operation is a thin surface coating of surfactant. Without this material, large alveoli would tend to enlarge and small alveoli would collapse. It is the present view that surfactant acts like a detergent, changing the stress-strain relationship of the alveolar wall and thereby stabilizing the lung.

Certain physical properties, such as compliance, elasticity, and surface tension, are characteristic of lungs. Compliance refers to the ease with which lungs can expand under pressure. A normal lung is about 100 times more distensible than a toy balloon. Elasticity refers to the ease with which the lungs and other thoracic structures return to their initial sizes after being distended. This aids in pushing air out of the lungs during expiration. Surface tension is exerted by the thin film of fluid in the alveoli and acts to resist distention. It creates a force that is directed inward and creates pressure in the alveolus, which is directly proportional to the surface tension and inversely proportional to the radius of the alveolus (Law of Laplace). Thus, the pressure inside an alveolus with a small radius would be higher than the pressure inside an adjacent alveolus with a larger radius and would result in air flowing from the smaller alveolus into the larger one. This could cause the smaller alveolus to collapse. This does not happen in normal lungs because the fluid inside the alveoli contains a phospholipid that acts as a surfactant. The surfactant lowers the surface tension in the alveoli and allows them to get smaller during expiration without collapsing. Premature babies often suffer from respiratory distress syndrome because their lungs lack sufficient surfactant to prevent their alveoli from collapsing. These babies can be kept alive with mechanical ventilators or surfactant sprays until their lungs mature enough to produce surfactant.

Breathing, or ventilation, is the mechanical process by which air is moved into (inspiration) and out of (expiration) the lungs. A normal adult takes about 15 to 20 breaths per minute. During inspiration, the inspiratory muscles contract and enlarge the thoracic cavity, the portion of the body where the lungs are located. This causes the alveoli to enlarge and the alveolar gas to expand. As the alveolar gas expands, the partial pressure within the respiratory system drops below atmospheric pressure by about 3 mmHg so air easily flows in (Boyle's Law). During expiration, the inspiratory muscles relax and return the thoracic cavity to its original volume. Since the volume of the gas inside the respiratory system

has decreased, its pressure increases to a value that is about 3 mmHg above atmospheric pressure. Air now moves out of the lungs and into the atmosphere.

The primary purpose of the respiratory system is gas exchange. In the gas-exchange process, gas must diffuse through the alveolar space, across tissue, and through plasma into the red blood cell, where it finally chemically joins to hemoglobin. A similar process occurs for carbon dioxide elimination.

As long as intermolecular interactions are small, most gases of physiologic significance can be considered to obey the ideal gas law:

$$pV = nRT$$

where

p = pressure, N/m^3
V = volume of gas, m^3
n = number of moles, mol
R = gas constant, $(N \times m)/(mol \times K)$
T = absolute temperature, K

The ideal gas law can be applied without error up to atmospheric pressure; it can be applied to a mixture of gases, such as air, or to its constituents, such as oxygen or nitrogen. All individual gases in a mixture are considered to fill the total volume and have the same temperature but reduced pressures. The pressure exerted by each individual gas is called the *partial pressure* of the gas.

Dalton's law states that the total pressure is the sum of the partial pressures of the constituents of a mixture:

$$P = \sum_{i=1}^{N} pi$$

where

p_i = partial pressure of the ith constituent, N/m^3
N = total number of constituents

Dividing the ideal gas law for a constituent by that for the mixture gives

$$\frac{P_i V}{PV} = \frac{n_i R_i T}{nRT}$$

so that

$$\frac{p_i}{p} = \frac{n_i R_i}{nR}$$

which states that the partial pressure of a gas may be found if the total pressure, mole fraction, and ratio of gas constants are known. For most respiratory calculations, p will be considered to be the pressure of 1 atmosphere, 101 kN/m2. Avogadro's principle states that different gases at the same temperature and pressure contain equal numbers of molecules:

$$\frac{V_1}{V_2} = \frac{nR_1}{nR_2} = \frac{R_1}{R_2}$$

Thus,

$$\frac{p_i}{p} = \frac{V_i}{V}$$

where V_i/V is the volume fraction of a constituent in air and is therefore dimensionless. Table 3.2 provides individual gas constants, as well as volume fractions of constituent gases of air.

Lung mechanics refers to the study of the mechanical properties of the lung and chest wall, whereas lung statics refers to the mechanical properties of a lung in which the volume is held constant over time. Understanding lung mechanics requires knowledge about the volumes within the lungs. Lung capacities contain two or more volumes. The tidal volume (TV) is the amount of air that moves in and out of the lungs during normal breathing (Figure 3.25). The total lung capacity (TLC) is the amount of gas contained within the lungs at the end of a maximum inspiration. The vital capacity (VC) is the maximum amount of air that can be exhaled from the lungs after inspiration to TLC. The residual volume (RV) is the amount of gas remaining in the lungs after maximum exhalation. The amount of gas that can be inhaled after inhaling during tidal breathing is called the inspiratory reserve volume (IRV). The amount of gas that can be expelled by a maximal exhalation after exhaling during tidal breathing is called the expiratory reserve volume (ERV). The inspiratory capacity (IC) is the maximum amount of gas that can be inspired after a normal exhalation during tidal breathing, and the functional residual capacity (FRC) is the amount of gas that remains in the lungs at this time (Table 3.3).

All of the volumes and capacities except those that include the residual volume can be measured with a spirometer. The classic spirometer is an air-filled container that is constructed from two drums of different sizes. One drum contains water, and the other air-filled drum is inverted over an air-filled tube and floats in the water. The tube is connected to a mouthpiece used by the patient. When the patient inhales, the level of the

TABLE 3.2 Molecular Masses, Gas Constants, and Volume Fractions for Air and Constituents

Constituent	Molecular Mass kg/mol	Gas Constant N m/(mol K)	Volume Fraction in Air m^3/m^3
Air	29.0	286.7	1.0000
Ammonia	17.0	489.1	0.0000
Argon	39.9	208.4	0.0093
Carbon Dioxide	44.0	189.0	0.0003
Carbon Monoxide	28.0	296.9	0.0000
Helium	4.0	2,078.6	0.0000
Hydrogen	2.0	4,157.2	0.0000
Nitrogen	28.0	296.9	0.7808
Oxygen	32.0	259.8	0.2095

Note: Universal gas constant is 8314.43 N m/kg K.

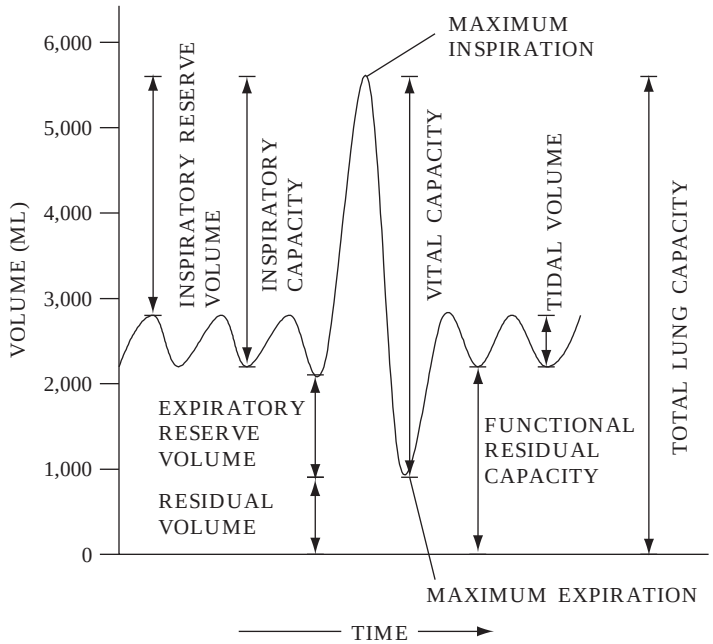


FIGURE 3.25 Lung volumes and capacities, except for residual volume, functional residual capacity, and total lung capacity, can be measured using spirometry.

floating drum drops. When the patient exhales, the level of the floating drum rises. These changes in floating drum position can be recorded and used to measure lung volumes.

EXAMPLE PROBLEM 3.9

The total lung capacity of a patient is 5.9 liters. If the patient's inspiratory capacity was found to be 3.3 liters using spirometry, what would be the patient's functional residual capacity? What would you need to measure in order to determine the patient's residual volume?

Solution

From Figure 3.25, total lung capacity (TLC) is equal to the sum of inspiratory capacity (IC) and functional residual capacity (FRC).

$$TLC = IC + FRC$$

$$5.9\,l = 3.3\,l + FRC$$

$$FRC = 2.6\,l$$

TLC, which cannot be determined by means of spirometry, and vital capacity (VC), which can be measured using spirometry, must be known in order to determine residual volume (RV), since

$$TLC - VC = RV$$

TABLE 3.3 Typical Lung Volumes for a Normal, Healthy Male

Lung Volume	Normal Values	
Total lung capacity (TLC)	$6.0 \times 10^{-3}m^3$	(6,000 cm^3)
Residual volume (RV)	$1.2 \times 10^{-3}m^3$	(1,200 cm^3)
Vital capacity (VC)	$4.8 \times 10^{-3}m^3$	(4,800 cm^3)
Inspiratory reserve volume (IRV)	$3.6 \times 10^{-3}m^3$	(3,600 cm^3)
Expiratory reserve volume (ERV)	$1.2 \times 10^{-3}m^3$	(1,200 cm^3)
Functional residual capacity (FRC)	$2.4 \times 10^{-3}m^3$	(2,400 cm^3)
Anatomic dead volume (V_D)	$1.5 \times 10^{-4}m^3$	(150 cm^3)
Upper airways volume	$8.0 \times 10^{-5}m^3$	(80 cm^3)
Lower airways volume	$7.0 \times 10^{-5}m^3$	(70 cm^3)
Physiologic dead volume (V_D)	$1.8 \times 10^{-4}m^3$	(180 cm^3)
Minute volume (V_e) at rest	$1.0 \times 10^{-4}m^3/s$	(6,000 cm^3/m)
Respiratory period (T) at rest	4s	
Tidal volume (V_T) at rest	$4.0 \times 10^{-4}m^3$	(400 cm^3)
Alveolar ventilation volume (V_A) at rest	$2.5 \times 10^{-4}m^3$	(250 cm^3)
Minute volume during heavy exercise	$1.7 \times 10^{-3}m^3/s$	(10,000 m^3/m)
Respiratory period during heavy exercise	1.2s	
Tidal volume during heavy exercise	$2.0 \times 10^{-3}m^3$	(2,000 cm^3)
Alveolar ventilation during heavy exercise	$1.8 \times 10^{-3}m^3$	(1,820 cm^3)

Since spirograms record changes in volume over time, flow rates can be determined for different maneuvers. For example, if a patient exhales as forcefully as possible to residual volume following inspiration to TLC, then the forced expiratory volume ($FEV_{1.0}$) is the total volume exhaled at the end of 1 s. The $FEV_{1.0}$ is normally about 80 percent of the vital capacity. Restrictive diseases, in which inspiration is limited by reduced compliance of the lung or chest wall or by weakness of the inspiratory muscles, result in reduced values for $FEV_{1.0}$ and vital capacity, but their ratio remains about the same. In obstructive diseases, such as asthma, the $FEV_{1.0}$ is reduced much more than the vital capacity. In these diseases, the TLC is abnormally large, but expiration ends prematurely. Another useful measurement is the forced expiratory flow rate ($FEF_{25–75}$ percent), which is the average flow rate measured over the middle half of the expiration—that is, from 25 to 75 percent of the vital capacity. Flow-volume loops provide another method for analyzing lung function by relating the rate of inspiration and expiration to the volume of air that is moved during each process.

The TLC can be measured using the gas dilution technique. In this method, patients inspire to TLC from a gas mixture containing a known amount of an inert tracer gas, such as helium, and hold their breath for 10 s. During this time, the inert gas becomes evenly distributed throughout the lungs and airways. Due to conservation of mass, the product of initial tracer gas concentration (which is known) times the amount inhaled (which is measured) equals the product of final tracer gas concentration (which is measured during expiration) times the TLC. Body plethysmography, which provides the most accurate method for measuring lung volumes, uses an airtight chamber in which the patient sits and breathes through a mouthpiece. This method uses Boyle's Law, which states that the product of pressure and volume for gas in a chamber is constant under isothermal conditions. Changes in lung volume and pressure at the mouth when the patient pants against a closed shutter can be used to calculate the functional residual capacity. Since the expiratory reserve volume can be measured, the residual volume can be calculated by subtracting it from the functional residual capacity.

EXAMPLE PROBLEM 3.10

A patient is allowed to breathe a mixture from a 2-liter reservoir that contains 10 percent of an inert gas—that is, a gas that will not cross from the lungs into the circulatory system. At the end of a period that is sufficient for the contents of the reservoir and the lungs to equilibrate, the concentration of the inert gas is measured and is found to be 2.7 percent. What is the patient's total lung capacity?

Solution

The total amount of inert gas is the same at the beginning and end of the measurement, but its concentration has changed from 10 percent (C_1) to 2.7 percent (C_2). At the beginning, it is confined to a 2-liter reservoir (V_1). At the end, it is in both the reservoir and the patient's lungs ($V_2 = V_1 + \text{TLC}$).

$$C_1 V_1 = C_2 V_2$$
$$(0.1)(2\,l) = (0.027)(2\,l + TLC)$$
$$0.2\,l - 0.054\,l = 0.027\,TLC$$
$$5.4\,l = TLC$$

External respiration occurs in the lungs when gases are exchanged between the blood and the alveoli (Figure 3.26). Each adult lung contains about 3.5×10^8 alveoli, which results in a large surface area (60–70 m^2) for gas exchange to occur. Each alveolus is only one cell layer thick, so the air-blood barrier is only two cells thick (an alveolar cell and a capillary endothelial cell), which is about 2 μm. The partial pressure of oxygen in the alveoli is higher than the partial pressure of oxygen in the blood, so oxygen moves from the alveoli into the blood. The partial pressure of carbon dioxide in the alveoli is lower than the partial pressure of carbon dioxide in the blood, so carbon dioxide moves from the blood into the alveoli. During internal respiration, carbon dioxide and oxygen move between the blood and the extracellular fluid surrounding the body's cells. The direction and rate of

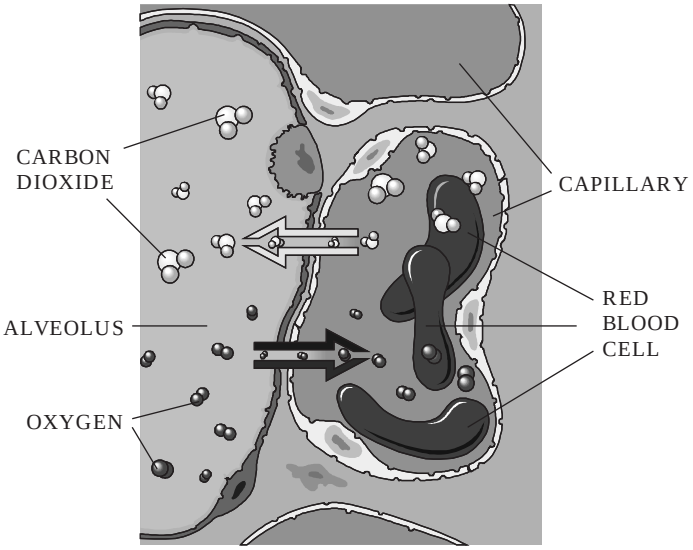


FIGURE 3.26 During external respiration, oxygen moves from the alveoli to the blood, and carbon dioxide moves from the blood to the air within the alveoli.

movement of a gas depends on the partial pressures of the gas in the blood and the extracellular fluid, the surface area available for diffusion, the thickness of the membrane that the gas must pass through, and a diffusion constant that is related to the solubility and molecular weight of the gas (Fick's Law).

Mechanical ventilators can be used to deliver air or oxygen to a patient. They can be electrically or pneumatically powered and can be controlled by microprocessors. Negative pressure ventilators, such as iron lungs, surround the thoracic cavity and force air into the lungs by creating a negative pressure around the chest. This type of ventilator greatly limits access to the patient. Positive pressure ventilators apply high-pressure gas at the entrance to the patient's lungs so air or oxygen flows down a pressure gradient and into the patient. These ventilators can be operated in control mode to breathe for the patient at all times or in assist mode to help with ventilation when the patient initiates the breathing cycle. This type of ventilation changes the pressure within the thoracic cavity to positive during inspiration, which affects venous return to the heart and cardiac output (the amount of blood the heart moves with each beat). High-frequency jet ventilators deliver very rapid (60–90 breaths per minute) low-volume bursts of air to the lungs. Oxygen and carbon dioxide are exchanged by molecular diffusion rather than by the mass movement of air. This method causes less interference with cardiac output than does positive pressure ventilation. Extracorporeal membrane oxygenation (ECMO) uses the technology that was developed for cardiopulmonary bypass machines. Blood is removed from the patient and passed through an artificial lung, where oxygen and carbon dioxide are exchanged. It is warmed to body temperature before being returned to the patient. This technique allows the patient's lungs to rest and heal themselves and has been used successfully on some cold water drowning victims and on infants with reversible pulmonary disease.

3.4.3 The Nervous System

The most exciting and mysterious part of the human body is the magical 3½ pounds of tissue we carry around inside our skulls: the brain. For centuries, the brain has frustrated those daring enough to explore its secrets. Encased not only in its bony protective covering but also in a shroud of mysticism, it has been an extremely difficult structure to study. Even with the invention of the microscope and the discovery of electricity, generations of Western scientists refrained from investigating the activity of the human brain out of respect for it as the seat of a human's immortal soul.

In recent years, this convoluted mass, the source of all thought and emotion, has been the focal point of intense scientific investigation. There has been a great flurry of activity in assembling interdisciplinary teams consisting of physiologists, psychologists, biochemists, and engineers in order to gain a better understanding of brain function. To many of these individuals, the brain represents a symbolic Mt. Everest, an obstacle to be scaled and conquered before it will be better understood. And yet, in spite of all the efforts to date, we are still only in the foothills of such a climb. The mechanisms and processes that enable the brain to convert the variety of electrical and chemical activity occurring within it into thoughts, feeling, dreams, and memories—the fundamental awareness of self—are still beyond our understanding.

However, in spite of the difficulties encountered and the frustrations experienced by explorers in this world of the mind, significant progress has been made in deciphering the cryptic flow of electrical energy coming from the brain. In reviewing these electrical signals, it has been possible to detect the presence of certain patterns or rhythms that occur in the brain that represent a "language" that can be recognized and understood by neural circuits in the brain itself. The fundamental building block of this neuronal communication network is the individual nerve cell: the neuron. Figure 3.16d is a schematic drawing of just such a cell. It consists of three major components: the cell body itself, or *soma*; the receptor zone, or *dendrites*; and a long fiber called the *axon*, which carries electrical signals from the main body of the cell to the muscles, glands, or other neurons. Numbering approximately 20 billion in each human being, these tiny cells come in a variety of sizes and shapes. However, nowhere is more variety displayed than in the length of the axonal terminating fiber. In the human body, it ranges from a few thousandths of an inch up to three feet or more, depending on the type of neuron involved. Consider, for example, the long pathways from the extremities to the brain. In these communication channels between the periphery and the "central data processor" that we call the brain, only a few neurons may be connected to one another. As a result, the axon of these nerve cells may be as long as 2 or 3 feet, even though the cell body is quite small. Some axons are surrounded by sheaths of myelin that are formed by specialized nonneural cells called Schwann cells. Gaps, called "Nodes of Ranvier," in the myelin sheath allow the action potential generated by the neuron to travel more rapidly by essentially jumping from one node to the next.

Neurons are anatomically distinct units with no physical continuity between them. The transmitting portion of a neuron, its axon, ends in a series of synapses, thereby making contact with other neurons (see Figure 3.28). Under the microscope this often stands out as a spherical enlargement at the end of the axon to which various names have been given, for example, boutons, end-feet, or synaptic terminals. This ending does not actually make contact with the soma or the dendrite but is separated by a narrow cleft (gap) that is, on

average, 100 to 200 Angstroms (10^{-9} meters) wide. This is known as the *synaptic cleft*. Each of these synaptic endings contains a large number of submicroscopic spherical structures (*synaptic vesicles*) that can be detected only under the electron microscope. These synaptic vesicles, in turn, are essentially "chemical carriers" containing transmitter substance that is released into the synaptic clefts on excitation. With this information in hand, let us consider the sequence of events that enables one neuron to communicate with another.

When an individual neuron is excited, an electrical signal is transmitted along its axon to many tiny branches, diverging fibers near its far end. These axonal terminals end or synapse close to the "input terminals" (the dendrites and cell body) of a large number of other neurons. When an electrical pulse arrives at the synapse, it triggers the release of a tiny amount of transmitter substance. This chemical carrier floats across the synaptic cleft between the axonal fiber and the cell body, thereby altering the status of the receiving neuron. For example, the chemical emissions may urge the receiving neuron into a state whereby this second cell is activated and conducts a similar electrical pulse to its axon. In this way, the initial electrical signal may be propagated to a still more remote part of the other hand. If the surfaces of muscle cells lie close enough to a number of such terminals to receive a substantial supply of these chemical carriers, the muscle will experience a resulting electrochemical reaction of its own that will cause it to contract and thereby perform some mechanical chore. In a similar manner, a gland can be stimulated to secrete the chemical characteristic to its activity. Neurons with the ability to cause a muscular or glandular reaction are known as effector neurons or motoneurons.

For most of us, the most pleasurable sensations come from our perception of the world around us. This sense of awareness is made possible by still another group of specialized neurons known as *receptor cells*. Acting as input devices, these neurons accept and convert various sensory information into appropriate electrical impulses that can then be properly processed within the nervous system. These receptors measure such quantities as pressure, warmth, cold, and displacement, as well as the presence of specific chemicals. Considering that every minute of one's life the brain is virtually bombarded by such a voluminous amount of incoming information, it is astounding that it can function at all.

So as we have seen, nerve cells are responsible for the following variety of essential functions:

1. Accepting and converting sensory information into a form that can be processed with the nervous system by other neurons.
2. Processing and analyzing this information so an "integrated portrait" of the incoming data can be obtained.
3. Translating the final outcome or "decision" of this analysis process into an appropriate electrical or chemical form needed to stimulate glands or activate muscles.

The nervous system, which is responsible for the integration and control of all the body's functions, has been divided by neuroscientists into the central nervous system (CNS) and the peripheral nervous system (PNS) (Figure 3.27). The former consists of all nervous tissue enclosed by bone (e.g., the brain and spinal cord), and the latter consists of all nervous tissue not enclosed by bone, which enables the body to detect and respond to both internal and external stimuli. The peripheral nervous system consists of the 12 pairs of cranial and 31 pairs of spinal nerves with afferent (sensory) and efferent (motor) neurons.

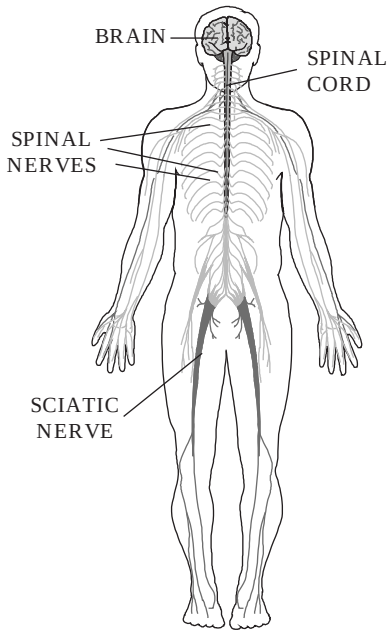


FIGURE 3.27 The central nervous system (CNS) consists of all nervous tissue that is enclosed by bone—that is, the brain and spinal cord—whereas the peripheral nervous system (PNS) consists of the nervous tissue that is not encased by bone.

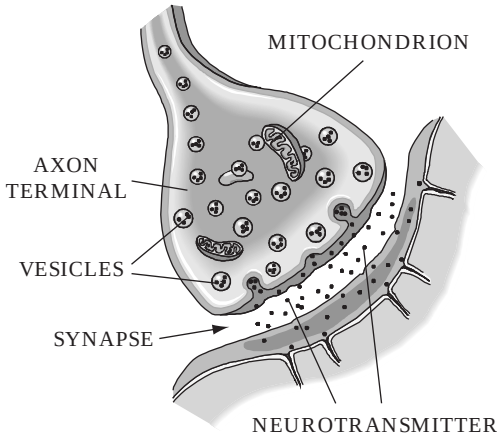


FIGURE 3.28 Following stimulation, vesicles in the axon terminal move to the synapse by means of exocytosis and release neurotransmitters into the space between the axon and the next cell, which could be the dendrite of another neuron, a muscle fiber, or a gland. The neurotransmitters diffuse across the synapse and elicit a response from the adjacent cell.

Clusters of nerve cells located in the CNS are called nuclei, and clusters of nerve cells in the PNS are called ganglion. On the other hand, nucleons located in the PNS have been designated as nerves, while those in the CNS are called tracts.

The nervous system has also been divided into the somatic and autonomic nervous systems. Each of these systems consists of components from both the central and peripheral nervous systems. For example, the somatic peripheral nervous system consists of the sensory neurons, which convey information from receptors for pain, temperature, and mechanical stimuli in the skin, muscles, and joints to the central nervous system, and the motor neurons, which return impulses from the central nervous system to these same areas of the body.

The autonomic nervous system is concerned with the internal meter of the body, including involuntary regulation of smooth muscle, cardiac muscle, and glands and is further divided into the sympathetic and parasympathetic divisions. The sympathetic division causes blood vessels in the viscera and skin to constrict, vessels in the skeletal muscles to dilate, and the heart rate to increase, whereas the parasympathetic division has the opposite effect on the vessels in the viscera and skin, provides no innervation to the skeletal muscles, and causes the heart rate to decrease. Thus, the sympathetic division prepares the body for "fight or flight," while the parasympathetic division returns the body to normal operating conditions.

Brain function is dependent on neuronal circuits. Neurons interconnect in several different types of circuits. In a divergent circuit, each branch in the axon of the presynaptic neuron connects with the dendrite of a different postsynaptic neuron. In a convergent circuit, axons from several presynaptic neurons meet at the dendrite(s) of a single postsynaptic neuron. In a simple feedback circuit, the axon of a neuron connects with the dendrite of an interneuron that connects back with the dendrites of the first neuron. A two-neuron circuit is one in which a sensory neuron synapses directly with a motor neuron, whereas a three-neuron circuit consists of a sensory neuron, an interneuron in the spinal cord, and a motor neuron. Both of these circuits can be found in reflex arcs (Figure 3.29). The reflex arc is a special type of neural circuit that begins with a sensory neuron at a receptor (e.g., a pain receptor in the fingertip) and ends with a motor neuron at an effector (e.g., a skeletal muscle).

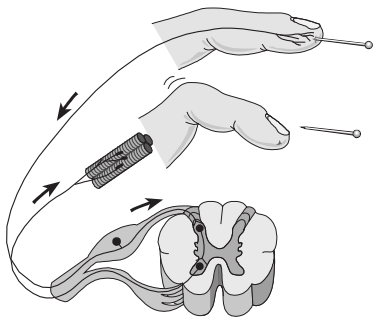


FIGURE 3.29 This reflex arc begins with a sensory neuron in the finger that senses pain when the fingertip is pricked by the pin. An action potential travels from the sensory neuron to an interneuron and then to a motor neuron that synapses with muscle fibers in the finger. The muscle fibers respond to the stimulus by contracting and removing the fingertip from the pin.

Withdrawal reflexes are elicited primarily by stimuli for pain and heat great enough to be painful and are also known as protective or escape reflexes. They allow the body to respond quickly to dangerous situations without taking additional time to send signals to and from the brain and to process the information.

The brain is a large, soft mass of nervous tissue and consists of the cerebrum, the diencephalon, the mesencephalon (midbrain), and the brain stem and cerebellum. The cerebrum (Figure 3.30a), which is divided into two hemispheres, is the largest and most obvious portion of the brain and consists of many convoluted ridges (gyri), narrow grooves (sulci), and deep fissures, which result in a total surface area of about 2.25 m^2. The outer layer of the cerebrum, the cerebral cortex, is composed of gray matter (neurons with unmyelinated

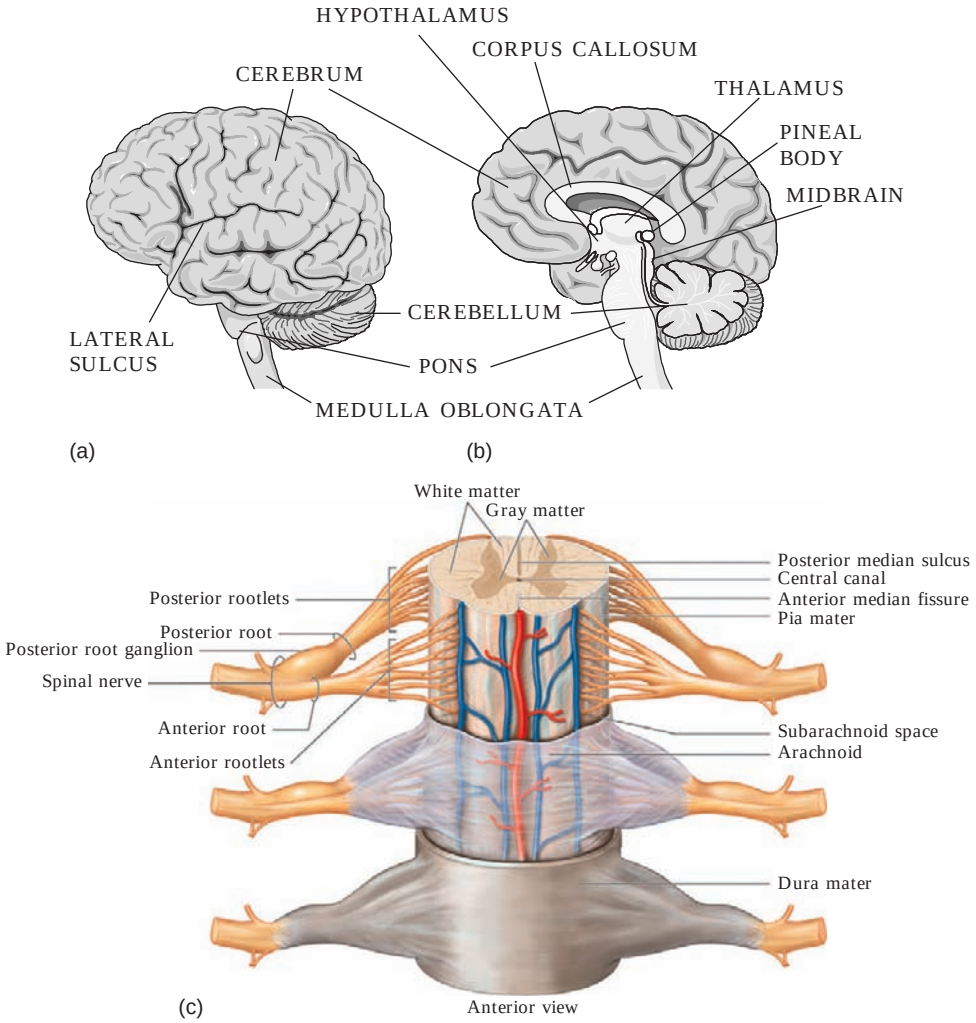


FIGURE 3.30 (a) The exterior surface of the brain. (b) A midsagittal section through the brain. (c) Structure of the spinal cord.

axons) that is 2–4 mm thick and contains over 50 billion neurons and 250 billion glial cells called neuroglia. The thicker inner layer is the white matter that consists of interconnecting groups of myelinated axons that project from the cortex to other cortical areas or from the thalamus (part of the diencephalon) to the cortex. The connection between the two cerebral hemispheres takes place via the corpus callosum (Figure 3.30b). The left side of the cortex controls motor and sensory functions from the right side of the body, whereas the right side controls the left side of the body. Association areas that interpret incoming data or coordinate a motor response are connected to the sensory and motor regions of the cortex.

Fissures divide each cerebral hemisphere into a series of lobes that include the frontal lobe, the parietal lobe, the temporal lobe, and the occipital lobe. Each of these lobes has different functions. The functions of the frontal lobe include initiating voluntary movement of the skeletal muscles, analyzing sensory experiences, providing responses relating to personality, and mediating responses related to memory, emotions, reasoning, judgment, planning, and speaking. The parietal lobe responds to stimuli from cutaneous (skin) and muscle receptors throughout the body. The temporal lobes interpret some sensory experiences, store memories of auditory and visual experiences, and contain auditory centers that receive sensory neurons from the cochlea of the ear. The occipital lobes integrate eye movements by directing and focusing the eye and are responsible for correlating visual images with previous visual experiences and other sensory stimuli. The insula is a deep portion of the cerebrum that lies under the parietal, frontal, and temporal lobes. Little is known about its function, but it seems to be associated with gastrointestinal and other visceral activities.

The diencephalon is the deep part of the brain that connects the midbrain of the brain stem with the cerebral hemispheres. Its main parts include the thalamus, hypothalamus, and epithalamus (Figure 3.30b). The thalamus, the major switchboard of the brain, is involved with sensory and motor systems, general neural background activity, and the expression of emotion and uniquely human behaviors. Due to its two-way communication with areas of the cortex, it is linked with thought, creativity, interpretation and understanding of spoken and written words, and identification of objects sensed by touch. The hypothalamus is involved with integration within the autonomic nervous system, temperature regulation, water and electrolyte balance, sleep-wake patterns, food intake, behavioral responses associated with emotion, endocrine control, and sexual responses. The epithalamus contains the pineal body that is thought to have a neuroendocrine function.

The brain stem connects the brain with the spinal cord and automatically controls vital functions such as breathing. Its principal regions include the midbrain, pons, and medulla oblongota (Figure 3.30b). The midbrain connects the pons and cerebellum with the cerebrum and is located at the upper end of the brain stem. It is involved with visual reflexes, the movement of eyes, focusing of the lenses, and the dilation of the pupils. The pons is a rounded bulge between the midbrain and medulla oblongata that functions with the medulla oblongata to control respiratory functions, acts as a relay station from the medulla oblongata to higher structures in the brain, and is the site of emergence of cranial nerve V. The medulla oblongata is the lowermost portion of the brain stem and connects the pons to the spinal cord. It contains vital centers that regulate heart rate, respiratory rate, constriction and dilation of blood vessels, blood pressure, swallowing, vomiting, sneezing, and coughing.

The cerebellum is located behind the pons and is the second largest part of the brain. It processes sensory information that is used by the motor systems and is involved with

coordinating skeletal muscle contractions and impulses for voluntary muscular movement that originate in the cerebral cortex. The cerebellum is a processing center that is involved with coordination of balance, body positions, and the precision and timing of movements.

The gray matter of the spinal cord is divided into the dorsal and ventral horns. In a human, standing upright, the "dorsal" horn is posterior and the "ventral" horn is anterior. Dorsal horn neurons receive and process sensory information from the skin, while ventral horn neurons participate in the control of skeletal muscle contraction. The gray matter is surrounded by columns (funiculi) of white matter containing ascending and descending axons. The spinal cord communicates with the periphery via the dorsal and ventral root fibers that exit between the bony vertebra. Dorsal root fibers bring information to the spinal cord, and ventral root fibers carry information away from the spinal cord (Figure 3.30c).

3.4.4 The Skeletal System

The average adult skeleton contains 206 bones, but the actual number varies from person to person and decreases with age as some bones become fused. Like the body, the skeletal system is divided into two parts: the axial skeleton and the appendicular skeleton (Figure 3.31). The axial skeleton contains 80 bones (skull, hyoid bone, vertebral column, and thoracic cage), whereas the appendicular skeleton contains 126 (pectoral and pelvic girdles and upper and lower extremities). The skeletal system protects and supports the

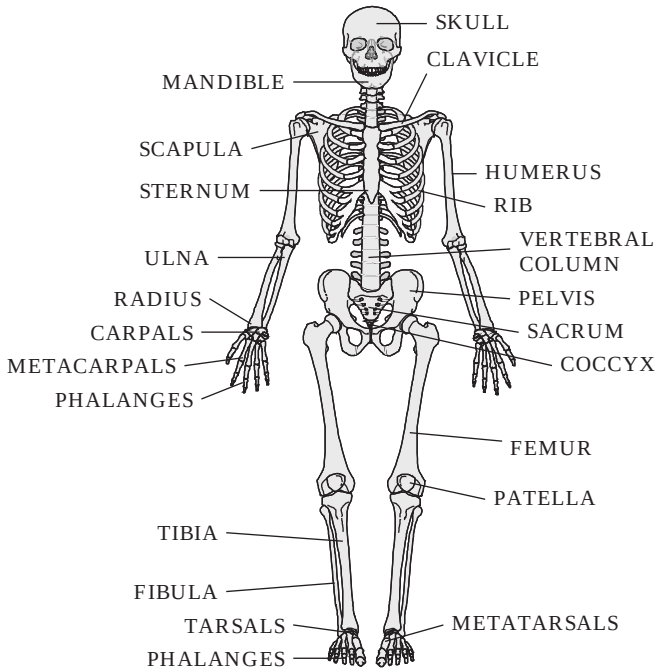


FIGURE 3.31 The skull, hyoid bone (not shown), vertebral column, and thoracic cage (ribs, cartilage, and sternum) make up the axial skeleton, whereas the pectoral (scapula and clavicle) and pelvic girdles and upper and lower extremities make up the appendicular skeleton.

body, helps with movement, produces blood cells, and stores important minerals. It is made up of strong, rigid bones that are composed of specialized connective tissue, bear weight, and form the major supporting elements of the body. Some support also comes from cartilage that is a smooth, firm, resilient, nonvascular type of connective tissue. Since the bones of the skeleton are hard, they protect the organs, such as the brain and abdominal organs, that they surround.

There are 8 cranial bones that support, surround, and protect the brain. Fourteen facial bones form the face and serve as attachments for the facial muscles that primarily move skin rather than bone. The facial bones, except for the lower jaw (mandible), are joined with each other and with the cranial bones. There are 6 auditory ossicles, 3 in each ear, that transmit sound waves from the external environment to the inner ear. The hyoid bone, which is near the skull but is not part of it, is a small U-shaped bone that is located in the neck just below the lower jaw. It is attached to the skull and larynx (voice box) by muscles and ligaments and serves as the attachment for several important neck and tongue muscles.

The vertebral column starts out with approximately 34 bones, but only 26 independent ones are left in the average human adult. There are 7 cervical bones including the axis, which acts as a pivot around which the head rotates, and the atlas, which sits on the axis and supports the "globe" of the head. These are followed by 5 cervical, 12 thoracic, and 5 lumbar vertebrae and then the sacrum and the coccyx. The last two consist of 5 fused vertebrae. The vertebral column supports the weight of and allows movement of the head and trunk, protects the spinal cord, and provides places for the spinal nerves to exit from the spinal cord. There are four major curves (cervical, thoracic, lumbar, and sacral/coccygeal) in the adult vertebral column that allow it to flex and absorb shock. While movement between any two adjacent vertebrae is generally quite limited, the total amount of movement provided by the vertebral column can be extensive. The thoracic cage consists of 12 thoracic vertebrae (which are counted as part of the vertebral column), 12 pairs of ribs and their associated cartilage, and the sternum (breastbone). It protects vital organs and prevents the collapse of the thorax during ventilation.

Bones are classified as long, short, flat, or irregular, according to their shape. Long bones, such as the femur and humerus, are longer than they are wide. Short bones, such as those found in the ankle and wrist, are as broad as they are long. Flat bones, such as the sternum and the bones of the skull, have a relatively thin and flattened shape. Irregular bones do not fit into the other categories and include the bones of the vertebral column and the pelvis.

Bones make up about 18 percent of the mass of the body and have a density of 1.9 g/cm^3. The two types of bone are spongy and compact (cortical). Spongy bone forms the ends (epiphyses) of the long bones and the interior of other bones and is quite porous. Compact bone forms the shaft (diaphysis) and outer covering of bones and has a tensile strength of 120 N/mm^2, compressive strength of 170 N/mm^2, and Young's modulus of 1.8×10^4 N/mm^2. The medullary cavity, a hollow space inside the diaphysis, is filled with fatty yellow marrow or red marrow that contains blood-forming cells.

Bone is a living organ that is constantly being remodeled. Old bone is removed by special cells, osteoclasts, and new bone is deposited by osteoblasts. Bone remodeling occurs during bone growth and in order to regulate calcium availability. The average skeleton is totally remodeled about three times during a person's lifetime. Osteoporosis is a disorder in which old bone is broken down faster than new bone is produced so the resulting bones are weak and brittle.

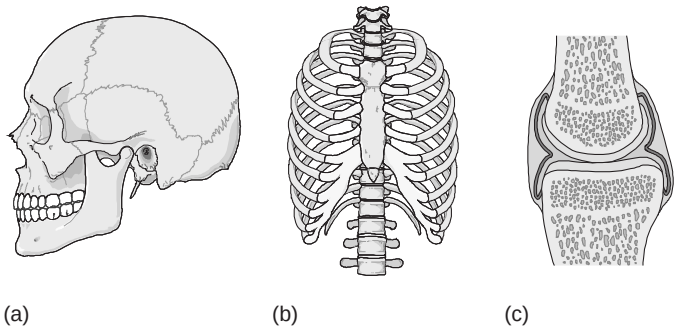


FIGURE 3.32 Bones of the skeletal system are attached to each other at (a) fibrous, (b) cartilaginous, or (c) synovial joints.

The bones of the skeletal system are attached to one another at fibrous, cartilaginous, or synovial joints (Figure 3.32). The articulating bones of fibrous joints are bound tightly together by fibrous connective tissue. These joints can be rigid and relatively immovable to slightly movable. This type of joint includes the suture joints in the skull. Cartilage holds together the bones in cartilaginous joints. These joints allow limited motion in response to twisting or compression and include the joints of the vertebral system and the joints that attach the ribs to the vertebral column and to the sternum. Synovial joints, such as the knee, are the most complex and varied and have fluid-filled joint cavities, cartilage that covers the articulating bones, and ligaments that help hold the joints together.

Synovial joints are classified into six types, depending on their structure and the type of motion they permit. Gliding joints (Figure 3.33) are the simplest type of synovial joint, allow

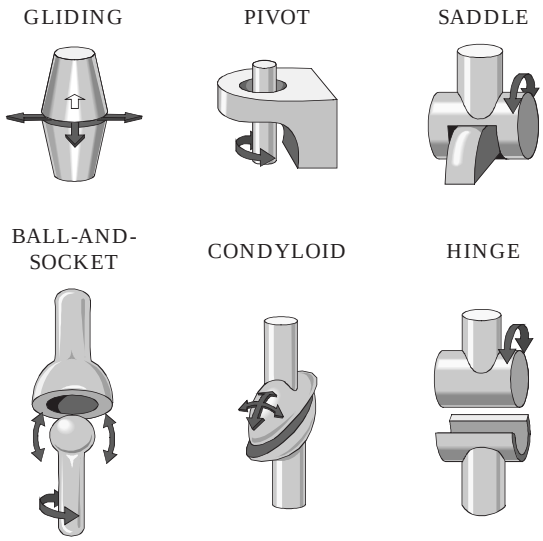


FIGURE 3.33 Synovial joints have fluid-filled cavities and are the most complex and varied types of joints. Each synovial joint is classified into one of six types, depending on its structure and type of motion.

back-and-forth or side-to-side movement, and include the intercarpal articulations in the wrist. Hinge joints, such as the elbow, permit bending in only one plane and are the most common type of synovial joint. The atlas and axis provide an example of a pivot joint that permits rotation. In condyloid articulations, an oval, convex surface of one bone fits into a concave depression on another bone. Condyloid joints, which include the metacarpophalangeal joints (knuckles) of the fingers, permit flexion-extension and rotation and are considered to be biaxial because rotation is limited to two axes of movement. The saddle joint, represented by the joint at the base of the thumb, is a modified condyloid joint that permits movement in several directions (multiaxial). Ball-and-socket joints allow motion in many directions around a fixed center. In these joints, the ball-shaped head of one bone fits into a cuplike concavity of another bone. This multiaxial joint is the most freely movable of all and includes the shoulder and hip joints. Biomedical engineers have helped develop artificial joints that are routinely used as replacements in diseased or injured hips, shoulders, and knees (Figure 3.34).

3.4.5 Muscular System

The muscular system (Figure 3.35) is composed of 600–700 skeletal muscles, depending on whether certain muscles are counted as separate or as pairs, and makes up 40 percent of the body's mass. The axial musculature makes up about 60 percent of the skeletal muscles in the body and arises from the axial skeleton (see Figure 3.31). It positions the head and spinal column and moves the rib cage during breathing. The appendicular musculature moves or stabilizes components of the appendicular skeleton.

The skeletal muscles in the muscular system maintain posture, generate heat to maintain the body's temperature, and provide the driving force that is used to move the bones and joints of the body and the skin of the face. Muscles that play a major role in accomplishing a movement are called prime movers, or agonists. Muscles that act in opposition to a prime

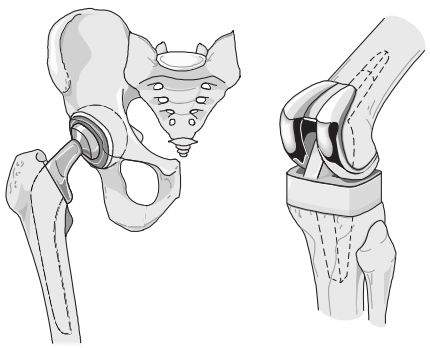


FIGURE 3.34 Diseased or damaged hip (a) and knee (b) joints that are nonfunctional or extremely painful can be replaced by prostheses. Artificial joints can be held in place by a special cement (polymethylmethacrylate [PMMA]) and by bone ingrowth. Special problems occur at the interfaces due to the different elastic moduli of the materials (110 GPa for titanium, 2.2 GPa for PMMA, and 20 GPa for bone).

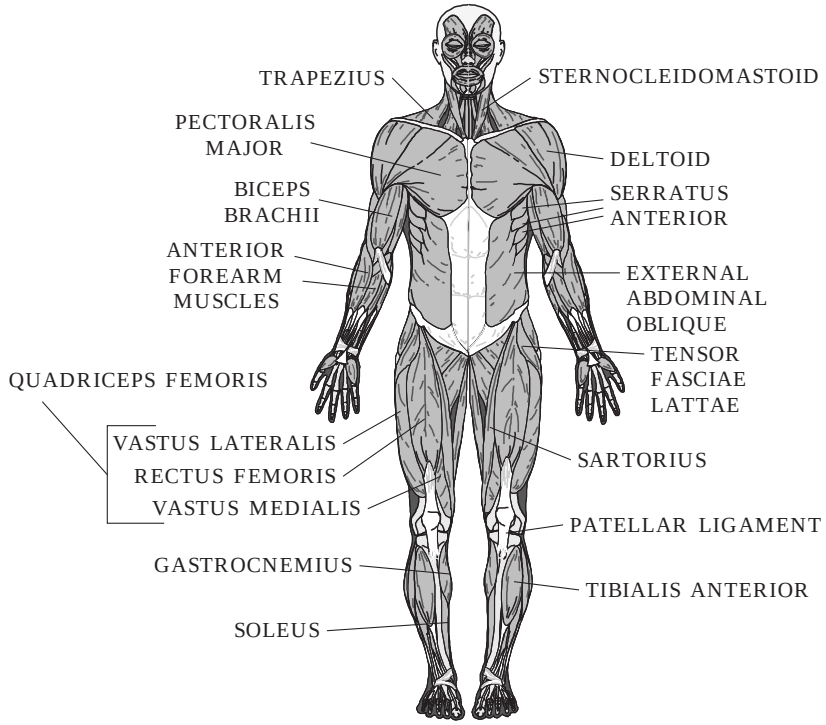


FIGURE 3.35 Some of the major skeletal muscles on the anterior side of the body.

mover are called antagonists, whereas muscles that assist a prime mover in producing a movement are called synergists. The continual contraction of some skeletal muscles helps maintain the body's posture. If all of these muscles relax, which happens when a person faints, the person collapses.

A system of levers, which consists of rigid lever arms that pivot around fixed points, is used to move skeletal muscle (Figure 3.36). Two different forces act on every lever: the weight to be moved—that is, the resistance to be overcome—and the pull or effort applied—that is, the applied force. Bones act as lever arms, and joints provide a fulcrum. The resistance to be overcome is the weight of the body part that is moved, and the applied force is generated by the contraction of a muscle or muscles at the insertion, the point of attachment of a muscle to the bone it moves. An example of a first-class lever, one in which the fulcrum is between the force and the weight, is the movement of the facial portion of the head when the face is tilted upward. The fulcrum is formed by the joint between the atlas and the occipital bone of the skull, and the vertebral muscles inserted at the back of the head generate the applied force that moves the weight, the facial portion of the head. A second-class lever is one in which the weight is between the force and the fulcrum. This can be found in the body when a person stands on "tip toe." The ball of the foot is the fulcrum, and the applied force is generated by the calf muscles on the back of the leg. The weight that is moved is that of the whole body. A third-class lever is one in which the force is between the weight and the fulcrum. When a person has a bent elbow and holds

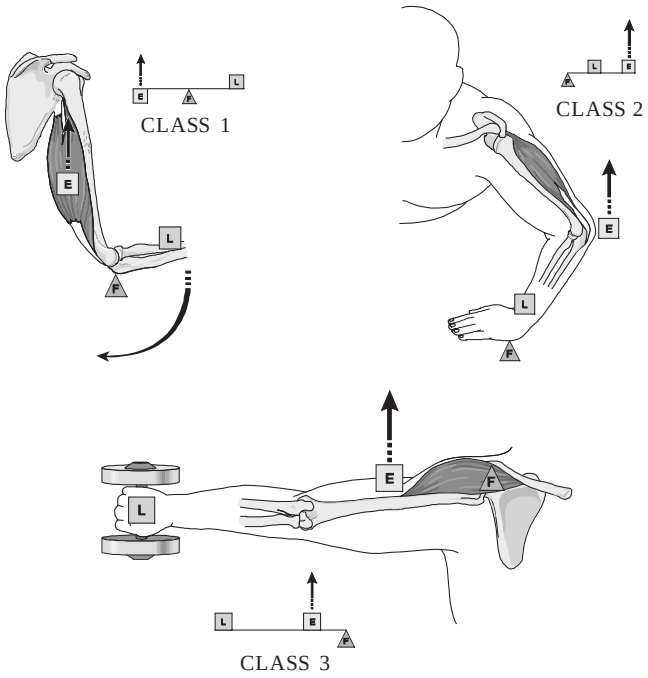


FIGURE 3.36 Depending on the muscle in use, the location of the load, and the location of the fulcrum, the humerus can act as a class 1 lever, a class 2 lever, or a class 3 lever.

a ball in front of the body, the applied force is generated by the contraction of the biceps brachii muscle. The weight to be moved includes the ball and the weight of the forearm and hand, and the elbow acts as the fulcrum.

The three types of muscle tissue—cardiac, skeletal, and smooth—share four important characteristics: contractility, the ability to shorten; excitability, the capacity to receive and respond to a stimulus; extensibility, the ability to be stretched; and elasticity, the ability to return to the original shape after being stretched or contracted. Cardiac muscle tissue is found only in the heart, whereas smooth muscle tissue is found within almost every other organ, where it forms sheets, bundles, or sheaths around other tissues. Skeletal muscles are composed of skeletal muscle tissue, connective tissue, blood vessels, and nervous tissue.

Each skeletal muscle is surrounded by a layer of connective tissue (collagen fibers) that separates the muscle from surrounding tissues and organs. These fibers come together at the end of the muscle to form tendons, which connect the skeletal muscle to bone, to skin (face), or to the tendons of other muscles (hand). Other connective tissue fibers divide the skeletal muscles into compartments called fascicles that contain bundles of muscle fibers. Within each fascicle, additional connective tissue surrounds each skeletal muscle fiber and ties adjacent ones together. Each skeletal muscle fiber has hundreds of nuclei just beneath the cell membrane. Multiple nuclei provide multiple copies of the genes that direct the production of enzymes and structural proteins needed for normal contraction so contraction can occur faster.

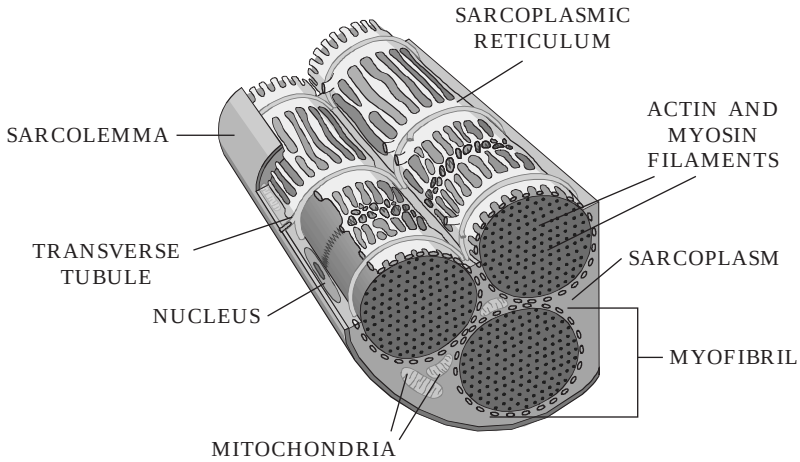


FIGURE 3.37 Skeletal muscles are composed of muscle fascicles that are composed of muscle fibers such as the one shown here. Muscle fibers have hundreds of nuclei just below the plasma membrane—the sarcolemma. Transverse tubules extend into the sarcoplasm, the cytoplasm of the muscle fiber, and are important in the contraction process because they deliver action potentials that result in the release of stored calcium ions. Calcium ions are needed to create active sites on actin filaments so cross-bridges can be formed between actin and myosin and the muscle can contract.

In muscle fibers, the plasma membrane is called the sarcolemma, and the cytoplasm is called the sarcoplasm (Figure 3.37). Transverse tubules (T tubules) begin at the sarcolemma and extend into the sarcoplasm at right angles to the surface of the sarcolemma. The T tubules, which play a role in coordinating contraction, are filled with extracellular fluid and form passageways through the muscle fiber. They make close contact with expanded chambers, cisternae, of the sarcoplasmic reticulum, a specialized form of the ER. The cisternae contain high concentrations of calcium ions that are needed for contraction to occur.

The sarcoplasm contains cylinders 1 or 2 μm in diameter that are as long as the entire muscle fiber and are called myofibrils. The myofibrils are attached to the sarcolemma at each end of the cell and are responsible for muscle fiber contraction. Myofilaments—protein filaments consisting of thin filaments (primarily actin) and thick filaments (mostly myosin)—are bundled together to make up myofibrils. Repeating functional units of myofilaments are called sarcomeres (Figure 3.38). The sarcomere is the smallest functional unit of the muscle fiber and has a resting length of about 2.6 μm. The thin filaments are attached to dark bands, called Z lines, which form the ends of each sarcomere. Thick filaments containing double-headed myosin molecules lie between the thin ones. It is this overlap of thin and thick filaments that gives skeletal muscle its banded, striated appearance. The I band is the area in a relaxed muscle fiber that just contains actin filaments, and the H zone is the area that just contains myosin filaments. The H zone and the area in which the actin and myosin overlap form the A band.

When a muscle contracts, myosin molecules in the thick filaments form cross-bridges at active sites in the actin of the thin filaments and pull the thin filaments toward the center of

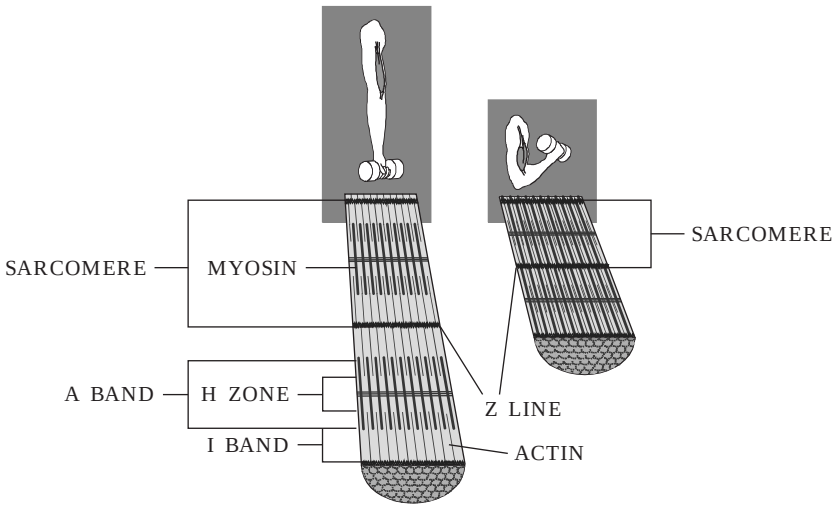


FIGURE 3.38 The sarcomere is the basic functional unit of skeletal muscles and extends from one Z line to the next. Actin filaments are attached to the Z lines and extend into the A band, where they overlap with the thicker myosin filaments. The H zone is the portion of the A band that contains no overlapping actin filaments. When the muscle changes from its extended, relaxed position (left panel) to its contracted state (right panel), the myosin filaments use cross-bridges to slide past the actin filaments and bring the Z lines closer together. This results in shorter sarcomeres and a contracted muscle.

the sarcomere. The cross-bridges are then released and reformed at a different active site further along the thin filament. This results in a motion that is similar to the hand-over-hand motion that is used to pull in a rope. This action, the sliding filament mechanism, is driven by ATP energy and results in shortening of the muscle. Shortening of the muscle components (contraction) results in bringing the muscle's attachments (e.g., bones) closer together (Figure 3.38).

Muscle fibers have connections with nerves. Sensory nerve endings are sensitive to length, tension, and pain in the muscle and send impulses to the brain via the spinal cord, whereas motor nerve endings receive impulses from the brain and spinal cord that lead to excitation and contraction of the muscle. Each motor axon branches and supplies several muscle fibers. Each of these axon branches loses its myelin sheath and splits up into a number of terminals that make contact with the surface of the muscle. When the nerve is stimulated, vesicles in the axon terminals release a neurotransmitter, acetylcholine, into the synapse between the neuron and the muscle. Acetylcholine diffuses across the synapse and binds to receptors in a special area, the motor end plate, of the sarcolemma. This causes the sodium channels in the sarcolemma to open up, and an action potential is produced in the muscle fiber. The resulting action potential spreads over the entire sarcolemmal surface and travels down all of the T tubules, where it triggers a sudden massive release of calcium by the cisternae. Calcium triggers the production of active sites on the thin filaments so cross-bridges with myosin can form and contraction occurs. Acetylcholinesterase breaks down the acetylcholine while the contraction process is under way so the original relatively low permeability of the sarcolemma to sodium is restored.

A motor unit consists of a motor neuron and the muscle fibers that it innervates. All the muscle fibers in a single motor unit contract at the same time, whereas muscle fibers in the same muscle but belonging to different motor units may contract at different times. When a contracted muscle relaxes, it returns to its original (resting) length if another contracting muscle moves it or if it is acted upon by gravity. During relaxation, ATP is expended to move calcium back to the cisternae. The active sites that were needed for cross-bridge formation become covered so actin and myosin can no longer interact. When the cross-bridges disappear, the muscle returns to its resting length—that is, it relaxes.

The human body contains two different types of skeletal muscle fibers: fast and slow. Fast fibers can contract in 10 ms or less following stimulation and make up most of the skeletal muscle fibers in the body. They are large in diameter and contain densely packed myofibrils, large glycogen reserves (used to produce ATP), and relatively few mitochondria. These fibers produce powerful contractions that use up massive amounts of ATP and fatigue (can no longer contract in spite of continued neural stimulation) rapidly. Slow fibers take about three times as long to contract as fast fibers. They can continue to contract for extended periods of time because they contain (1) a more extensive network of capillaries, so they can receive more oxygen; (2) a special oxygen-binding molecule called myoglobin; and (3) more mitochondria, which can produce more ATP than fast fibers. Muscles contain different amounts of slow and fast fibers. Those that are dominated by fast fibers (e.g., chicken breast muscles) appear white, while those that are dominated by slow fibers (e.g., chicken legs) appear red. Most human muscles appear pink because they contain a mixture of both. Genes determine the percentage of fast and slow fibers in each muscle, but the ability of fast muscle fibers to resist fatigue can be increased through athletic training.

3.5 HOMEOSTASIS

Organ systems work together to maintain a constant internal environment within the body. Homeostasis is the process by which physical and chemical conditions within the internal environment of the body are maintained within tolerable ranges even when the external environment changes. Body temperature, blood pressure, and breathing and heart rates are some of the functions that are controlled by homeostatic mechanisms that involve several organ systems working together.

Extracellular fluid—the fluid that surrounds and bathes the body's cells—plays an important role in maintaining homeostasis. It circulates throughout the body and carries materials to and from the cells. It also provides a mechanism for maintaining optimal temperature and pressure levels, the proper balance between acids and bases, and concentrations of oxygen, carbon dioxide, water, nutrients, and many of the chemicals that are found in the blood.

Three components—sensory receptors, integrators, and effectors—interact to maintain homeostasis (Figure 3.39). Sensory receptors, which may be cells or cell parts, detect stimuli—that is, changes to their environment—and send information about the stimuli to integrators. Integrators are control points that pull together information from one or more sensory receptors. Integrators then elicit a response from effectors. The brain is an

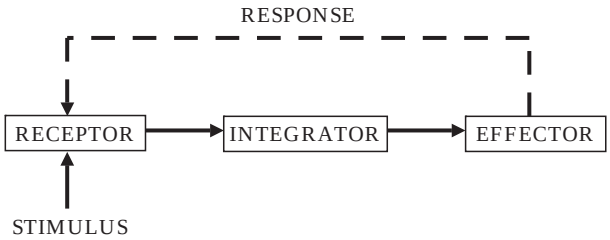


FIGURE 3.39 Feedback mechanisms are used to help maintain homeostatis. A stimulus is received by a receptor that sends a signal (messenger) to an effector or to an integrator that sends a signal to an effector. The effector responds to the signal. The response feeds back to the receptor and modifies the effect of the stimulus. In negative feedback, the response subtracts from the effect of the stimulus on the receptor. In positive feedback, the response adds to the effect of the stimulus on the receptor.

integrator that can send messages to muscles or glands or both. The messages result in some type of response from the effectors. The brain receives information about how parts of the body are operating and can compare this to information about how parts of the body should be operating.

Positive feedback mechanisms are those in which the initial stimulus is reinforced by the response. There are very few examples of this in the human body, since it disrupts homeostasis. Childbirth provides one example. Pressure from the baby's head in the birth canal stimulates receptors in the cervix, which send signals to the hypothalamus. The hypothalamus responds to the stimulus by releasing oxytocin, which enhances uterine contractions. Uterine contractions increase in intensity and force the baby further into the birth canal, which causes additional stretching of the receptors in the cervix. The process continues until the baby is born, the pressure on the cervical stretch receptors ends, and the hypothalamus is no longer stimulated to release oxytocin.

Negative feedback mechanisms result in a response that is opposite in direction to the initiating stimulus. For example, receptors in the skin and elsewhere in the body detect the body's temperature. Temperature information is forwarded to the hypothalamus in the brain, which compares the body's current temperature to what the temperature should be (approximately 37°C). If the body's temperature is too low, messages are sent to contract the smooth muscles in blood vessels near the skin (reducing the diameter of the blood vessels and the heat transferred through the skin), to skeletal muscles to start contracting rapidly (shivering), and to the arrector pili muscles (see Figure 3.16a) to erect the hairs and form "goose bumps." The metabolic activity of the muscle contractions generates heat and warms the body. If the body's temperature is too high, messages are sent to relax the smooth muscles in the blood vessels near the skin (increasing the diameter of the blood vessels and the amount of heat transferred through the skin) and to sweat glands to release moisture and thus increase evaporative cooling of the skin. When the temperature of circulating blood changes to such an extent in the appropriate direction that it reaches the set point of the system, the hypothalamus stops sending signals to the effector muscles and glands.

Another example of a negative feedback mechanism in the body involves the regulation of glucose in the bloodstream by clusters of cells, the pancreatic islets (Figure 3.40). There are between 2×10^5 and 2×10^6 pancreatic islets scattered throughout the adult

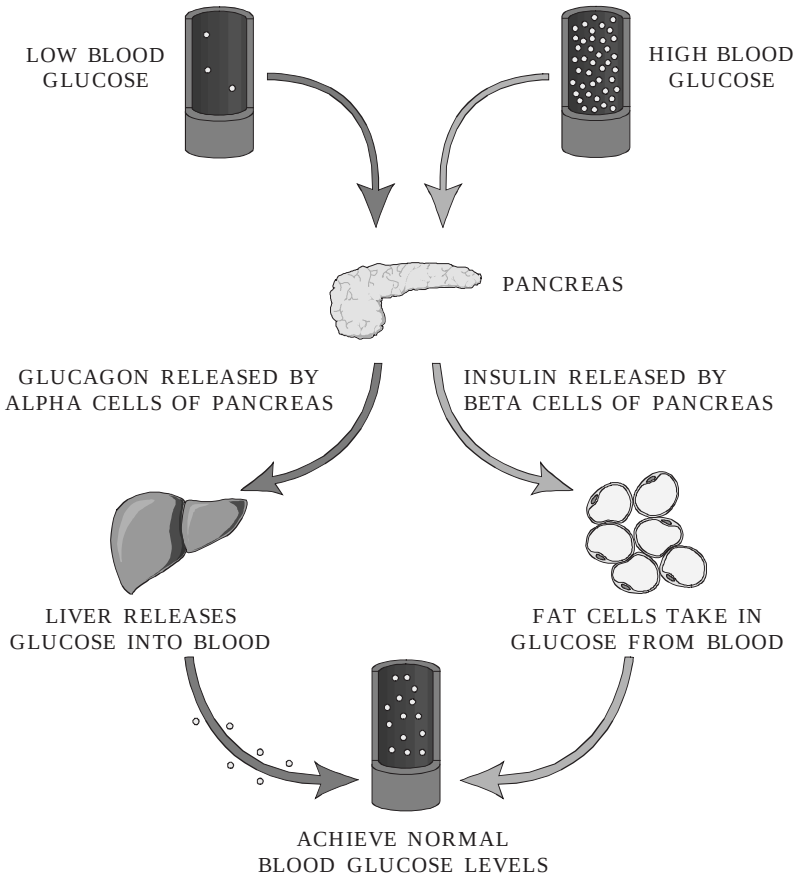


FIGURE 3.40 Two negative feedback mechanisms help control the level of glucose in the blood. When blood glucose levels are higher than the body's set point (stimulus), beta cells (receptors) in the pancreatic islets produce insulin (messenger), which facilitates glucose transport across plasma membranes and enhances the conversion of glucose into glycogen for storage in the liver (effector). This causes the level of glucose in the blood to drop. When the level equals the body's set point, the beta cells stop producing insulin. When blood glucose levels are lower than the body's set point (stimulus), alpha cells (receptors) in the pancreatic islets produce glucagon (messenger), which stimulates the liver (effector) to convert glycogen into glucose. This causes the level of glucose in the blood to increase. When the level equals the body's set point, the alpha cells stop producing glucagon.

pancreas. When glucose levels are high, beta cells in the islets produce insulin, which facilitates glucose transport across plasma membranes and into cells and enhances the conversion of glucose into glycogen that is stored in the liver. During periods of fasting, or whenever the concentration of blood glucose drops below normal (70–110 mg/dl), alpha cells produce glucagon, which stimulates the liver to convert glycogen into glucose and the formation of glucose from noncarbohydrate sources such as amino acids and lactic acid. When glucose levels return to normal, the effector cells in the pancreatic islets stop producing their respective hormone—that is, insulin or glucagon. Some biomedical

engineers are working on controlled drug delivery systems that can sense blood glucose levels and emulate the responses of the pancreatic islet cells, whereas other biomedical engineers are trying to develop an artificial pancreas that would effectively maintain appropriate blood glucose levels.

3.6 EXERCISES

1. Using as many appropriate anatomical terms as apply, write sentences that describe the positional relationship between your mouth and (a) your left ear, (b) your nose, and (c) the big toe on your right foot.
2. Using as many appropriate anatomical terms as apply, describe the position of the stomach in the body and its position relative to the heart.
3. Search the Internet to find a transverse section of the body that was imaged using computerized tomography (CT) or magnetic resonance imaging (MRI). Print the image and indicate its web address.
4. Search the Internet to find a frontal section of the body that was imaged using CT or MRI. Print the image and indicate its web address.
5. Name and give examples of the four classes of biologically important organic compounds. What are the major functions of each of these groups?
6. What are the molarity and osmolarity of a 1-liter solution that contains half a mole of calcium chloride? How many molecules of chloride would the solution contain?
7. Consider a simple model cell, such as the one in Figure 3.6, that consists of cytoplasm and a plasma membrane. The cell's initial volume is 2 nl and contains 0.2 M protein. The cell is placed in a large volume of 0.05 M $CaCl_2$. Neither Ca^{++} nor Cl^- can cross the plasma membrane and enter the cell. Is the 0.05 M $CaCl_2$ solution hypotonic, isotonic, or hypertonic relative to the osmolarity inside the cell? Describe what happens to the cell as it achieves equilibrium in this new environment. What will be the final osmolarity of the cell? What will be its final volume?
8. What does the principle of electrical neutrality mean in terms of the concentration of ions within a cell?
9. Consider the same model cell that was used in problem 7, but instead of being placed in 0.05 M $CaCl_2$, the cell is placed in 0.2 M urea. Unlike Ca^{++} and Cl^-, urea can cross the plasma membrane and enter the cell. Describe what happens to the cell as it achieves equilibrium in this environment. What will be the final osmolarity of the cell? What will be its final volume?
10. Briefly describe the path that a protein, such as a hormone, that is manufactured on the rough ER would take in order to leave the cell.
11. What major role do mitochondria have in the cell? Why might it be important to have this process contained within an organelle?
12. List and briefly describe three organelles that provide structural support and assist with cell movement.
13. Find a location on the Internet that describes the Human Genome Project. Print its home page and indicate its web address. Find and print an ideogram of a chromosome that shows a gene that causes cystic fibrosis.

Continued

14. Briefly describe the major differences between replication and transcription.
15. Describe how the hereditary information contained in genes within the cell's DNA is expressed as proteins that direct the cell's activities.
16. Six different codons code for leucine, while only one codes for methionine. Why might this be important for regulating translation and producing proteins?
17. Insulin (Figure 3.41) was the first protein to be sequenced biochemically. Assuming that there were no introns involved in the process, what are the possible DNA sequences that produced the last four amino acids in the molecule?

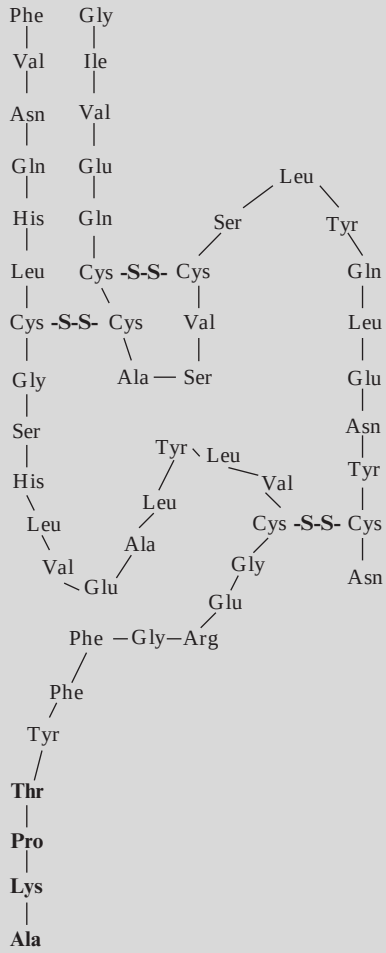


FIGURE 3.41 Bovine insulin consists of two polypeptide chains that are joined by two disulfide bonds (-S-S-). Hydrogen bonds also exist between the chains and between segments of the same chain. The three-letter names stand for different amino acids (see Table 3.1).

18. Copy the title page and abstract of five peer-reviewed journal articles that discuss engineering applications for five different organ systems in the body (one article per organ system). Review articles, conference proceeding papers, copies of keynote addresses and other speeches, book chapters, articles from the popular press and newspapers, and editorials are not acceptable. Good places to look are the *Annals of Biomedical Engineering*, the *IEEE Transactions on Biomedical Engineering*, the *IEEE Engineering in Medicine and Biology Magazine*, and *Medical and Biological Engineering and Computing*. What information in the article indicates that it was peer-reviewed?
19. Trace the path of a single red blood cell from a capillary bed in your right hand to the capillary beds of your right lung and back. What gases are exchanged? Where are they exchanged during this process?
20. Draw and label a block diagram of pulmonary and systemic blood flow that includes the chambers of the heart, valves, major veins and arteries that enter and leave the heart, the lungs, and the capillary bed of the body. Use arrows to indicate the direction of flow through each component.
21. Find an example of an ECG representing normal sinus rhythm on the Internet and use it to demonstrate how heart rate is determined.
22. Why are R waves (Figure 3.22) used to determine heart rate rather than T waves?
23. How could the stroke volume be determined if a thermal dilution technique is used to determine cardiac output?
24. What would be the pulse pressure and mean arterial pressure for a hypertensive person with a systolic pressure of 145 mmHg and a diastolic pressure of 98 mmHg?
25. The total lung capacity of a patient is 5.5 liters. Find the patient's inspiratory reserve volume if the patient's vital capacity was 4.2 liters, the tidal volume was 500 ml, and the expiratory reserve volume was 1.2 liters.
26. What would you need to know or measure in order to determine the residual volume of the patient described in Example Problem 3.10?
27. Briefly describe the functions and major components of the central, peripheral, somatic, automatic, sympathetic, and parasympathetic nervous systems. Which ones are subsets of others?
28. Explain how sarcomeres shorten and how that results in muscle contraction.
29. How do the muscular and skeletal systems interact to produce movement?
30. Draw a block diagram to show the negative feedback mechanisms that help regulate glucose levels in the blood. Label the inputs, sensors, integrators, effectors, and outputs.

Suggested Readings

B.H. Brown, R.H. Smallwood, D.C. Barber, P.V. Lawford, D.R. Hose, Medical Physics and Biomedical Engineering, Institute of Physics Publishing, Bristol and Philadelphia, 1999.

G.M. Cooper, The Cell—A Molecular Approach, second ed., ASM Press, Washington, D.C., 2000.

S. Deutsch, A. Deutsch, Understanding the Nervous System: An Engineering Perspective, IEEE Press, New York, 1993.

S.I. Fox, Human Physiology, eighth ed., McGraw-Hill, Boston, 2004.

W.J. Germann, C.L. Stanfield, Principles of Human Physiology, second ed., Pearson Benjamin Cummings, San Francisco, 2005.

A.C. Guyton, Basic Neuroscience. Anatomy & Physiology, W. B. Saunders Company, Philadelphia, 1991.
A.C. Guyton, J.E. Hall, Textbook of Medical Physiology, tenth ed., W. B. Saunders Company, Philadelphia, 2000.
F.M. Harold, The Way of the Cell—Molecules, Organisms and the Order of Life, Oxford University Press, Inc., New York, 2001.
G. Karp, Cell and Molecular Biology—Concepts and Experiments, third ed., John Wiley & Sons, Inc., New York, 2002.
A.M. Katz, Physiology of the Heart, Raven Press, New York, 1986.
R.D. Keynes, D.J. Aidley, Nerve & Muscle, second ed., Cambridge University Press, Cambridge, 1991.
A.R. Leff, P.T. Schumacker, Respiratory Physiology. Basics and Applications, W. B. Saunders Company, Philadelphia, 1993.
H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, Molecular Cell Biology, fourth ed., W. H. Freeman and Company, New York, 2000.
F.H. Martini, Fundamentals of Anatomy & Physiology, fifth ed., Prentice Hall, Upper Saddle River, NJ, 2001.
G.G. Matthews, Cellular Physiology of Nerve and Muscle, Blackwell Scientific Publications, Boston, 1991.
G.H. Pollack, Cells, Gels and the Engines of Life—A New Unifying Approach to Cell Function, Ebner & Sons, Seattle, WA, 2001.
R. Rhoades, R. Pflanzer, Human Physiology, fourth ed., Thomson Learning, Inc., Pacific Grove, CA, 2003.
D.U. Silverthorn, Human Physiology—An Integrated Approach, third ed., Pearson Benjamin Cummings, San Francisco, 2004.
A. Tözeren, S.W. Byers, New Biology for Engineers and Computer Scientists, Pearson Education, Inc., Upper Saddle River, NJ, 2004.
K.M. Van De Graaff, S.I. Fox, K.M. LaFleur, Synopsis of Human Anatomy & Physiology, Wm. C. Brown Publishers, Dubuque, IA, 1997.
J.B. West, Respiratory Physiology—The Essentials, fourth ed., Williams & Wilkins, Baltimore, 1990.
E.P. Widmaier, H. Raff, K.T. Strang, Vander, Sherman, & Luciano's Human Physiology—The Mechanisms of Body Function, McGraw-Hill, Boston, 2004.

CHAPTER

4

Biomechanics

Joseph L. Palladino, PhD, and Roy B. Davis III, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Understand the application of engineering kinematic relations to biomechanical problems.
- Understand the application of engineering kinetic relations to biomechanical problems.
- Understand the application of engineering mechanics of materials to biological structures.
- Use MATLAB to write and solve biomechanical static and dynamic equations.
- Use Simulink to study viscoelastic properties of biological tissues.
- Understand how kinematic equations of motion are used in clinical analysis of human gait.
- Understand how kinetic equations of motion are used in clinical analysis of human gait.
- Explain how biomechanics applied to human gait is used to quantify pathological conditions, to suggest surgical and clinical treatments, and to quantify their effectiveness.
- Understand basic rheology of biological fluids.
- Understand the development of models that describe blood vessel mechanics.

- Understand basic heart mechanics.
- Explain how biomechanics applied to the cardiovascular system is used to quantify the effectiveness of the heart as a pump, to study heart-vessel interaction, and to develop clinical applications.

4.1 INTRODUCTION

Biomechanics combines engineering and the life sciences by applying principles from classical mechanics to the study of living systems. This relatively new field covers a broad range of topics, including strength of biological materials, biofluid mechanics in the cardiovascular and respiratory systems, material properties and interactions of medical implants and the body, heat and mass transfer into biological tissues, biocontrol systems regulating metabolism or voluntary motion, and kinematics and kinetics applied to study human gait. The great breadth of the field of biomechanics arises from the complexities and variety of biological organisms and systems.

The goals of this chapter are twofold: to apply basic engineering principles to biological structures and to develop clinical applications. Section 4.2 provides a review of concepts from introductory statics and dynamics. Section 4.3 presents concepts from mechanics of material that are fundamental for engineers and accessible to those with only a statics/dynamics background. Section 4.4 introduces viscoelastic complexities characteristic of biological materials, with the concepts further applied in Section 4.5. The last two sections bring all of this information together in two "real-world" biomechanics applications: human gait analysis and cardiovascular dynamics. The human body is a complex machine, with the skeletal system and ligaments forming the framework and the muscles and tendons serving as the motors and cables. Human gait biomechanics may be viewed as a structure (skeleton) composed of levers (bones) with pivots (joints) that move as the result of net forces produced by pairs of agonist and antagonist muscles, a concept with origins as early as 1680, as depicted in Figure 4.1 from Borelli's *De Motu Animalium* (*On the Motion of Animals*). Consequently, the strength of the structure and the action of muscles will be of fundamental importance. Using a similar functional model, the cardiovascular system may be viewed as a complex pump (heart) pumping a complex fluid (blood) into a complex set of pipes (blood vessels). An extensive suggested reading list for both gait and cardiovascular dynamics permits the reader to go beyond the very introductory nature of this textbook.

The discipline of mechanics has a long history. For lack of more ancient records, the history of mechanics starts with the ancient Greeks and Aristotle (384–322 BC). Hellenic mechanics devised a correct concept of statics, but those of dynamics, fundamental in living systems, did not begin until the end of the Middle Ages and the beginning of the modern era. Starting in the sixteenth century, the field of dynamics advanced rapidly with work by Kepler, Galileo, Descartes, Huygens, and Newton. Dynamic laws were subsequently codified by Euler, LaGrange, and LaPlace (see *A History of Mechanics* by Dugas).

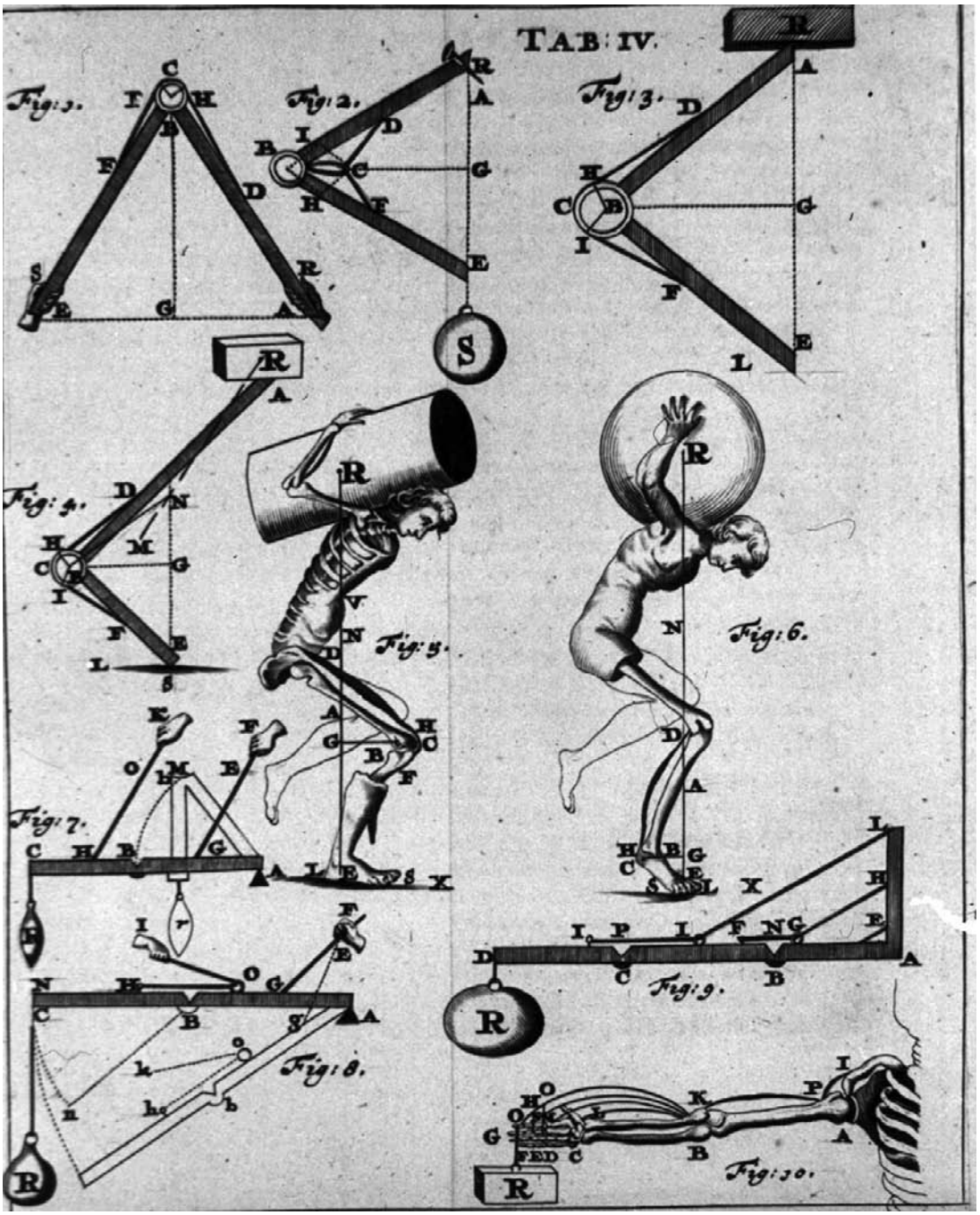


FIGURE 4.1 Plate reproduced from Borelli's *De Motu Animalium*, showing animal (human) motion resulting from the action of muscle pairs on bones, serving as levers, allowed to move at joints. *From* Images from the History of Medicine (IHM), *National Library of Medicine, National Institutes of Health, http://www.nlm.nih.gov/hmd/ihm/.*

In Galileo's *Two New Sciences* (1638), the subtitle *Attenenti all Mecanica & i Movimenti Locali* (*Pertaining to Mechanics and Local Motions*) refers to force, motion, and strength of materials. Since then, "mechanics" has been extended to describe the forces and motions of any system, ranging from quanta, atoms, molecules, gases, liquids, solids, structures, stars, and galaxies. The biological world is consequently a natural object for the study of mechanics.

The relatively new field of biomechanics applies mechanical principles to the study of living systems. The eminent professor of biomechanics Dr. Y. C. Fung describes the role of biomechanics in biology, physiology, and medicine as follows:

> Physiology can no more be understood without biomechanics than an airplane can without aerodynamics. For an airplane, mechanics enables us to design its structure and predict its performance. For an organ, biomechanics helps us to understand its normal function, predict changes due to alteration, and propose methods of artificial intervention. Thus, diagnosis, surgery, and prosthesis are closely associated with biomechanics.[1]

Clearly, biomechanics is essential to assessing and improving human health.

The following is a brief list of biomechanical milestones, especially those related to the topics in this chapter:

- **Galen of Pergamon (129–199)** Published extensively in medicine, including *De Motu Muscularum* (*On the Movements of Muscles*). He realized that motion requires muscle contraction.
- **Leonardo da Vinci (1452–1519)** Made the first accurate descriptions of ball-and-socket joints, such as the shoulder and hip, calling the latter the "*polo dell'omo*" (pole of man). His drawings depicted mechanical force acting along the line of muscle filaments.
- **Andreas Vesalius (1514–1564)** Published *De Humani Corporis Fabrica* (*The Fabric of the Human Body*). Based on human cadaver dissections, his work led to a more accurate anatomical description of human musculature than Galen's and demonstrated that motion results from the contraction of muscles that shorten and thicken.
- **Galileo Galilei (1564–1642)** Studied medicine and physics, integrated measurement and observation in science, and concluded that mathematics is an essential tool of science. His analyses included the biomechanics of jumping and the gait analysis of horses and insects, as well as dimensional analysis of animal bones.
- **Santorio Santorio (1561–1636)** Used Galileo's method of measurement and analysis and found that the human body changes weight with time. This observation led to the study of metabolism and, thereby, ushered in the scientific study of medicine.
- **William Harvey (1578–1657)** Developed an experimental basis for the modern circulation concept of a closed path between arteries and veins. The structural basis, the capillary, was discovered by Malpighi in 1661.
- **Giovanni Borelli (1608–1679)** A mathematician who studied body dynamics, muscle contraction, animal movement, and motion of the heart and intestines. He published *De Motu Animalium* (*On the Motion of Animals*) in 1680.

[1] *Biomechanics: Mechanical Properties of Living Tissues,* 2nd ed., Y. C. Fung, 1993.

- **Jan Swammerdam (1637–1680)** Introduced the nerve-muscle preparation, stimulating muscle contraction by pinching the attached nerve in the frog leg. He also showed that muscles contract with little change in volume, refuting the previous belief that muscles contract when "animal spirits" fill them, causing bulging.
- **Robert Hooke (1635–1703)** Devised Hooke's Law, relating the stress and elongation of elastic materials, and used the term "cell" in biology.
- **Isaac Newton (1642–1727)** Not known for biomechanics work, but he developed calculus, the classical laws of motion, and the constitutive equation for viscous fluid, all of which are fundamental to biomechanics.
- **Nicholas André (1658–1742)** Coined the term "orthopaedics" at the age of 80 and believed that muscular imbalances cause skeletal deformities.
- **Stephen Hales (1677–1761)** Was likely the first to measure blood pressure, as described in his book *Statistical Essays: Containing Haemostaticks, or an Account of Some Hydraulick and Hydrostatical Experiments Made on the Blood and Blood-Vessels of Animals; etc.*, in 1733.
- **Leonard Euler (1707–1783)** Generalized Newton's laws of motion to continuum representations that are used extensively to describe rigid body motion, and studied pulse waves in arteries.
- **Thomas Young (1773–1829)** Studied vibrations and voice, wave theory of light and vision, and devised Young's modulus of elasticity.
- **Ernst Weber (1795–1878) and Eduard Weber (1806–1871)** Published *Die Mechanik der meschlichen Gerwerkzeuge* (*On the Mechanics of the Human Gait Tools*) in 1836, pioneering the scientific study of human gait.
- **Hermann von Helmholtz (1821–1894)** Studied an immense array of topics, including optics, acoustics, thermodynamics, electrodynamics, physiology, and medicine, including ophthalmoscopy, fluid mechanics, nerve conduction speed, and the heat of muscle contraction.
- **Etienne Marey (1830–1904)** Analyzed the motion of horses, birds, insects, fish, and humans. His inventions included force plates to measure ground reaction forces and the "*Chronophotographe a pellicule*," or motion picture camera.
- **Wilhelm Braune and Otto Fischer (research conducted from 1895–1904)** Published *Der Gang des Menschen* (*The Human Gait*), containing the mathematical analysis of human gait and introducing methods still in use. They invented "cyclography" (now called interrupted-light photography with active markers), pioneered the use of multiple cameras to reconstruct 3-D motion data, and applied Newtonian mechanics to estimate joint forces and limb accelerations.

4.2 BASIC MECHANICS

This section reviews some of the main points from any standard introductory mechanics (statics and dynamics) course. Good references abound, such as *Engineering Mechanics* by Merriam and Kraige (2008). A review of vector mathematics is followed by matrix coordinate transformations, a topic new to some students. Euler's equations of motion (see Section 4.2.5) may also be new material. For both topics, *Principles of Dynamics* by Greenwood provides a comprehensive reference.

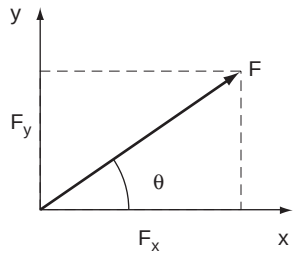


FIGURE 4.2 Two-dimensional representation of vector **F**.

4.2.1 Vector Mathematics

Forces may be written in terms of scalar components and unit vectors, of magnitude equal to one, or in polar form with magnitude and direction. Figure 4.2 shows that the two-dimensional vector **F** is composed of the **i** component, F_x, in the x-direction, and the **j** component, F_y, in the y-direction, or

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} \tag{4.1}$$

as in $20\mathbf{i} + 40\mathbf{j}$ lb. In this chapter, vectors are set in bold type. This same vector may be written in polar form in terms of the vector's magnitude $|F|$, also called the *norm*, and the vector's angle of orientation, θ:

$$|F| = \sqrt{F_x^2 + F_y^2} \tag{4.2}$$

$$\theta = \arctan\frac{F_y}{F_x} \tag{4.3}$$

yielding $|F| = 44.7$ lb and $\theta = 63.4^\circ$. Vectors are similarly represented in three dimensions in terms of their **i**, **j**, and **k** components:

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k} \tag{4.4}$$

with **k** in the z-direction.

Often, a vector's magnitude and two points along its line of action are known. Consider the three-dimensional vector in Figure 4.3. **F** has magnitude of 10 lb, and its line of action passes from the origin (0,0,0) to the point (2,6,4). **F** is written as the product of the magnitude $|F|$ and a unit vector $\mathbf{e_F}$ that points along its line of action:

$$\begin{aligned}\mathbf{F} &= |F|\mathbf{e_F} \\ &= 10 \text{ lb}\left(\frac{2\mathbf{i} + 6\mathbf{j} + 4\mathbf{k}}{\sqrt{2^2 + 6^2 + 4^2}}\right) \\ \mathbf{F} &= 2.67\mathbf{i} + 8.02\mathbf{j} + 5.34\mathbf{k} \text{ lb}\end{aligned}$$

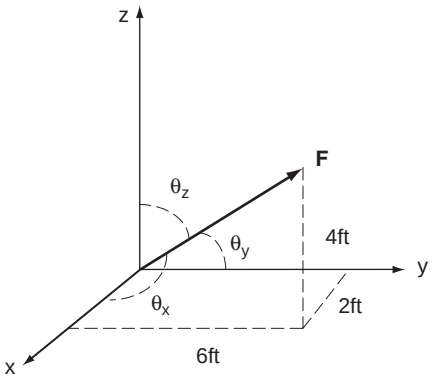


FIGURE 4.3 Three-dimensional vector defined by its magnitude and line of action.

The quantity in parentheses is the unit vector of **F**, or

$$\mathbf{e_F} = \left(\frac{2\mathbf{i} + 6\mathbf{j} + 4\mathbf{k}}{\sqrt{2^2 + 6^2 + 4^2}}\right) = 0.267\mathbf{i} + 0.802\mathbf{j} + 0.534\mathbf{k}$$

and the magnitude of **F** is

$$\begin{aligned} |F| &= \sqrt{2.67^2 + 8.02^2 + 5.34^2} \\ &= 10 \text{ lb} \end{aligned}$$

The vector **F** in Figure 4.3 may also be defined in 3-D space in terms of the angles between its line of action and each coordinate axis. Consider the angles θ_x, θ_y, and θ_z that are measured from the positive x, y, and z axes, respectively, to **F**. Then

$$\cos\theta_x = \frac{F_x}{|F|} \tag{4.5}$$

$$\cos\theta_y = \frac{F_y}{|F|} \tag{4.6}$$

$$\cos\theta_z = \frac{F_z}{|F|} \tag{4.7}$$

These ratios are termed the *direction cosines* of **F**. The unit vector $\mathbf{e_F}$ is equivalent to

$$\mathbf{e_F} = \cos\theta_x\mathbf{i} + \cos\theta_y\mathbf{j} + \cos\theta_z\mathbf{k} \tag{4.8}$$

or, in general

$$\mathbf{e_F} = \left(\frac{F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}}{\sqrt{F_x^2 + F_y^2 + F_z^2}}\right) \tag{4.9}$$

The angles θ_x, θ_y, and θ_z for this example are consequently

$$\theta_x = \arccos\left(\frac{2.67}{10}\right) = 74.5^\circ$$

$$\theta_y = \arccos\left(\frac{8.02}{10}\right) = 36.7^\circ$$

$$\theta_z = \arccos\left(\frac{5.34}{10}\right) = 57.7^\circ$$

Vectors are added by summing their components:

$$\mathbf{A} = A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k}$$
$$\mathbf{B} = B_x\mathbf{i} + B_y\mathbf{j} + B_z\mathbf{k}$$
$$\mathbf{C} = \mathbf{A} + \mathbf{B} = (A_x + B_x)\mathbf{i} + (A_y + B_y)\mathbf{j} + (A_z + B_z)\mathbf{k}$$

In general, a set of forces may be combined into an equivalent force denoted the resultant **R**, where

$$\mathbf{R} = \sum F_x\mathbf{i} + \sum F_y\mathbf{j} + \sum F_z\mathbf{k} \tag{4.10}$$

as will be illustrated in subsequent sections. Vectors are subtracted similarly by subtracting vector components.

Vector multiplication consists of two distinct operations: the *dot* and *cross* products. The dot, or scalar, product of vectors **A** and **B** produces a scalar via

$$\mathbf{A} \cdot \mathbf{B} = AB\cos\theta \tag{4.11}$$

where θ is the angle between the vectors. For an orthogonal coordinate system, where all axes are 90° apart, all like terms alone remain, since

$$\begin{aligned} \mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} = 1 \\ \mathbf{i}\cdot\mathbf{j} = \mathbf{j}\cdot\mathbf{k} = \mathbf{k}\cdot\mathbf{i} = \cdots = 0 \end{aligned} \tag{4.12}$$

For example:

$$\mathbf{A} = 3\mathbf{i} + 2\mathbf{j} + \mathbf{k}\ \text{ft}$$
$$\mathbf{B} = -2\mathbf{i} + 3\mathbf{j} + 10\mathbf{k}\ \text{lb}$$
$$\mathbf{A}\cdot\mathbf{B} = 3(-2) + 2(3) + 1(10) = 10\ \text{ft lb}$$

Note that the dot product is commutative—that is, $\mathbf{A}\cdot\mathbf{B} \equiv \mathbf{B}\cdot\mathbf{A}$.

The physical interpretation of the dot product $\mathbf{A}\cdot\mathbf{B}$ is the projection of **A** onto **B**, or, equivalently, the projection of **B** onto **A**. For example, *work* is defined as the force that acts in the same direction as the motion of a body. Figure 4.4 (left) shows a force vector **F** dotted with a direction of motion vector **d**. The work W done by **F** is given by $\mathbf{F}\cdot\mathbf{d} \equiv Fd\cos\theta$. Dotting **F** with **d** yields the component of **F** acting in the same direction as **d**.

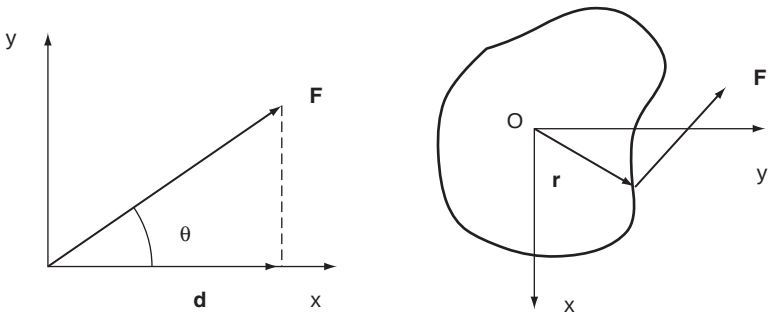


FIGURE 4.4 (Left) The dot, or scalar, product of vectors **F** and **d** is equivalent to the projection of **F** onto **d**. (Right) The cross, or vector, product of vectors **r** and **F** is a vector that points along the axis of rotation, the *z*-axis coming out of the page.

The *moment* of a force about a point or axis is a measure of its tendency to cause rotation. The cross, or vector, product of two vectors yields a new vector that points along the axis of rotation. For example, Figure 4.4 (right) shows a vector **F** acting in the *x*-*y* plane at a distance from the body's coordinate center O. The vector **r** points from O to the line of action of **F**. The cross product $\mathbf{r} \times \mathbf{F}$ is a vector that points in the *z* direction along the body's axis of rotation. If **F** and **r** are three-dimensional, thereby including **k** components, their cross product will have additional components of rotation about the *x* and *y* axes. The moment **M** resulting from crossing **r** into **F** is written

$$\mathbf{M} = M_x\mathbf{i} + M_y\mathbf{j} + M_z\mathbf{k} \tag{4.13}$$

where M_x, M_y, and M_z cause rotation of the body about the *x*, *y*, and *z* axes, respectively.

Cross products may be taken by crossing each vector component term by term—for example:

$$\begin{aligned}\mathbf{A} \times \mathbf{B} = {} & 3(-2)\mathbf{i} \times \mathbf{i} + 3(3)\mathbf{i} \times \mathbf{j} + 3(10)\mathbf{i} \times \mathbf{k} \\ & + 2(-2)\mathbf{j} \times \mathbf{i} + 2(3)\mathbf{j} \times \mathbf{j} + 2(10)\mathbf{j} \times \mathbf{k} \\ & + 1(-2)\mathbf{k} \times \mathbf{i} + 1(3)\mathbf{k} \times \mathbf{j} + 1(10)\mathbf{k} \times \mathbf{k}\end{aligned}$$

The magnitude $|\mathbf{A} \times \mathbf{B}| = AB \sin\theta$, where θ is the angle between **A** and **B**. Consequently, for an orthogonal coordinate system, the cross products of all like terms equal zero, and $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, $\mathbf{j} \times \mathbf{k} = \mathbf{i}$, $\mathbf{k} \times \mathbf{i} = \mathbf{j}$, $\mathbf{i} \times \mathbf{k} = -\mathbf{j}$, and so on. The previous example yields

$$\begin{aligned}\mathbf{A} \times \mathbf{B} &= 9\mathbf{k} - 30\mathbf{j} + 4\mathbf{k} + 20\mathbf{i} - 2\mathbf{j} - 3\mathbf{i} \\ &= 17\mathbf{i} - 32\mathbf{j} + 13\mathbf{k} \text{ lb ft}\end{aligned}$$

Note that the cross product is *not* commutative—in other words, $\mathbf{A} \times \mathbf{B} \neq \mathbf{B} \times \mathbf{A}$.

Cross products of vectors are commonly computed using matrices. The previous example $\mathbf{A} \times \mathbf{B}$ is given by the matrix

$$
\begin{aligned}
\mathbf{A} \times \mathbf{B} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \\
&= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 3 & 2 & 1 \\ -2 & 3 & 10 \end{vmatrix} \\
&= \mathbf{i}[(2)(10) - (1)(3)] - \mathbf{j}[(3)(10) - (1)(-2)] + \mathbf{k}[(3)(3) - (2)(-2)] \\
&= \mathbf{i}(20 - 3) - \mathbf{j}(30 + 2) + \mathbf{k}(9 + 4) \\
&= 17\mathbf{i} - 32\mathbf{j} + 13\mathbf{k} \text{ lb ft}
\end{aligned} \tag{4.14}
$$

EXAMPLE PROBLEM 4.1

The vector **F** in Figure 4.5 has a magnitude of 10 kN and points along the dashed line as shown. (a) Write **F** as a vector. (b) What is the component of **F** in the *x*-*z* plane? (c) What moment does **F** generate about the origin (0,0,0)?

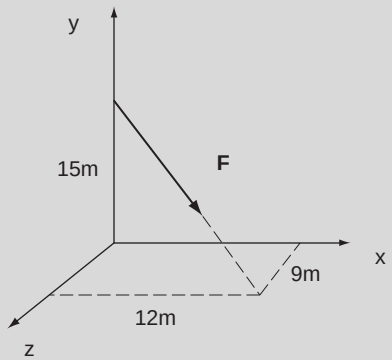


FIGURE 4.5 Force vector **F** has magnitude of 10 kN.

Solution

This example problem is solved using MATLAB. The >> prompt denotes input, and the percent sign, %, precedes comments ignored by MATLAB. Lines that begin without the >> prompt are MATLAB output. Some spaces in the following output were omitted to conserve space.

```
>> %(a) First write the direction vector d that points along F
>> % as a 1D array:
>> d = [12  -15  9]

d = 12  -15  9
>> % Now write the unit vector of F, giving its direction:
>> unit_vector = d/norm(d)
unit_vector = 0.5657  -0.7071  0.4243
```

```
>> % F consists of the magnitude 10 kN times this unit vector
>> F = 10*unit_vector

F = 5.6569  -7.0711  4.2426

>> % Or, more directly
>> F = 10*(d/norm(d))

F = 5.6569  -7.0711  4.2426

>> % (b) First write the vector r_xz that points from the origin
>> % to the intersection of F and the xz plane:
>> r_xz = [12 0 9]
r_xz = 12  0  9

>> % The dot product is given by the sum of all the term by term
>> % multiplications of elements of vectors F and r_xz
>> % F_dot_r_xz = sum(F.*r_xz)
>> % or simply, dot(F,r_xz)
>> F_dot_r_xz = dot(F,r_xz)

F_dot_r_xz = 106.0660

>> % (c) Cross F with a vector that points from the origin to
>> % any point along the line of action of vector F.
>> % The cross product is given by the cross function
>> r_xz_cross_F = cross(r_xz,F)

r_xz_cross_F = 63.6396  0  -84.8528

>> % Note that the cross product is not commutative
>> % resulting in different +- signs.
>> cross(F,r_xz)
ans = -63.6396  0  84.8528
```

EXAMPLE PROBLEM 4.2

Pointers are sometimes used in biomechanics labs to measure the location of a point in space. The pointer in Figure 4.6 consists of a rod equipped with two reflective markers, A and B. The locations of the two reflective markers are provided by a camera-based motion capture system. Given marker locations A = (629, −35, 190) mm and B = (669, 191, 120) mm, determine the location of the pointer tip, T, if marker B is a fixed distance, D, of 127 mm from the pointer tip.

Solution

Given marker locations

A = (629, –35, 190) mm
B = (669, 191, 120) mm

Continued

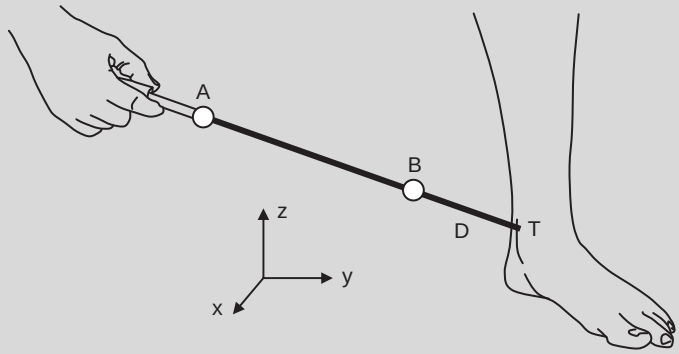


FIGURE 4.6 Pointer with reflective markers A and B and with tip T located a fixed distance D = 127 mm from marker B.

then the vector from marker A to marker B is

$$\begin{aligned}\mathbf{r_{B/A}} &= \mathbf{r_B} - \mathbf{r_A} \\ &= (669\mathbf{i} + 191\mathbf{j} + 120\mathbf{k}) - (629\mathbf{i} - 35\mathbf{j} + 190\mathbf{k})\ \text{mm} \\ &= 40\mathbf{i} + 226\mathbf{j} - 70\mathbf{k}\ \text{mm}\end{aligned}$$

with an associated unit vector,

$$\begin{aligned}\mathbf{e_{B/A}} &= \frac{\mathbf{r_{B/A}}}{|\mathbf{r_{B/A}}|} \\ &= \frac{40\mathbf{i} + 226\mathbf{j} - 70\mathbf{k}}{\sqrt{(40)^2 + (226)^2 + (-70)^2}} \\ &= 0.167\mathbf{i} + 0.942\mathbf{j} - 0.292\mathbf{k}\end{aligned}$$

Pointer tip T (Figure 4.7) is located by

$$\begin{aligned}\mathbf{r_T} &= \mathbf{r_B} + \mathbf{r_{T/B}} \\ &= \mathbf{r_B} + \mathrm{D}\,\mathbf{e_{T/B}} \equiv \mathbf{r_B} + \mathrm{D}\,\mathbf{e_{B/A}} \\ &= (669\mathbf{i} + 191\mathbf{j} + 120\mathbf{k}) + 127(0.167\mathbf{i} + 0.942\mathbf{j} - 0.292\mathbf{k})\ \text{mm} \\ &= 690\mathbf{i} + 311\mathbf{j} + 83\mathbf{k}\ \text{mm}\end{aligned}$$

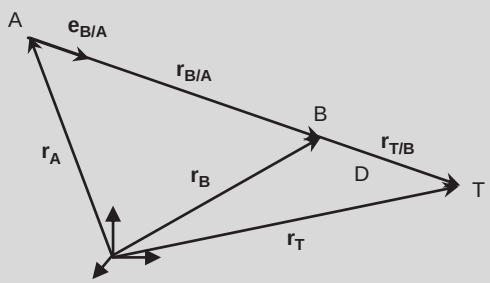


FIGURE 4.7 Pointer tip T can be located using vector r_B, the unit vector $\mathbf{e_{B/A}}$ and the distance D from B to T.

4.2.2 Coordinate Transformations

3-D Direction Cosines

When studying the kinematics of human motion, it is often necessary to transform body or body segment coordinates from one coordinate system to another. For example, coordinates corresponding to a coordinate system determined by markers on the body, a moving coordinate system, must be translated to coordinates with respect to the fixed laboratory, an inertial coordinate system. These three-dimensional transformations use direction cosines that are computed as follows.

Consider the vector **A** measured in terms of the uppercase coordinate system *XYZ*, shown in Figure 4.8 in terms of the unit vectors **I**, **J**, **K**:

$$\mathbf{A} = A_x\mathbf{I} + A_y\mathbf{J} + A_z\mathbf{K} \tag{4.15}$$

The unit vectors **I**, **J**, **K** can be written in terms of **i**, **j**, **k** in the *xyz* system

$$\mathbf{I} = \cos\theta_{xX}\mathbf{i} + \cos\theta_{yX}\mathbf{j} + \cos\theta_{zX}\mathbf{k} \tag{4.16}$$

$$\mathbf{J} = \cos\theta_{xY}\mathbf{i} + \cos\theta_{yY}\mathbf{j} + \cos\theta_{zY}\mathbf{k} \tag{4.17}$$

$$\mathbf{K} = \cos\theta_{xZ}\mathbf{i} + \cos\theta_{yZ}\mathbf{j} + \cos\theta_{zZ}\mathbf{k} \tag{4.18}$$

where θ_{xX} is the angle between **i** and **I**, and similarly for the other angles.

Substituting Eqs. (4.16)–(4.18) into Eq. (4.15) gives

$$\begin{aligned}\mathbf{A} = {} & A_x\left[\cos\theta_{xX}\mathbf{i} + \cos\theta_{yX}\mathbf{j} + \cos\theta_{zX}\mathbf{k}\right] \\ & + A_y\left[\cos\theta_{xY}\mathbf{i} + \cos\theta_{yY}\mathbf{j} + \cos\theta_{zY}\mathbf{k}\right] \\ & + A_z\left[\cos\theta_{xZ}\mathbf{i} + \cos\theta_{yZ}\mathbf{j} + \cos\theta_{zZ}\mathbf{k}\right]\end{aligned} \tag{4.19}$$

or

$$\begin{aligned}\mathbf{A} = {} & \left(A_x\cos\theta_{xX} + A_y\cos\theta_{xY} + A_z\cos\theta_{xZ}\right)\mathbf{i} \\ & + \left(A_x\cos\theta_{yX} + A_y\cos\theta_{yY} + A_z\cos\theta_{yZ}\right)\mathbf{j} \\ & + \left(A_x\cos\theta_{zX} + A_y\cos\theta_{zY} + A_z\cos\theta_{zZ}\right)\mathbf{k}\end{aligned} \tag{4.20}$$

Consequently, **A** may be represented in terms of **I**, **J**, **K** or **i**, **j**, **k**.

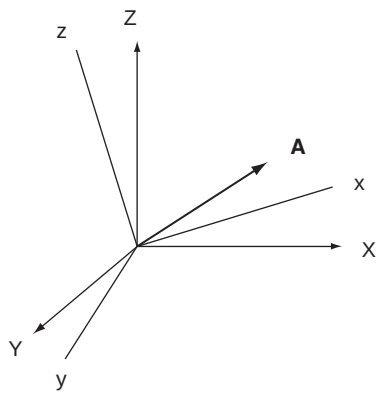


FIGURE 4.8 Vector **A**, measured with respect to coordinate system *XYZ* is related to coordinate system *xyz* via the nine direction cosines of Eq. (4.20).

Euler Angles

The coordinates of a body in one orthogonal coordinate system may be related to another orthogonal coordinate system via Euler angle transformation matrices. For example, one coordinate system might correspond to markers placed on the patient's pelvis, and the other coordinate system might correspond to the patient's thigh. The two coordinate systems are related by a series of rotations about each original axis in turn. Figure 4.9 shows

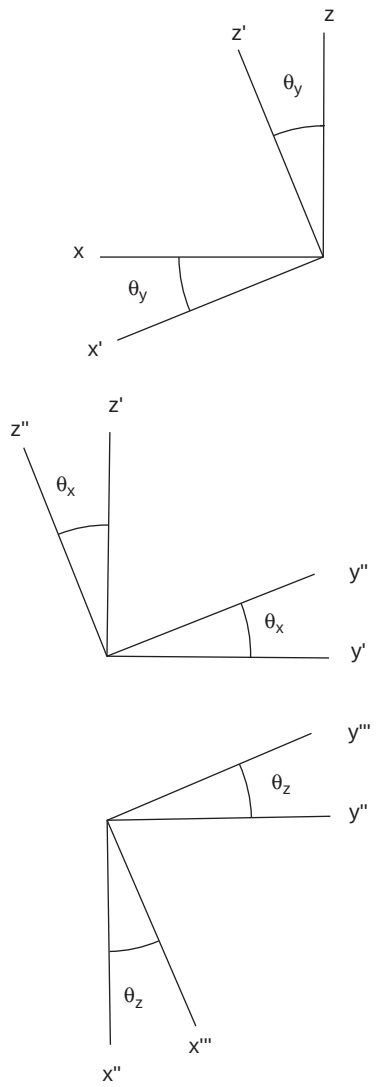


FIGURE 4.9 The unprimed coordinate system *xyz* undergoes three rotations: about the *y*-axis (top), about the *x*-axis (middle), and about the *z*-axis (bottom), yielding the new triple-primed coordinate system $x''' y''' z'''$ for a *y*-*x*-*z* rotation sequence.

the xyz coordinate axes with a y-x-z rotation sequence. First, xyz is rotated about the y-axis (top), transforming the **ijk** unit vectors into the $\mathbf{i}'\mathbf{j}'\mathbf{k}'$ unit vectors, via the equations

$$\mathbf{i}' = \cos\theta_y \mathbf{i} - \sin\theta_y \mathbf{k} \tag{4.21}$$

$$\mathbf{j}' = \mathbf{j} \tag{4.22}$$

$$\mathbf{k}' = \sin\theta_y \mathbf{i} + \cos\theta_y \mathbf{k} \tag{4.23}$$

This new primed coordinate system is then rotated about the x-axis (Figure 4.9, middle), giving the double-primed system

$$\mathbf{i}'' = \mathbf{i}' \tag{4.24}$$

$$\mathbf{j}'' = \cos\theta_x \mathbf{j}' + \sin\theta_x \mathbf{k}' \tag{4.25}$$

$$\mathbf{k}'' = -\sin\theta_x \mathbf{j}' + \cos\theta_x \mathbf{k}' \tag{4.26}$$

Finally, the double-primed system is rotated about the z-axis, giving the triple-primed system

$$\mathbf{i}''' = \cos\theta_z \mathbf{i}'' + \sin\theta_z \mathbf{j}'' \tag{4.27}$$

$$\mathbf{j}''' = -\sin\theta_z \mathbf{i}'' + \cos\theta_z \mathbf{j}'' \tag{4.28}$$

$$\mathbf{k}''' = \mathbf{k}'' \tag{4.29}$$

The three rotations may be written in matrix form to directly translate **ijk** into $\mathbf{i}'''\mathbf{j}'''\mathbf{k}'''$:

$$\begin{bmatrix} \mathbf{i}''' \\ \mathbf{j}''' \\ \mathbf{k}''' \end{bmatrix} = \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix} \tag{4.30}$$

$$= \begin{bmatrix} \cos\theta_z & \sin\theta_z\cos\theta_x & \sin\theta_z\sin\theta_x \\ -\sin\theta_z & \cos\theta_z\cos\theta_x & \cos\theta_z\sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{i}''' \\ \mathbf{j}''' \\ \mathbf{k}''' \end{bmatrix} = \begin{bmatrix} \cos\theta_z\cos\theta_y + \sin\theta_z\sin\theta_x\sin\theta_y & \sin\theta_z\cos\theta_x & -\cos\theta_z\sin\theta_y + \sin\theta_z\sin\theta_x\cos\theta_y \\ -\sin\theta_z\cos\theta_y + \cos\theta_z\sin\theta_x\sin\theta_y & \cos\theta_z\cos\theta_x & \sin\theta_z\sin\theta_y + \cos\theta_z\sin\theta_x\cos\theta_y \\ \cos\theta_x\sin\theta_y & -\sin\theta_x & \cos\theta_x\cos\theta_y \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix} \tag{4.31}$$

If the angles of coordinate system rotation $(\theta_x, \theta_y, \theta_z)$ are known, coordinates in the xyz system can be transformed into the $x'''y'''z'''$ system. Alternatively, if both the unprimed and triple-primed coordinates are known, the angles may be computed as follows:

$$\begin{aligned} \mathbf{k}''' \cdot \mathbf{j} &= -\sin\theta_x \\ \theta_x &= -\arcsin(\mathbf{k}''' \cdot \mathbf{j}) \end{aligned} \tag{4.32}$$

$$\mathbf{k}''' \cdot \mathbf{i} = \cos\theta_x \sin\theta_y$$
$$\theta_y = \arcsin\left[\frac{\mathbf{k}''' \cdot \mathbf{i}}{\cos\theta_x}\right] \tag{4.33}$$

$$\mathbf{i}''' \cdot \mathbf{j} = \sin\theta_z \cos\theta_x$$
$$\theta_z = \arcsin\left[\frac{\mathbf{i}''' \cdot \mathbf{j}}{\cos\theta_x}\right] \tag{4.34}$$

Walking produces tri-planar hip, knee, and ankle motion: flexion/extension (FE), abduction/adduction (AA), and internal/external transverse rotation (TR). Euler angles offer an opportunity to quantify these coordinated motions. The order of the Euler angle rotation sequence FE-AA-TR, corresponding to a *y-x-z* axis rotation sequence, or tilt-obliquity-rotation, was chosen to correlate to the largest to smallest joint excursions during walking [6]. More recently, research suggests that for pelvic motion, an Euler angle *z-x-y* rotation sequence corresponding to rotation-obliquity-tilt is more consistent with clinical observations [1].

EXAMPLE PROBLEM 4.3

Write the Euler angle transformation matrices for the *y-x-z* rotation sequence using the MATLAB symbolic math toolbox.

Solution

The following MATLAB script, or m-file, is a collection of MATLAB commands that can be run by invoking the m-file name "eulerangles" in the command line.

```
% eulerangles.m
%
% Euler angles for y-x-z rotation sequence
% using MATLAB symbolic math toolbox
%
% x, y and z are thetax, thetay and thetaz, respectively
% First define them as symbolic variables

syms x y z

% Writing equations 4.21-23 as a matrix A

A = [ cos(y),  0,  -sin(y);
      0,       1,        0;
      sin(y),  0,   cos(y)]

% equations 4.24-26 as matrix B

B = [ 1     0,         0;
      0,  cos(x), sin(x);
      0, -sin(x), cos(x)]
```

```
% and equations 4.27-29 as matrix C
C = [ cos(z), sin(z), 0;
    -sin(z), cos(z), 0;
       0,      0,    1]

% The matrix equation 4.30 is created by multiplying matrices C, B
% and A

D=C*B*A
```

The resulting transformation matrix from the preceding m-file is

```
D =
[cos(z)*cos(y)+sin(z)*sin(x)*sin(y), sin(z)*cos(x), -cos(z)
*sin(y)+sin(z)*sin(x)*cos(y)]
[-sin(z)*cos(y)+cos(z)*sin(x)*sin(y), cos(z)*cos(x), sin(z)
*sin(y)+cos(z)*sin(x)*cos(y)]
[cos(x)*sin(y),   -sin(x),   cos(x)*cos(y)]
```

which is the same as Eq. (4.31).

The Euler transformation matrices are used differently depending on the available data. For example, if the body coordinates in both the fixed (unprimed) and body (triple primed) systems are known, the body angles θ_x, θ_y, and θ_z can be computed, for example, using Eqs. (4.32)–(4.34) for a *y*-*x*-*z* rotation sequence. Alternatively, the body's initial position and the angles θ_x, θ_y, and θ_z may be used to compute the body's final position.

EXAMPLE PROBLEM 4.4

An aircraft undergoes 30 degrees of pitch (θ_x), then 20 degrees of roll (θ_y), and finally 10 degrees of yaw (θ_z). Write a MATLAB function that computes the Euler angle transformation matrix for this series of angular rotations.

Solution

Since computers use radians for trigonometric calculations, first write two simple functions to compute cosines and sines in degrees:

```
function y = cosd(x)
%COSD(X) cosines of the elements of X measured in degrees.
y = cos(pi*x/180);

function y = sind(x)
%SIND(X) sines of the elements of X measured in degrees.
y = sin(pi*x/180);
```

Next write the *x*-*y*-*z* rotation sequence transformation matrix:

```
function D = eulangle(thetax, thetay, thetaz)
%EULANGLE matrix of rotations by Euler's angles.
% EULANGLE(thetax, thetay, thetaz) yields the matrix of
```

Continued

```
% rotation of a system of coordinates by Euler's
% angles thetax, thetay and thetaz, measured in degrees.

% Now the first rotation is about the x-axis, so we use eqs. 4.24-26

A = [ 1   0              0
      0   cosd(thetax)   sind(thetax)
      0   -sind(thetax)  cosd(thetax) ];

% Next is the y-axis rotation (eqs. 4.21-23)

B = [ cosd(thetay)   0   -sind(thetay)
      0              1   0
      sind(thetay)   0   cosd(thetay) ];

% Finally, the z-axis rotation (eqs. 4.27-29)

C = [ cosd(thetaz)    sind(thetaz)   0
      -sind(thetaz)   cosd(thetaz)   0
      0               0              1 ];
% Multiplying rotation matrices C, B and A as in Eq. 4.30 gives the
solution:

D=C*B*A;
```

Now use this function to compute the numerical transformation matrix:

```
>> eulangle(30,20,10)
ans =
  0.9254    0.3188   -0.2049
 -0.1632    0.8232    0.5438
  0.3420   -0.4698    0.8138
```

This matrix can be used to convert any point in the initial coordinate system (premaneuver) to its position after the roll, pitch, and yaw maneuvers have been executed.

4.2.3 Static Equilibrium

Newton's equations of motion applied to a structure in static equilibrium reduce to the following vector equations:

$$\sum \mathbf{F} = 0 \tag{4.35}$$

$$\sum \mathbf{M} = 0 \tag{4.36}$$

These equations are applied to biological systems in the same manner as standard mechanical structures. Analysis begins with a drawing of the free-body diagram of the body segments of interest with all externally applied loads and reaction forces at the supports. Orthopedic joints can be modeled with appropriate ideal joints, such as hinge, ball-and-socket, and so forth, as discussed in Chapter 3 (see Figure 3.33).

EXAMPLE PROBLEM 4.5

Figure 4.10 (top) shows a Russell's traction rig used to apply an axial tensile force to a fractured femur for immobilization. (a) What magnitude weight w must be suspended from the free end of the cable to maintain the leg in static equilibrium? (b) Compute the average tensile force applied to the thigh under these conditions.

Solution

The free-body diagram for this system is shown in the lower panel of Figure 4.10. If the pulleys are assumed frictionless and of small radius, the cable tension T is constant throughout. Using Eq. (4.35),

$$\mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3 + \mathbf{F}_{\text{femur}} - mg\mathbf{j} = 0$$

Writing each force in vector form,

$$\mathbf{F}_1 = -F_1\mathbf{i} = -T\mathbf{i}$$

$$\begin{aligned}\mathbf{F}_2 &= (-F_2\cos 30^\circ)\mathbf{i} + (F_2\sin 30^\circ)\mathbf{j} \\ &= (-T\cos 30^\circ)\mathbf{i} + (T\sin 30^\circ)\mathbf{j}\end{aligned}$$

$$\begin{aligned}\mathbf{F}_3 &= (F_3\cos 40^\circ)\mathbf{i} + (F_3\sin 40^\circ)\mathbf{j} \\ &= (T\cos 40^\circ)\mathbf{i} + (T\sin 40^\circ)\mathbf{j}\end{aligned}$$

$$\mathbf{F}_{\text{femur}} = (F_{\text{femur}}\cos 20^\circ)\mathbf{i} = (F_{\text{femur}}\sin 20^\circ)\mathbf{j}$$

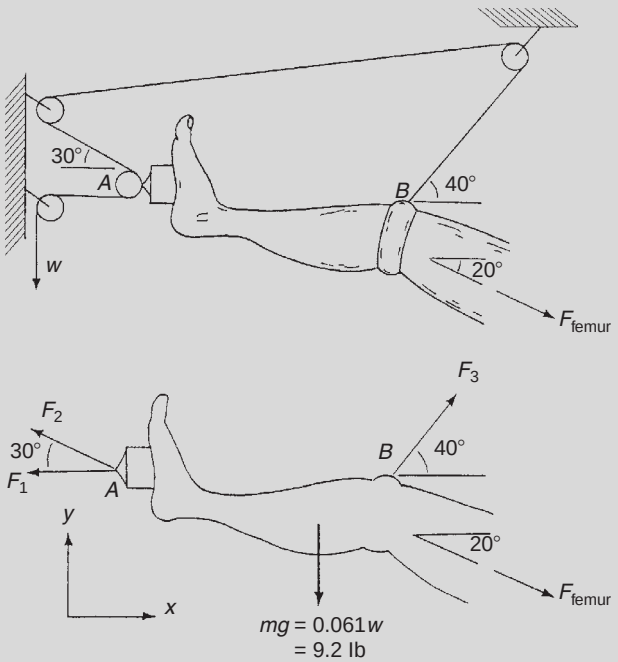


FIGURE 4.10 (Top) Russell's traction mechanism for clinically loading lower-extremity limbs. (Bottom) Free-body diagram of the leg in traction. *Adapted from [5].*

Continued

Using Table 4.1, and neglecting the weight of the thigh, the weight of the foot and leg is 0.061 multiplied by total body weight, yielding

$$mg\mathbf{j} = (0.061)(150\mathbf{j}\text{ lb}) = 9.2\mathbf{j}\text{ lb}$$

Summing the x components gives

$$-T - T\cos 30^\circ + T\cos 40^\circ + F_{\text{femur}}\cos 20^\circ = 0$$

Summing the y components gives

$$T\sin 30^\circ + T\sin 40^\circ - F_{\text{femur}}\sin 20^\circ - mg = 0$$

The last two expressions may be solved simultaneously, giving both T, which is equal to the required externally applied weight, and the axial tensile force, F_{femur}

$$T = 12.4\text{ lb}$$
$$F_{\text{femur}} = 14.5\text{ lb}$$

EXAMPLE PROBLEM 4.6

The force plate depicted in Figure 4.11 has four sensors, one at each corner, that read the vertical forces F_1, F_2, F_3, and F_4. If the plate is square with side of length ℓ and forces $F_1 - F_4$ are known, write two expressions that will give the x and y locations of the resultant force R.

Solution

The resultant magnitude R can be computed from the sum of forces in the z-direction:

$$\sum F_z = 0$$
$$F_1 + F_2 + F_3 + F_4 - R = 0$$
$$R = F_1 + F_2 + F_3 + F_4$$

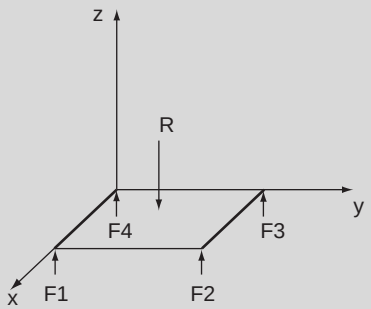


FIGURE 4.11 A square force plate with sides of length ℓ is loaded with resultant force R and detects the vertical forces at each corner, $F_1 - F_4$.

The force plate remains horizontal, so the sum of the moments about the x and y axes must each be zero. Taking moments about the x-axis,

$$\sum M_x = 0$$
$$F_2\ell + F_3\ell - Ry = 0$$
$$y = \frac{(F_2 + F_3)\ell}{R}$$

Similarly, summing moments about the y-axis,

$$\sum M_y = 0$$

$$F_1\ell + F_2\ell - Rx = 0$$

$$x = \frac{(F_1 + F_2)\ell}{R}$$

The coordinates x and y locate the resultant R.

4.2.4 Anthropomorphic Mass Moments of Inertia

A body's mass resists linear motion; its mass moment of inertia resists rotation. The resistance of a body, or a body segment such as a thigh in gait analysis, to rotation is quantified by the body or body segment's moment of inertia I:

$$I = \int_m r^2 dm \tag{4.37}$$

where m is the body mass and r is the moment arm to the axis of rotation. The elemental mass dm can be written ρdV. For a body with constant density ρ, the moment of inertia can be found by integrating over the body's volume V:

$$I = \rho \int_V r^2 dV \tag{4.38}$$

This general expression can be written in terms of rotation about the x, y, and z axes:

$$\begin{aligned} I_{xx} &= \textstyle\int_V (y^2 + z^2)\rho dV \\ I_{yy} &= \textstyle\int_V (x^2 + z^2)\rho dV \\ I_{zz} &= \textstyle\int_V (x^2 + y^2)\rho dV \end{aligned} \tag{4.39}$$

The *radius of gyration* k is the moment arm between the axis of rotation and a single point where all of the body's mass is concentrated. Consequently, a body segment may be treated as a point mass with moment of inertia

$$I = mk^2 \tag{4.40}$$

where m is the body segment mass. The moment of inertia with respect to a parallel axis I is related to the moment of inertia with respect to the body's center of mass I_{cm} via the *parallel axis theorem*:

$$I = I_{cm} + md^2 \tag{4.41}$$

where d is the perpendicular distance between the two parallel axes. Anthropomorphic data for various body segments are listed in Table 4.1.

TABLE 4.1 Anthropomorphic Data

Segment	Definition	Segment Weight/ Body Weight	Center Mass/ Segment Length		Radius Gyration/ Segment Length	
			Proximal	Distal	Proximal	Distal
Hand	Wrist axis/knuckle II middle finger	0.006	0.506	0.494	0.587	0.577
Forearm	Elbow axis/ulnar styloid	0.016	0.430	0.570	0.526	0.647
Upper arm	Glenohumeral axis/elbow axis	0.028	0.436	0.564	0.542	0.645
Forearm and hand	Elbow axis/ulnar styloid	0.022	0.682	0.318	0.827	0.565
Total arm	Glenohumeral joint/ulnar styloid	0.050	0.530	0.470	0.645	0.596
Foot	Lateral malleolus/head metatarsal II	0.0145	0.50	0.50	0.690	0.690
Leg	Femoral condyles/medial malleolus	0.0465	0.433	0.567	0.528	0.643
Thigh	Greater trochanter/femoral condyles	0.100	0.433	0.567	0.540	0.653
Foot and leg	Femoral condyles/medial malleolus	0.061	0.606	0.394	0.735	0.572
Total leg	Greater trochanter/medial malleolus	0.161	0.447	0.553	0.560	0.650
Head and neck	C7-T1 and 1st rib/ear canal	0.081	1.000		1.116	
Shoulder mass	Sternoclavicular joint/ glenohumeral axis		0.712	0.288		
Thorax	C7-T1/T12-L1 and diaphragm	0.216	0.82	0.18		
Abdomen	T12-L1/L4-L5	0.139	0.44	0.56		
Pelvis	L4-L5/greater trochanter	0.142	0.105	0.895		
Thorax and abdomen	C7-T1/L4-L5	0.355	0.63	0.37		
Abdomen and pelvis	T12-L1/greater trochanter	0.281	0.27	0.73		
Trunk	Greater trochanter/glenohumeral joint	0.497	0.50	0.50		
Trunk, head, neck	Greater trochanter/glenohumeral joint	0.578	0.66	0.34	0.830	0.607
Head, arm, trunk	Greater trochanter/glenohumeral joint	0.678	0.626	0.374	0.798	0.621

Adapted from Winter, 2009.

EXAMPLE PROBLEM 4.7

A person weighing 150 pounds has a thigh length of 17 inches. Find the moment of inertia of this body segment with respect to its center of mass in SI units.

Solution

Thigh length in SI units is

$$\ell_{\text{thigh}} = 17 \text{ in} = 0.432 \text{ m}$$

Table 4.1 lists ratios of segment weight to body weight for different body segments. Starting with body mass,

$$m_{\text{body}} = (150 \text{ lb})(0.454 \text{ kg/lb}) = 68.1 \text{ kg}$$

the thigh segment mass is

$$m_{\text{thigh}} = (0.100)(68.1 \text{ kg}) = 6.81 \text{ kg}$$

Table 4.1 also lists body segment center of mass and radius of gyration as ratios with respect to segment length for each body segment. Table 4.1 gives both proximal and distal segment length ratios. Note that "proximal" for the thigh refers toward the hip and "distal" refers toward the knee. Consequently, the proximal thigh segment length is the distance between the thigh center of mass and the hip, and the distal thigh segment length is the distance between the thigh center of mass and the knee. The moment of inertia of the thigh with respect to the hip is therefore

$$I_{\text{thigh/hip}} = mk^2 = (6.81 \text{ kg})[(0.540)(0.432 \text{ m})]^2 = 0.371 \text{ kg m}^2$$

The thigh's moment of inertia with respect to the hip is related to the thigh's moment of inertia with respect to its center of mass via the parallel axis theorem (Eq. (4.41)). Consequently, Table 4.1 data can be used to compute segment moments of inertia with respect to their centers of mass:

$$I_{\text{thigh/hip}} = I_{\text{thigh/cm}} + md^2$$

so

$$I_{\text{thigh/cm}} = I_{\text{thigh/hip}} - md^2$$

In this case, distance d is given by the proximal segment length data:

$$d = (0.432 \text{ m})(0.433) = 0.187 \text{ m}$$

and the final result is

$$I_{\text{thigh/cm}} = 0.371 \text{ kg m}^2 - (6.81 \text{ kg})(0.187 \text{ m})^2 = 0.133 \text{ kg m}^2$$

EXAMPLE PROBLEM 4.8

A person weighing 160 pounds is holding a 10-lb weight in his palm, with the elbow fixed at 90° flexion (Figure 4.12 (top)). (a) What force must the biceps generate to hold the forearm in static equilibrium? (b) What force(s) does the forearm exert on the humerus?

Continued

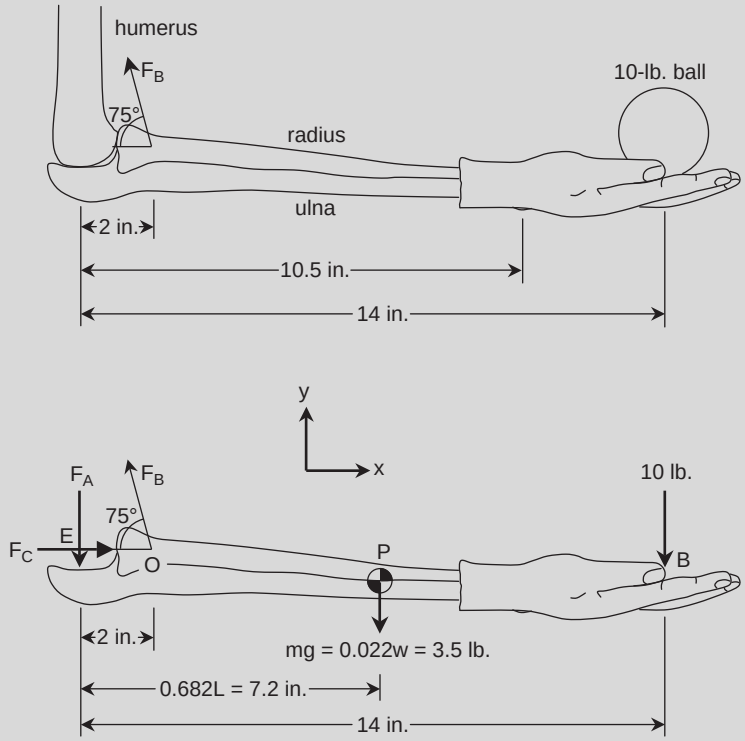


FIGURE 4.12 (Top) The forearm held statically fixed in 90° flexion while holding a 10 lb weight at the hand. (Bottom) Free-body diagram of the forearm system. *Adapted from [5].*

Solution

Figure 4.12 (bottom) shows the free-body diagram of this system. Due to the increased number of unknowns compared to the previous example, both Eqs. (4.35) and (4.36) will be used. From the anthropometric relationships in Table 4.1, the segment weight (forearm and hand) is approximated as 2.2 percent of total body weight, with the segment mass located 68.2 percent of the segment length away from the elbow axis. Note that the segment length for the "forearm and hand" segment in Table 4.1 is defined as the distance between the elbow axis and the ulnar styloid.

Summing moments about the elbow at point O, the equilibrium equation $\sum \mathbf{M} = 0$ can be written as

$$\mathbf{r}_{OE} \times \mathbf{F}_A + \mathbf{r}_{OB} \times (-10\ \text{lb})\mathbf{j} + \mathbf{r}_{OP} \times (-3.5\ \text{lb})\mathbf{j} = 0$$
$$(-2\ \text{in})\mathbf{i} \times (-F_A)\mathbf{j} + (12\ \text{in})\mathbf{i} \times (-10\ \text{lb})\mathbf{j} + (5.2\ \text{in})\mathbf{i} \times (-3.5\ \text{lb})\mathbf{j} = 0$$
$$(2\ \text{in})F_A\mathbf{k} - (120\ \text{lb in})\mathbf{k} - (18.2\ \text{lb in})\mathbf{k} = 0$$

Solving this last expression for the one unknown, F_A, the vertical force at the elbow:

$$F_A = 69.1 \text{ lb}$$

To find the unknown horizontal force at the elbow, F_C, and the unknown force the biceps must generate, F_B, the other equation of equilibrium $\sum \mathbf{F} = 0$ is used:

$$F_C\mathbf{i} - F_A\mathbf{j} + (-F_B \cos 75^\circ \mathbf{i} + F_B \sin 75^\circ \mathbf{j}) - 10 \text{ lb}\mathbf{j} - 3.5 \text{ lb}\mathbf{j} = 0$$

Summing the x and y components gives

$$F_C - F_B \cos(75^\circ) = 0$$
$$-F_A + F_B \sin(75^\circ) - 10 \text{ lb} - 3.5 \text{ lb} = 0$$

Solving these last two equations simultaneously and using $F_A = 69.1$ lb gives the force of the biceps muscle, F_B, and the horizontal elbow force, F_C:

$$F_B = 85.5 \text{ lb}$$
$$F_C = 22.1 \text{ lb}$$

4.2.5 Equations of Motion

Vector equations of motion are used to describe the translational and rotational kinetics of bodies.

Newton's Equations of Motion

Newton's second law relates the net force $\mathbf{F}$ and the resulting translational motion as

$$\mathbf{F} = m\mathbf{a} \tag{4.42}$$

where $\mathbf{a}$ is the linear acceleration of the body's center of mass for translation. For rotation

$$\mathbf{M} = \mathbf{I}\boldsymbol{\alpha} \tag{4.43}$$

where $\mathbf{I}\boldsymbol{\alpha}$ is the body's angular momentum. Hence, the rate of change of a body's angular momentum is equal to the net moment $\mathbf{M}$ acting on the body. These two vector equations of motion are typically written as a set of six x, y, and z component equations.

Euler's Equations of Motion

Newton's equations of motion describe the motion of the center of mass of a body. More generally, Euler's equations of motion describe the motion of a rigid body with respect to its center of mass. For the special case where the xyz coordinate axes are chosen to coincide with the body's principal axes, that is, a cartesian coordinate system whose origin is located at the body's center of mass, Euler's equations are

$$\sum M_x = I_{xx}\alpha_x + (I_{zz} - I_{yy})\omega_y\omega_z \tag{4.44}$$

$$\sum M_y = I_{yy}\alpha_y + (I_{xx} - I_{zz})\omega_z\omega_x \tag{4.45}$$

$$\sum M_z = I_{zz}\alpha_z + (I_{yy} - I_{xx})\omega_x\omega_y \tag{4.46}$$

M_i is the net moment, I_{ii} is the body's moment of inertia with respect to the principal axes, and α_i and ω_i are the body's angular acceleration and angular velocity, respectively. Euler's equations require angular measurements in radians. Their derivation is outside the scope of this chapter but may be found in an intermediate dynamics book—for example, [12]. Equations (4.44)–(4.46) will be used in Section 4.6 to compute intersegmental or joint moments.

4.3 MECHANICS OF MATERIALS

Just as kinematic and kinetic relations may be applied to biological bodies to describe their motion and its associated forces, concepts from mechanics of materials may be used to quantify tissue deformation, to study distributed orthopedic forces, and to predict the performance of orthopedic implants and prostheses and of surgical corrections. Since this topic is very broad, some representative concepts will be illustrated with the following examples.

An orthopedic bone plate is a flat segment of stainless steel used to screw two failed sections of bone together. The bone plate in Figure 4.13 has a rectangular cross section, A, measuring 4.17 mm by 12 mm and made of 316L stainless steel. An applied axial load, F, of 500 N produces axial *stress*, σ, (force/area):

$$\begin{aligned}\sigma &= \frac{F}{A} \\ &= \frac{500\ \text{N}}{(4.17 \times 10^{-3}\ \text{m})(12 \times 10^{-3}\ \text{m})} = 10\ \text{MPa}\end{aligned} \tag{4.47}$$

The maximum shear stress, $\tau_{\max}$, occurs at a 45° angle to the applied load

$$\begin{aligned}\tau\,\text{max} &= \frac{F_{45^\circ}}{A_{45^\circ}} \\ &= \frac{(500\ \text{N})\cos 45^\circ}{\left[\dfrac{(0.00417\ \text{m})(0.012\ \text{m})}{\cos 45^\circ}\right]} = 5\ \text{MPa}\end{aligned} \tag{4.48}$$

which is 0.5σ, as expected from mechanics of materials principles. Prior to loading, two points were punched 15 mm apart on the long axis of the plate, as shown. After the 500 N load is applied, those marks are an additional 0.00075 mm apart. The plate's *strain*, ε, relates the change in length, $\Delta\ell$ to the original length, ℓ:

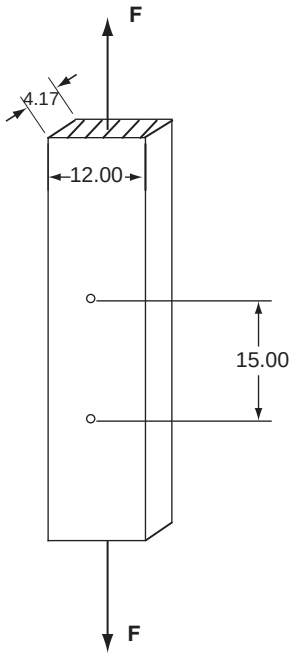


FIGURE 4.13 Bone plate used to fix bone fractures, with applied axial load. Dimensions are in mm. *Adapted from [2].*

$$\varepsilon = \frac{\Delta \ell}{\ell} = \frac{0.00075 \text{ mm}}{15 \text{ mm}} = 50 \times 10^{-6} \tag{4.49}$$

often reported as 50 μ, where μ denotes microstrain (10^{-6}).

The *elastic modulus*, E, relates stress and strain and is a measure of a material's resistance to distortion by a tensile or compressive load. For linearly elastic (Hookean) materials, E is a constant, and a plot of σ as a function of ε is a straight line with slope E:

$$E = \frac{\sigma}{\varepsilon} \tag{4.50}$$

For the bone plate,

$$E = \frac{10 \times 10^6 \text{ Pa}}{50 \times 10^{-6}} = 200 \text{ GPa}$$

Materials such as metals and plastics display linearly elastic properties only in limited ranges of applied loads. Biomaterials have even more complex elastic properties. Figure 4.14 shows tensile stress-strain curves measured from longitudinal and transverse sections of

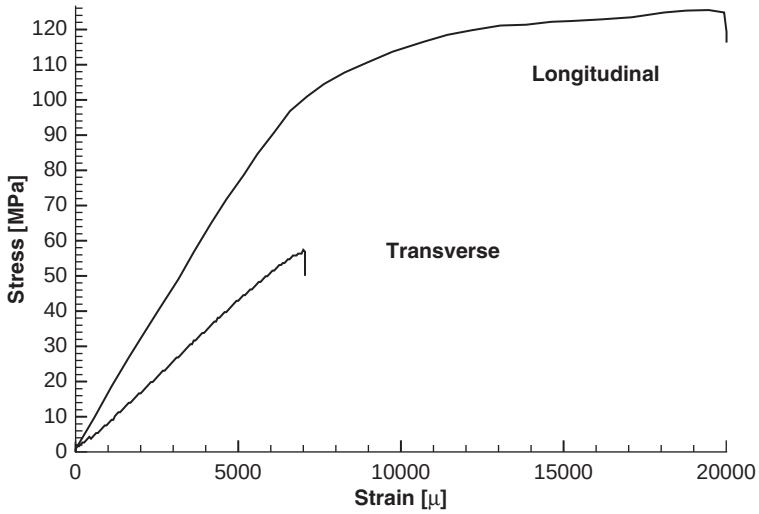


FIGURE 4.14 Tensile stress-strain curves for longitudinal and transverse sections of bone. *Adapted from [2].*

TABLE 4.2 Tensile Yield and Ultimate Stresses, and Elastic Moduli *E* for Some Common Orthopedic Materials

Material	σ_{yield} [MPa]	$\sigma_{ultimate}$ [MPa]	E [GPa]
Stainless steel	700	850	180
Cobalt alloy	490	700	200
Titanium alloy	1,100	1,250	110
Bone	85	120	18
PMMA (fixative)		35	5
UHMWPE (bearing)	14	27	1
Patellar ligament		58	

Data from [2].

bone. Taking the longitudinal curve first, from 0 to 7,000μ, bone behaves as a purely elastic solid, with E $\approx$ 12 GPa. At a tensile stress of approximately 90 MPa, the stress-strain curve becomes nonlinear, yielding into the plastic region of deformation. This sample ultimately fails around 120 MPa. Table 4.2 shows elastic moduli, yield stresses, and ultimate stresses for some common orthopedic materials, both natural and implant.

Figure 4.14 also shows that the elastic properties of bone differ depending on whether the sample is cut in the longitudinal or transverse direction—that is, bone is anisotropic. Bone is much weaker and less stiff in the transverse compared to the longitudinal direction, as is illustrated by the large differences in the yield and ultimate stresses and the slopes of the stress-strain curves for the two samples.

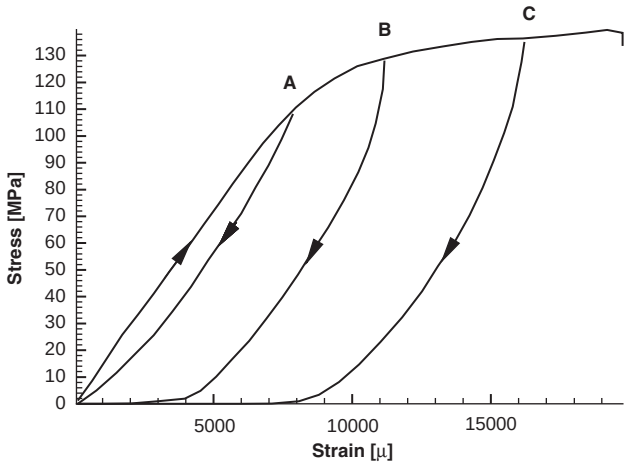


FIGURE 4.15 Bone shows hysteresis and shifting of stress-strain curves with repeated loading and unloading. *Adapted from [2].*

Figure 4.15 shows that the elastic properties of bone also vary depending on whether the load is being applied or removed, displaying hysteresis. From a thermodynamic view, the energy stored in the bone during loading is not equal to the energy released during unloading. This energy difference becomes greater as the maximum load increases (curves A to B to C). The "missing" energy is dissipated as heat due to internal friction and damage to the material at high loads.

The anisotropic nature of bone is sufficient in that its ultimate stress in compression is 200 MPa, while in tension it is only 140 MPa and in torsion 75 MPa. For torsional loading, the *shear modulus* or *modulus of rigidity*, denoted G, relates the shear stress to the shear strain. The modulus of rigidity is related to the elastic modulus via *Poisson's ratio*, v, where

$$v = \frac{\varepsilon_{\text{transverse}}}{\varepsilon_{\text{longitudinal}}} \tag{4.51}$$

Typically, $v \approx 0.3$, meaning that longitudinal deformation is three times greater than transverse deformation. For linearly elastic materials, E, G, and v are related by

$$G = \frac{E}{2(1+v)} \tag{4.52}$$

One additional complexity of predicting biomaterial failure is the complexity of physiological loading. For example, consider "boot-top" fractures in skiing. If the forward motion of a skier is abruptly slowed or stopped, such as by suddenly running into wet or soft snow, his forward momentum causes a moment over the ski boot top, producing three-point bending of the tibia. In this bending mode the anterior tibia undergoes compression, while the posterior is in tension and potentially in failure, since bone is much stronger in compression than in tension. Contraction of the triceps surae muscle produces high compressive stress at the posterior side, reducing the amount of bone tension and helping to prevent injury. Example Problem 4.9 shows how topics from statics and mechanics of materials may be applied to biomechanical problems.

EXAMPLE PROBLEM 4.9

Figure 4.16 (left) shows an orthopedic nail-plate used to fix an intertrochanteric fracture. The hip applies an external force of 400 N during static standing, as shown. The nail-plate is rectangular stainless steel with cross-sectional dimensions of 10 mm (width) by 5 mm (height), and is well-fixed with screws along its vertical axis and friction fit into the trochanteric head (along the x-axis). What forces, moments, stresses, and strains will develop in this orthopedic device?

Solution

As for any statics problem, the first task is constructing a free-body diagram, including all applied forces and moments and all reaction forces and moments that develop at the supports. Because of the instability at the fracture site, the nail-plate may be required to carry the entire 400 N load. Consequently, one reasonable model of the nail-plate is a cantilever beam of length 0.06 m with a combined loading, as depicted in Figure 4.16 (right, top). The applied 400 N load consists of both axial and transverse components:

$$F_x = 400\ \text{N}\cos 20^\circ = 376\ \text{N}$$
$$F_y = 400\ \text{N}\sin 20^\circ = 137\ \text{N}$$

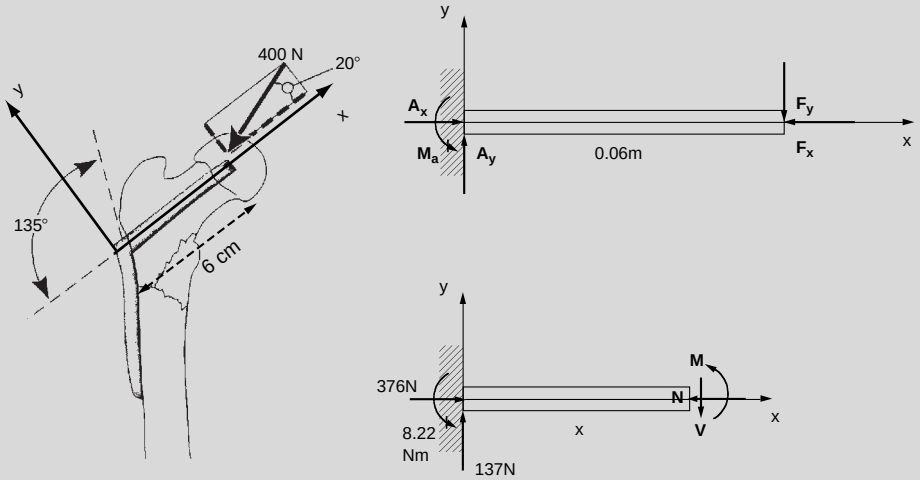


FIGURE 4.16 (Left) An intertrochanteric nail plate used in bone repair. To the right is the free-body diagram of the upper section of this device, and below it is the free-body diagram of a section of this beam cut at a distance x from the left-hand support. *Adapted from [2].*

The axial load produces compressive normal stress; from Eq. (4.47),

$$\sigma_x = \frac{F_x}{A}$$
$$= \frac{376\ \text{N}}{(0.005\ \text{m})(0.01\ \text{m})} = 7.52\ \text{MPa}$$

in compression, which is only about 1 percent of the yield stress for stainless steel (see Table 4.2). The maximum shear stress due to the axial load is

$$\tau_{\max} = \frac{\sigma_x}{2} = 3.76 \text{ MPa}$$

and occurs at 45° from the long axis. The axial strain can be computed using the elastic modulus for stainless steel,

$$E = \frac{\sigma}{\varepsilon} = \frac{F/A}{\Delta\ell/\ell}$$

giving an expression for strain:

$$\varepsilon = \frac{F}{EA}$$

$$= \frac{376 \text{ N}}{180 \times 10^9 \text{Pa}(0.005 \text{ m})(0.01 \text{ m})} = 41.8 \times 10^{-6}$$

From this strain the axial deformation can be computed

$$\Delta\ell_{\text{axial}} = \varepsilon\ell = 2.51 \times 10^{-6}\text{m}$$

which is negligible.

The transverse load causes the cantilever section to bend. The equations describing beam bending can be found in any mechanics of materials text (e.g., [26]). Consider the beam in the left panel of Figure 4.17. If this beam is fixed at the left-hand side and subjected to a downward load on the right, it will bend with the top of the beam elongating and the bottom shortening. Consequently, the top of the beam is in tension, and the bottom in compression. The point of transition, where there is no bending force, is denoted the neutral axis, located at distance c. For a symmetric rectangular beam of height h, c is located at the midline $h/2$. The beam resists bending via its area moment of inertia I. For a rectangular cross section of width b and height h, $I = \frac{1}{12}bh^3$, depicted in the right panel of Figure 4.17.

Beam tip deflection δy is equal to

$$\delta y = \frac{Fx^2}{6EI}(3L - x) \tag{4.53}$$

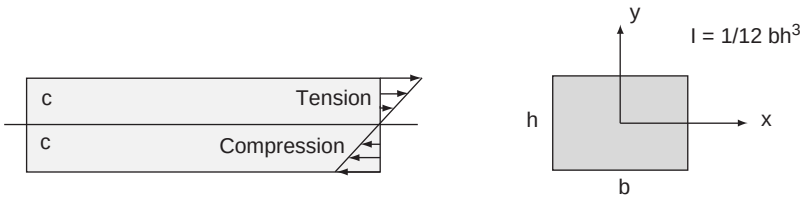


FIGURE 4.17 (Left) A beam fixed on the left and subjected to a downward load on the right undergoes bending, with the top of the beam in tension and the bottom in compression. The position where tension changes to compression is denoted the neutral axis, located at c. (Right) A beam of rectangular cross section with width b and height h resists bending via the area moment of inertia $I = \frac{1}{12}bh^3$.

where x is the axial distance along the beam, L is the total beam length, and I is the beam's cross-sectional area moment of inertia. For this example,

$$I = \frac{1}{12}(10 \times 10^{-3}\ \text{m})(5 \times 10^{-3}\ \text{m})^3 = 1.042 \times 10^{-10}\ \text{m}^4$$

Maximum deflection will occur at $x = L$,

$$\begin{aligned} \delta y_{\max} &= \frac{FL^3}{3EI} \\ &= \frac{137\ \text{N}(0.06\ \text{m})^3}{3(180 \times 10^9 \text{N/m}^2)(1.042 \times 10^{-10}\text{m}^4)} \\ &= 5.26 \times 10^{-4}\ \text{m} = 0.526\ \text{mm} \end{aligned} \tag{4.54}$$

which is also negligible.

Computation of maximum shear and bending stresses requires maximum shear force V and bending moment M. Starting by static analysis of the entire free-body

$$\begin{aligned} \textstyle\sum F_x &: A_x - 376\ \text{N} = 0 \\ \textstyle\sum F_y &: A_y - 137\ \text{N} = 0 \\ \textstyle\sum M_A &: M_a - 137\ \text{N}(0.06\ \text{m}) = 0 \end{aligned}$$

Solving these equations gives $A_x = 376$ N, $A_y = 137$ N, and $M_a = 8.22$ N m. Taking a cut at any point x to the right of A and isolating the left-hand section gives the free-body in Figure 4.16 (right, bottom). Applying the equations of static equilibrium to this isolated section yields

$$\begin{aligned} \textstyle\sum F_x &: 376\ \text{N} - N = 0 \\ N(x) &= 376\ \text{N} \\ \textstyle\sum F_y &: 137\ \text{N} - V = 0 \\ V(x) &= 137\ \text{N} \\ \textstyle\sum M_A &: 8.22\ \text{N m} - (137\ \text{N})(x\ \text{m}) + M = 0 \\ M(x) &= (137\ \text{N m})x - 8.22\ \text{N m} \end{aligned}$$

These last equations can be plotted easily using MATLAB, giving the axial force, shear force, and bending moment diagrams shown in Figure 4.18.

```
% Use MATLAB to plot axial force, shear force and bending moment
% diagrams for Example Problem 9

x = [0:0.01:0.06];
N = x.*0 + 376;
```

```
V = x.*0 + 137;
M = 137.*x - 8.22;

figure
subplot(3,1,1), plot(x,N,x,N,'x')
xlabel('x [m]')
ylabel('N [N]')
title('Axial Force N')
subplot(3,1,2), plot(x,V,x,V,'x')
xlabel('x [m]')
ylabel('V [N]')
title('Shear Force V')
subplot(3,1,3), plot(x,M,x,M,'x')
xlabel('x [m]')
ylabel('M [N-m]')
title('Bending Moment M')
```

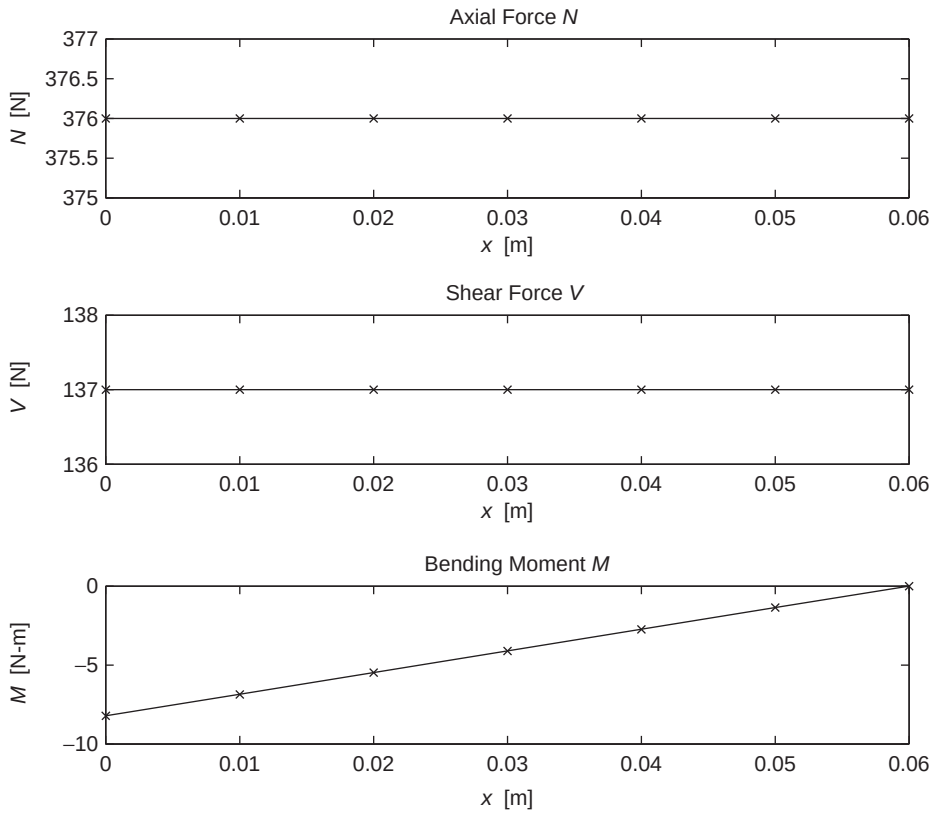


FIGURE 4.18 Axial force N (top), shear force V (middle), and bending moment M (bottom) computed for the nail plate in Figure 4.16 as functions of the distance x along the plate.

The maximum bending and shear stresses follow as

$$\sigma_{b\max} = \frac{M_{\max} c}{I} \tag{4.55}$$

where c, the distance to the beam's neutral axis, is $h/2$ for this beam:

$$\sigma_{b\max} = \frac{-8.22\ \text{Nm}[0.5(5 \times 10^{-3}\ \text{m})]}{1.042 \times 10^{-10}\ \text{m}^4} = -197\ \text{MPa}$$

$$\tau_{b\max} = \frac{V_{\max} h^2}{8I} \tag{4.56}$$

$$= \frac{137\ \text{N}(5 \times 10^{-3}\ \text{m})^2}{8(1.042 \times 10^{-10}\ \text{m}^4)} = 4.11\ \text{MPa}$$

All of these stresses are well below $\sigma_{\text{yield}} = 700$ MPa for stainless steel.

4.4 VISCOELASTIC PROPERTIES

The Hookean elastic solid is a valid description of materials only within a narrow loading range. For example, an ideal spring that relates force and elongation by a spring constant k is invalid in nonlinear low-load and high-load regions. Further, if this spring is coupled to a mass and set into motion, the resulting perfect harmonic oscillator will vibrate forever, which experience shows does not occur. Missing is a description of the system's viscous or damping properties. In this case, energy is dissipated as heat in the spring and air friction on the moving system.

Similarly, biomaterials all display viscoelastic properties. Different models of viscoelasticity have been developed to characterize materials with simple constitutive equations. For example, Figure 4.19 shows three such models that consist of a series ideal spring and dashpot (Maxwell), a parallel spring and dashpot (Voight), and a series spring and dashpot with a parallel spring (Kelvin). Each body contains a dashpot, which generates force in proportion to the derivative of its elongation. Consequently, the resulting models exhibit stress and strain properties that vary in time.

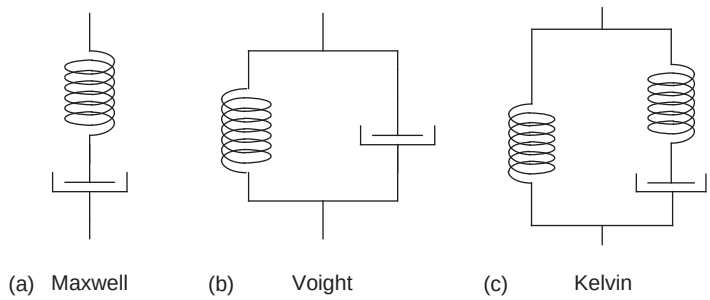


FIGURE 4.19 Three simple viscoelastic models: (a) the Maxwell model, (b) the Voight model, and (c) the Kelvin body or standard linear solid model.

The dynamic response of each model can be quantified by applying a step change in force F and noting the model's resulting change in length, or position x, denoted the *creep* response. The converse experiment applies a step change in x and measures the resulting change in F, denoted *stress relaxation*. Creep and stress relaxation tests for each dynamic model can be carried out easily using the Simulink program. Figure 4.20 shows a purely elastic material subjected to a step change in applied force F. The material's subsequent position x follows the change in force directly. This material exhibits no creep. Figure 4.21 shows the purely elastic material subjected to a step change in position x. Again, the material responds immediately with a step change in F (i.e., no stress relaxation is observed).

James Clerk Maxwell (1831–1879) used a series combination of ideal spring and dashpot to describe the viscoelastic properties of air. Figure 4.22 shows the Maxwell viscoelastic model subjected to a step change in applied force, and Figure 4.23 shows the Maxwell model's stress relaxation response. The latter exhibits an initial high stress followed by stress relaxation back to the initial stress level. The creep response, however, shows that this model is not bounded in displacement, since an ideal dashpot may be extended forever.

Woldemar Voight (1850–1919) used the parallel combination of an ideal spring and dashpot in his work with crystallography. Figure 4.24 shows the creep test of the Voight viscoelastic model. Figure 4.25 shows that this model is unbounded in force. That is, when a step change in length is applied, force goes to infinity, since the dashpot cannot immediately respond to the length change.

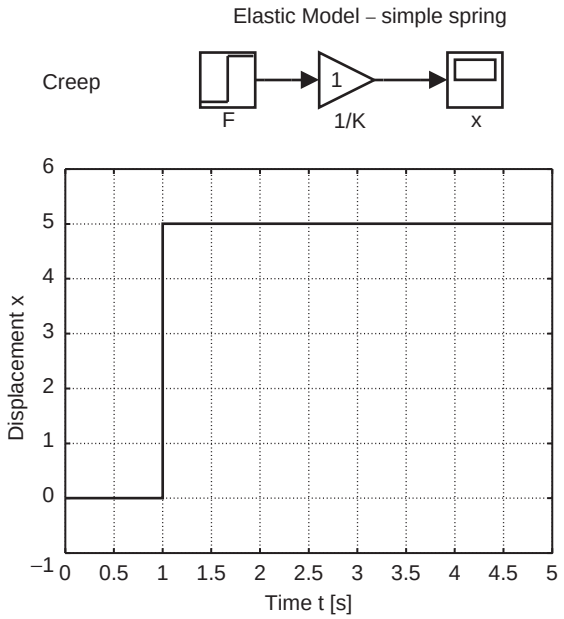


FIGURE 4.20 Simulink model of the creep test for a purely elastic material (an ideal spring). This model solves the equation $x=F/K$, where x is displacement, F is applied force, and K is the spring constant. Below is the elastic creep response to a step increase in applied force F with $K=1$ and force changed from 0 to 5 (arbitrary units). The displacement x linearly follows the applied force.

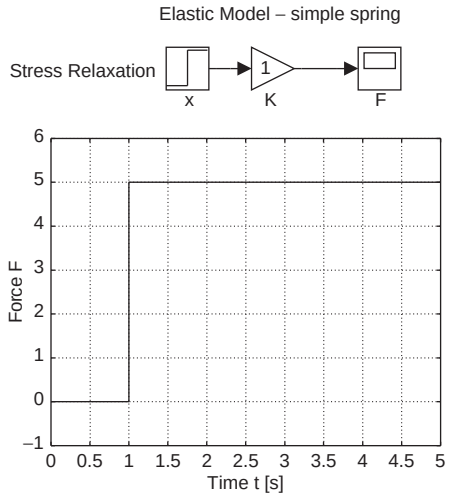


FIGURE 4.21 Simulink model (top) and stress relaxation test of the purely elastic model, which solves $F=Kx$. Applied step displacement $x=5$ and spring constant $K=1$. Force linearly follows displacement.

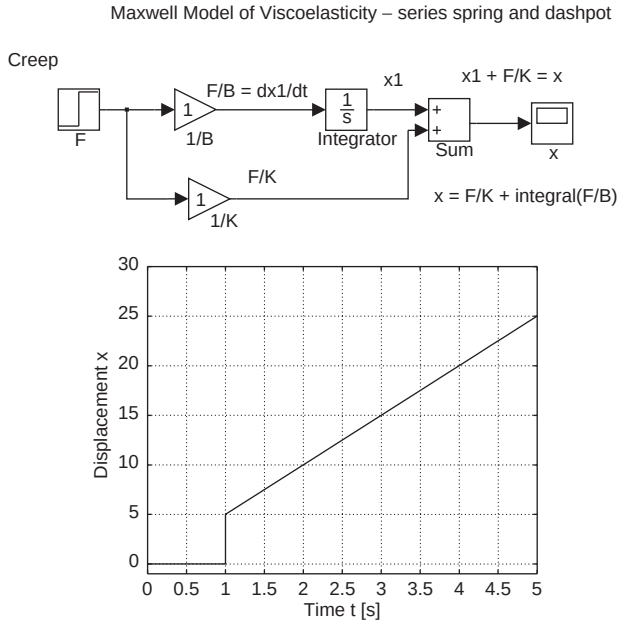


FIGURE 4.22 Creep of the Maxwell viscoelastic model, a series combination of ideal spring and dashpot (Figure 4.19a). The spring constant $K=1$, and dashpot damping coefficient $B=1$ (arbitrary units). This system is subjected to a step change in force, and displacement x arises by solving $x = F/K + \int F/B$. The spring instantly responds, followed by creep of the ideal dashpot, which may extend as long as force is applied.

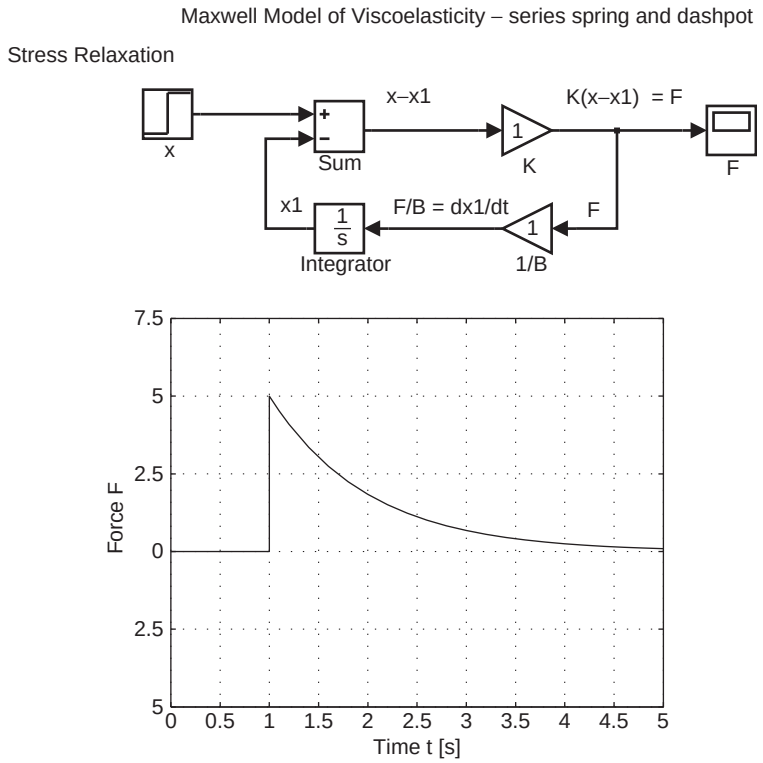


FIGURE 4.23 Stress relaxation of the Maxwell viscoelastic model. This model solves $F = K[x - \int F/B]$, again with K=B=1 and arbitrary units. The ideal spring instantly responds followed by stress relaxation via the dashpot to the steady-state force level.

William Thompson (Lord Kelvin, 1824–1907) used the three-element viscoelastic model (Figure 4.19c) to describe the mechanical properties of different solids in the form of a torsional pendulum. Figure 4.26 shows the three-element Kelvin model's creep response. This model has an initial rapid jump in position with subsequent slow creep. Figure 4.27 shows the Kelvin model stress relaxation test. Initially, the material is very stiff, with subsequent stress decay to a nonzero steady-state level that is due to the extension of the dashpot. The three-element Kelvin model is the simplest lumped viscoelastic model that is bounded both in extension and force.

The three-element viscoelastic model describes the basic features of stress relaxation and creep. Biological materials often exhibit more complex viscoelastic properties. For example, plotting hysteresis as a function of frequency of applied strain gives discrete curves for the lumped viscoelastic models. Biological tissues demonstrate broad, distributed hysteresis properties. One solution is to describe biomaterials with a distributed network of three-element models. A second method is to use the generalized viscoelastic model of Westerhof and Noordergraaf (1970) to describe the viscoelastic wall properties of blood vessels. Making the elastic modulus mathematically complex yields a model that includes the frequency dependent elastic modulus, stress relaxation, creep, and hysteresis exhibited by arteries. Further, the Voight and Maxwell models emerge as special (limited) cases of this general approach.

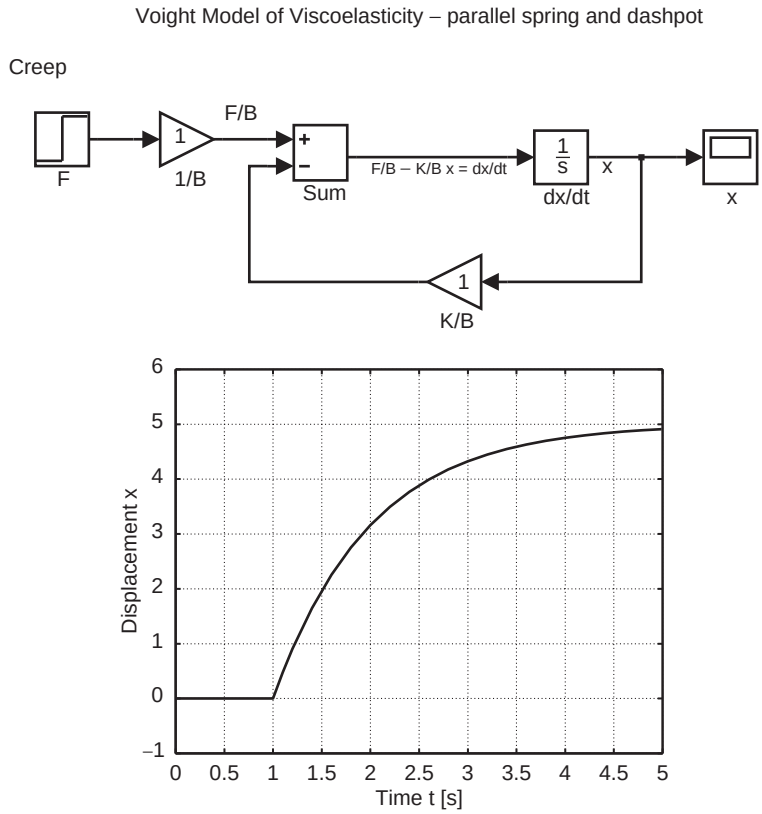


FIGURE 4.24 Creep of the Voight viscoelastic model, a parallel combination of ideal spring and dashpot (Figure 4.19b). This model solves the differential equation $dx/dt = 1/B[F - Kx]$ for x. K=B=1, and the step applied force is 5 arbitrary units. Displacement slowly creeps toward its steady-state value.

4.5 CARTILAGE, LIGAMENT, TENDON, AND MUSCLE

The articulating surfaces of bones are covered with articular cartilage, a biomaterial composed mainly of collagen. Collagen is the main structural material of hard and soft tissues in animals. Isolated collagen fibers have high tensile strength that is comparable to nylon (50–100 MPa) and an elastic modulus of approximately 1 GPa. Elastin is a protein found in vertebrates and is particularly important in blood vessels and the lungs. Elastin is the most linearly elastic biosolid known, with an elastic modulus of approximately 0.6 MPa. It gives skin and connective tissue their elasticity.

4.5.1 Cartilage

Cartilage serves as the bearing surfaces of joints. It is porous, and its complex mechanical properties arise from the motion of fluid in and out of the tissue when subjected to joint loading. Consequently, articular cartilage is strongly viscoelastic, with stress relaxation times

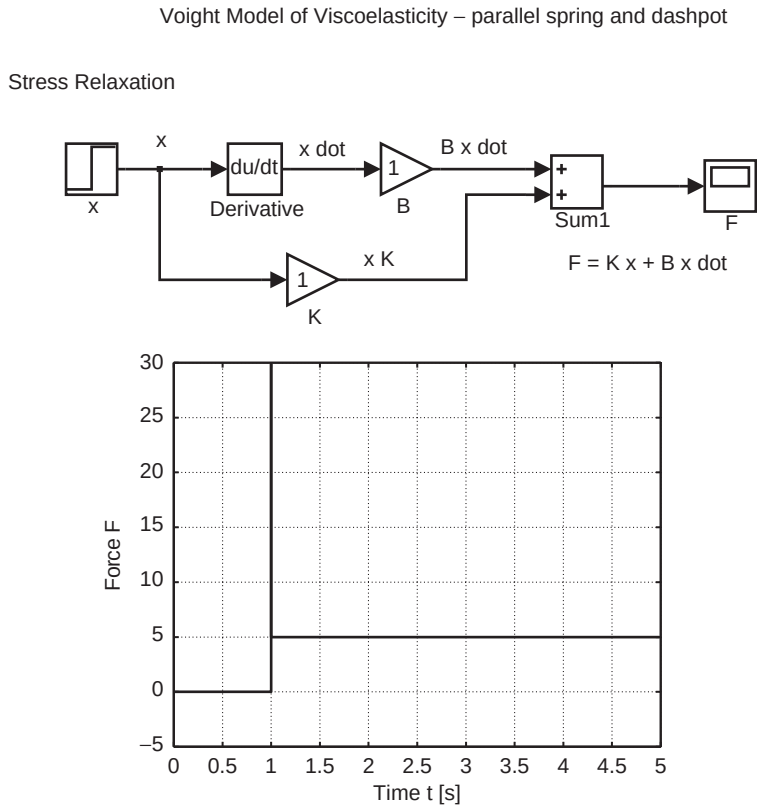


FIGURE 4.25 Stress relaxation of the Voight viscoelastic model. This model solves the equation $F = Kx + Bdx/dt$. Since the dashpot is in parallel with the spring, and since it cannot respond immediately to a step change in length, the model force goes to infinity.

in compression on the order of 1 to 5 seconds. Cartilage is anisotropic and displays hysteresis during cyclical loading. The ultimate compressive stress of cartilage is on the order of 5 MPa.

4.5.2 Ligaments and Tendons

Ligaments join bones together and consequently serve as part of the skeletal framework. Tendons join muscles to bones and transmit forces generated by contracting muscles to cause movement of the jointed limbs. Tendons and ligaments primarily transmit tension, so they are composed mainly of parallel bundles of collagen fibers and have similar mechanical properties. Human tendon has an ultimate stress of 50–100 MPa and exhibits very nonlinear stress-strain curves. The middle stress-strain range is linear, with an elastic modulus of approximately 1–2 GPa. Both tendons and ligaments exhibit hysteresis, viscoelastic creep, and stress relaxation. These materials may also be "preconditioned," whereby initial tensile loading can affect subsequent load-deformation curves. The material properties shift due to changes in the internal tissue structure with repeated loading.

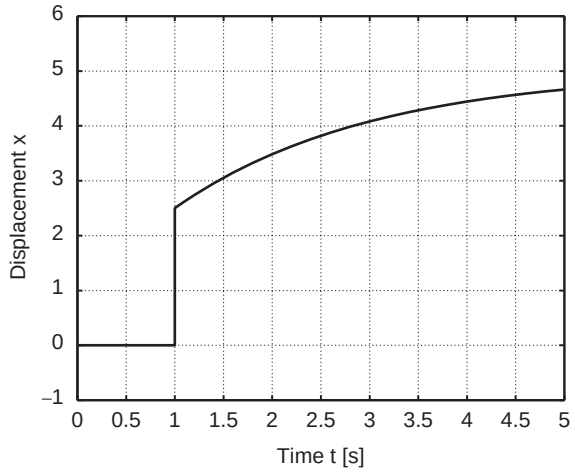


FIGURE 4.26 Creep of the Kelvin three-element viscoelastic model. This model's equations of motion are left to the reader to derive. After a step change in force, this model has an initial immediate increase in displacement, with a subsequent slow creep to a steady-state level.

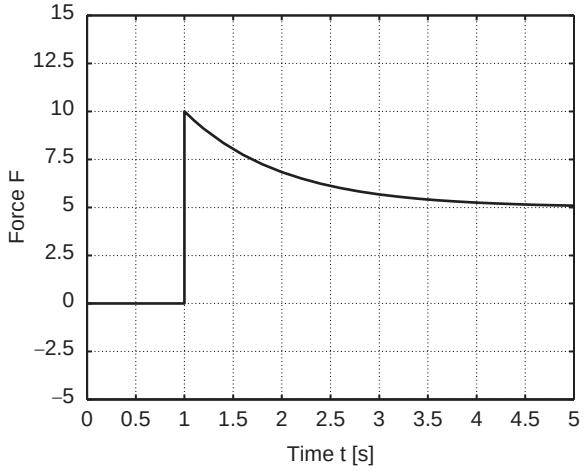


FIGURE 4.27 Stress relaxation of the Kelvin viscoelastic model. This model has an initial immediate increase in force followed by slower stress relaxation to a steady-state force level.

4.5.3 Muscle Mechanics

Chapter 3 introduced muscle as an active, excitable tissue that generates force by forming cross-bridge bonds between the interdigitating actin and myosin myofilaments. The quantitative description of muscle contraction has evolved into two separate foci: lumped descriptions based on A. V. Hill's contractile element and cross-bridge models based on A. F. Huxley's description of a single sarcomere [22]. The earliest quantitative descriptions of muscle are lumped whole muscle models, with the simplest mechanical description being a

purely elastic spring. Potential energy is stored when the spring is stretched, and shortening occurs when it is released. The idea of muscle elastance can be traced back to Ernst Weber [29], who considered muscle as an elastic material that changes state during activation via conversion of chemical energy. Subsequently, investigators retained the elastic description but ignored metabolic alteration of muscle stiffness. A purely elastic model of muscle can be refuted on thermodynamic grounds, since the potential energy stored during stretching is less than the sum of the energy released during shortening as work and heat. Still, efforts to describe muscle by a combination of traditional springs and dashpots continued. In 1922, Hill coupled the spring with a viscous medium, thereby reintroducing viscoelastic muscle descriptions that can be traced back to the 1840s.

Quick stretch and release experiments show that muscle's viscoelastic properties are strongly time dependent. In general, the faster a change in muscle length occurs, the more severely the contractile force is disturbed. Muscle contraction clearly arises from a more sophisticated mechanism than a damped elastic spring. In 1935, Fenn and Marsh added a series elastic element to Hill's damped elastic model and concluded that "muscle cannot properly be treated as a simple mechanical system." Subsequently, Hill embodied the empirical hyperbolic relation between load and initial velocity of shortening for skeletal muscle as a model building block, denoted the contractile element. Hill's previous viscoelastic model considered muscle to possess a fixed amount of potential energy whose rate of release is controlled by viscosity. Energy is now thought to be controlled by some undefined internal mechanism rather than by friction. This new feature of muscle dynamics varying with load was a step in the right direction; however, subsequent models, including heart studies, built models based essentially on the hyperbolic curve that was measured for tetanized skeletal muscle. This approach can be criticized on two grounds: (1) embodiment of the contractile element by a single force-velocity relation sets a single, fixed relation between muscle energetics and force; and (2) it yields no information on the contractile mechanism behind this relation. Failure of the contractile element to describe a particular loading condition led investigators to add passive springs and dashpots liberally, with the number of elements reaching at least nine by the late 1960s. Distributed models of muscle contraction, generally, have been conservative in design and have depended fundamentally on the Hill contractile element. Recent models are limited to tetanized, isometric contractions or to isometric twitch contractions.

A second, independent focus of muscle contraction research works at the ultrastructural level, with the sliding filament theory serving as the most widely accepted contraction mechanism. Muscle force generation is viewed as the result of crossbridge bonds formed between thick and thin filaments at the expense of biochemical energy. The details of bond formation and detachment are under considerable debate, with the mechanism for relaxation particularly uncertain. Prior to actual observation of crossbridges, A. F. Huxley [13] devised the crossbridge model based on structural and energetic assumptions. Bonds between myofilaments are controlled via rate constants f and g that dictate attachment and detachment, respectively. One major shortcoming of this idea was the inability to describe transients resulting from rapid changes in muscle length or load, similar to the creep and stress relaxation tests previously discussed.

Subsequent models adopt increasingly complex bond attachment and detachment rate functions and are often limited in scope to description of a single pair of myofilaments. Each tends

to focus on description of a single type of experiment. No model has been shown to broadly describe all types of contractile loading conditions. Crossbridge models have tended to rely on increasingly complex bond attachment and detachment rate functions. This trend has reversed the issue of describing complex muscle dynamics from the underlying, simpler crossbridges to adopting complex crossbridge dynamics to describe a particular experiment.

Alternatively, Palladino and Noordergraaf [22] proposed a large-scale, distributed muscle model that manifests both contraction and relaxation as the result of fundamental mechanical properties of crossbridge bonds. As such, muscle's complex contractile properties emerge from its underlying ultrastructure dynamics—that is, function follows from structure. Bonds between myofilaments, which are biomaterials, are described as viscoelastic material. The initial stimulus for contraction is electrical. Electrical propagation through cardiac muscle occurs at finite speed, implying spatial asynchrony of stimulation. Furthermore, Ca^{++} release from the sarcoplasmic reticulum depends on diffusion for availability at the myosin heads. These effects, as well as nonuniformity of structure, strongly suggest that contraction is asynchronous throughout the muscle. Recognition of muscle's distributed properties by abandoning the assumption of perfect synchrony in contraction and consideration of myofilament mass allow for small movements of thick with respect to thin filaments. Such movements lead to bond detachment and heat production. Gross movement such as muscle shortening exacerbates this process. Quick transients in muscle length or applied load have particularly strong effects and have been observed experimentally. Muscle relaxation is thereby viewed as a consequence of muscle's distributed properties.

The distributed muscle model is built from the following main features: sarcomeres consist of overlapping thick and thin filaments connected by crossbridge bonds that form during activation and detach during relaxation. Figure 4.28 shows a schematic of a muscle fiber composed of a string of series sarcomeres. Crossbridge bonds are each described as three-element viscoelastic solids, and myofilaments as masses. Force is generated due to viscoelastic crossbridge bonds that form and are stretched between the interdigitating matrix of myofilaments.

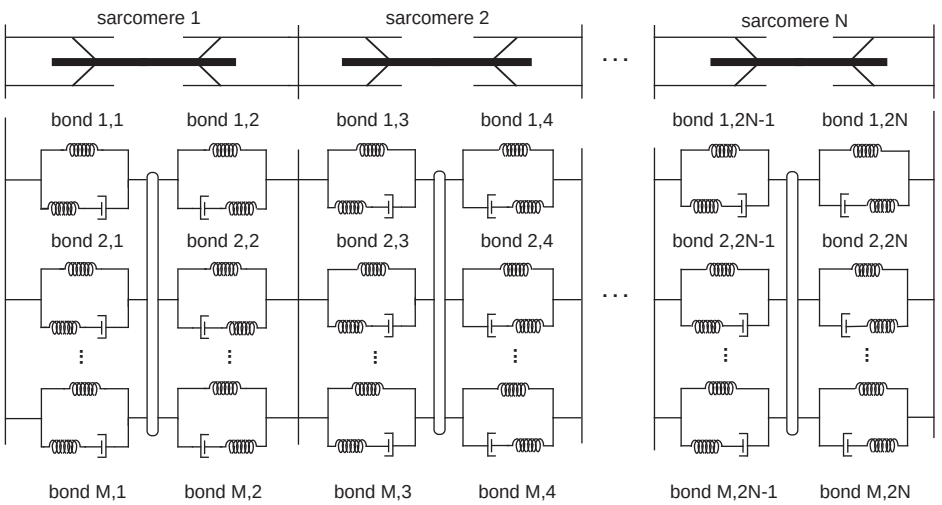


FIGURE 4.28 Schematic diagram of a muscle fiber built from a distributed network of N sarcomeres. Each sarcomere has M parallel pairs of crossbridge bonds. *Adapted from [22].*

The number of bonds formed depends on the degree of overlap between thick and thin filaments and is dictated spatially and temporally due to finite electrical and chemical activation rates. Asynchrony in bond formation and unequal numbers of bonds formed in each half sarcomere, as well as mechanical disturbances such as muscle shortening and imposed length transients, cause small movements of the myofilaments. Since myofilament masses are taken into account, these movements take the form of damped vibrations with a spectrum of frequencies due to the distributed system properties. When the stress in a bond goes to zero, the bond detaches. Consequently, myofilament motion and bond stress relaxation lead to bond detachment and produce relaxation without assumption of bond detachment rate functions. In essence, relaxation results from inherent system instability. Although the model is built from linear, time-invariant components (springs, dashpots, and masses), the highly dynamic structure of the model causes its mechanical properties to be highly nonlinear and time-varying, as is found in muscle fibers and strips.

Sensitivity of the model to mechanical disturbances is consistent with experimental evidence from muscle force traces, aequorin measurements of free calcium ion, and high-speed x-ray diffraction studies, which all suggest enhanced bond detachment. The model is also consistent with sarcomere length feedback studies in which reduced internal motion delays relaxation, and it predicted muscle fiber (cell) dynamics prior to their experimental measurement.

This model proposes a structural mechanism for the origin of muscle's complex mechanical properties and predicts new features of the contractile mechanism—for example, a mechanism for muscle relaxation and prediction of muscle heat generation. This approach computes muscle's complex mechanical properties from physical description of muscle anatomical structure, thereby linking subcellular structure to organ-level function.

This chapter describes some of the high points of biological tissues' mechanical properties. More comprehensive references include Fung's *Biomechanics: Mechanical Properties of Living Tissues*, Nigg and Herzog's *Biomechanics of the Musculo-Skeletal System*, and Mow and Hayes's *Basic Orthopaedic Biomechanics*. Muscle contraction research has a long history, as chronicled in the book *Machina Carnis* by Needham. For a more comprehensive history of medicine, see Singer and Underwood's book. The next two sections apply biomechanics concepts introduced in Sections 4.2–4.5 to human gait analysis and to the quantitative study of the cardiovascular system.

4.6 CLINICAL GAIT ANALYSIS

An example of applied dynamics in human movement analysis is clinical gait analysis. Clinical gait analysis involves the measurement of the parameters that characterize a patient's gait pattern, the interpretation of the collected and processed data, and the recommendation of treatment alternatives. It is a highly collaborative process that requires the cooperation of the patient and the expertise of a multidisciplinary team that typically includes a physician, a physical therapist or kinesiologist, and an engineer or technician. The engineer is presented with a number of challenges. The fundamental objective in data collection is to monitor the patient's movements accurately and with sufficient precision for clinical use without altering the patient's typical performance. While measurement devices for clinical gait analysis are established to some degree and are commercially available, the

protocols for the use of the equipment continue to develop. The validity of these protocols and associated models and the care with which they are applied ultimately dictate the meaning and quality of the resulting data provided for interpretation. This is one area in which engineers in collaboration with their clinical partners can have a significant impact on the clinical gait analysis process.

Generally, data collection for clinical gait analysis involves the placement of highly reflective markers on the surface of the patient's skin. These external markers reflect light to an array of video-based motion cameras that surround the measurement volume. The instantaneous location of each of these markers can then be determined stereometrically based on the images obtained simultaneously from two or more cameras. Other aspects of gait can be monitored as well, including ground reactions via force platforms embedded in the walkway and muscle activity via electromyography with either surface or intramuscular fine wire electrodes, depending on the location of the particular muscle.

In keeping with the other material presented in this chapter, the focus of this section will pertain to the biomechanical aspects of clinical gait analysis and includes an outline of the computation of segmental and joint kinematics and joint kinetics, and a brief illustration of how the data are interpreted.

4.6.1 The Clinical Gait Model

The gait model is the algorithm that transforms the data collected during walking trials into the information required for clinical interpretation. For example, the gait model uses the data associated with the three-dimensional displacement of markers on the patient to compute the angles that describe how the patient's body segment and lower-extremity joints are moving. The design of the gait model is predicated on a clear understanding of the needs of the clinical interpretation team—for example, the specific aspects of gait dynamics of interest. To meet these clinical specifications, gait model development is constrained both by the technical limitations of the measurement system and by the broad goal of developing protocols that may be appropriate for a wide range of patient populations that vary in age, gait abnormality, walking ability, and so on. An acceptable model must be sufficiently general to be used for many different types of patients (e.g., adults and children with varying physical and cognitive involvement), be sufficiently sophisticated to allow detailed biomechanical questions to be addressed, and be based on repeatable protocols that are feasible in a clinical setting.

4.6.2 Kinematic Data Analysis

Reflective markers placed on the surface of the patient's skin are monitored or tracked in space and time by a system of video-based cameras. These marker trajectories are used to compute coordinate systems that are anatomically aligned and embedded in each body segment under analysis. These anatomical coordinate systems provide the basis for computing the absolute spatial orientation, or attitude, of the body segment or the angular displacement of one segment relative to another, such as joint angles. For this analysis, at least three noncollinear markers or points of reference must be placed on or identified for each body segment included in the analysis. These markers form a plane from which a segmentally fixed coordinate system may be derived. Any three markers will allow the segment motion to be monitored, but unless these markers are referenced to the subject's anatomy, such

kinematic quantification is of limited clinical value. Markers must either be placed directly over palpable bony landmarks on the segment or at convenient (i.e., visible to the measurement cameras) locations on the segment that are referenced to the underlying bones. An examination of the pelvic and thigh segments illustrates these two alternatives.

Pelvic Anatomical Coordinate System

For the pelvis, markers placed over the right and left anterior-superior-iliac-spine (ASIS) and either the right or left posterior-superior-iliac-spine (PSIS) will allow for the computation of an anatomically aligned coordinate system, as described in Example Problem 4.10.

EXAMPLE PROBLEM 4.10

Given the following three-dimensional locations in meters for a set of pelvic markers expressed relative to an inertially fixed laboratory coordinate system (Figure 4.29),

$$\begin{aligned} \text{Right ASIS}: \mathbf{RASIS} &= -0.850\mathbf{i} - 0.802\mathbf{j} + 0.652\mathbf{k} \\ \text{Left ASIS}: \mathbf{LASIS} &= -0.831\mathbf{i} - 0.651\mathbf{j} + 0.652\mathbf{k} \\ \mathbf{PSIS} &= -1.015\mathbf{i} - 0.704\mathbf{j} + 0.686\mathbf{k} \end{aligned}$$

compute an anatomical coordinate system for the pelvis.

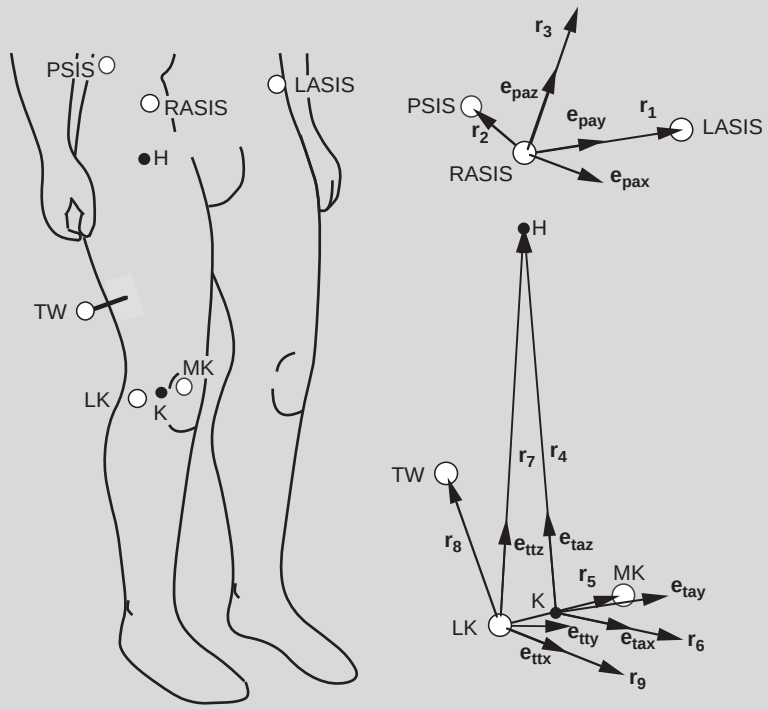


FIGURE 4.29 Kinematic markers used to define pelvis and thigh coordinate systems. For the pelvis, PSIS denotes posterior-superior-iliac-spine, H is hip center, and RASIS and LASIS denote right and left anterior-superior-iliac-spine markers, respectively. For the thigh, TW is thigh wand, K is knee center, and MK and LK are medial and lateral knee (femoral condyle) markers, respectively.

Continued

Solution

These three anatomical markers form a plane. The line between the right ASIS and left ASIS represents one coordinate system axis. Another coordinate axis is perpendicular to the pelvic plane. The third coordinate axis is computed to be orthogonal to the first two:

1. Subtract vector **RASIS** from vector **LASIS**,

$$\mathbf{LASIS} - \mathbf{RASIS} = (-0.831 - (-0.850))\mathbf{i} + (-0.651 - (-0.802))\mathbf{j} + (0.652 - 0.652)\mathbf{k}$$

to find

$$\mathbf{r}_1 = 0.0190\mathbf{i} + 0.1510\mathbf{j} + 0.0000\mathbf{k}$$

and its associated unit vector:

$$\mathbf{e}_{r1} = \frac{0.019\mathbf{i} + 0.151\mathbf{j} + 0.000\mathbf{k}}{\sqrt{0.019^2 + 0.151^2 + 0.000^2}}$$
$$\mathbf{e}_{r1} = 0.125\mathbf{i} + 0.992\mathbf{j} + 0.000\mathbf{k}$$

Unit vector $\mathbf{e_{r1}}$ represents the medial-lateral direction or *y*-axis for the pelvic anatomical coordinate system $\mathbf{e_{pay}}$ (Figure 4.29).

2. A second vector in the pelvic plane is required to compute the coordinate axis that is perpendicular to the plane. Consequently, subtract vector **RASIS** from vector **PSIS** to find

$$\mathbf{r}_2 = -0.165\mathbf{i} + 0.098\mathbf{j} + 0.034\mathbf{k}$$

3. Take the vector cross product $\mathbf{e_{pay}} \times \mathbf{r}_2$ to yield

$$\mathbf{r}_3 = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0.125 & 0.992 & 0.000 \\ -0.165 & 0.098 & 0.034 \end{vmatrix}$$
$$= [(0.992)(0.034) - (0.000)(0.098)]\mathbf{i}$$
$$+ [(0.000)(-0.165) - (0.125)(0.034)]\mathbf{j}$$
$$+ [(0.125)(0.098) - (0.992)(-0.165)]\mathbf{k}$$
$$= 0.034\mathbf{i} - 0.004\mathbf{j} + 0.176\mathbf{k}$$

and its associated unit vector:

$$\mathbf{e}_{r3} = \mathbf{e_{paz}} = 0.188\mathbf{i} - 0.024\mathbf{j} + 0.982\mathbf{k}$$

Unit vector $\mathbf{e_{r3}}$ represents the anterior-superior direction or *z*-axis of the pelvic anatomical coordinate system $\mathbf{e_{paz}}$ (Figure 4.29).

4. The third coordinate axis is computed to be orthogonal to the first two. Take the vector cross product $\mathbf{e_{pay}} \times \mathbf{e_{paz}}$ to compute the fore-aft direction, or *x*-axis, of the pelvic anatomical coordinate system:

$$\mathbf{e_{pax}} = 0.974\mathbf{i} - 0.123\mathbf{j} - 0.190\mathbf{k}$$

For this example, the anatomical coordinate system for the pelvis can be expressed as follows:

$$\{\mathbf{e_{pa}}\} = \begin{bmatrix} \mathbf{e_{pax}} \\ \mathbf{e_{pay}} \\ \mathbf{e_{paz}} \end{bmatrix} = \begin{bmatrix} 0.974 & -0.123 & -0.190 \\ 0.125 & 0.992 & 0.000 \\ 0.188 & -0.024 & 0.982 \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix}$$

Note that the coefficients associated with these three axes represent the direction cosines that define the orientation of the pelvic coordinate system relative to the laboratory coordinate system.

In summary, by monitoring the motion of the three pelvic markers, the instantaneous orientation of an anatomical coordinate system for the pelvis, $\{\mathbf{e_{pa}}\}$, comprising axes $\mathbf{e_{pax}}$, $\mathbf{e_{pay}}$, and $\mathbf{e_{paz}}$, can be determined. The absolute angular displacement of this coordinate system can then be computed via Euler angles as pelvic tilt, obliquity, and rotation using Eqs. (4.32)–(4.34). An example of these angle computations is presented later in this section.

Thigh Anatomical Coordinate System

The thigh presents a more significant challenge than the pelvis because three bony anatomical landmarks are not readily available as reference points during gait. A model based on markers placed over the medial and lateral femoral condyles and the greater trochanter is appealing but ill-advised. A marker placed over the medial femoral condyle is not always feasible during gait—for example, with patients whose knees make contact while walking. A marker placed over the greater trochanter is often described in the literature but should not be used as a reference because of its significant movement relative to the underlying greater trochanter during gait—that is, skin motion artifact [3].

In general, the approach used to quantify thigh motion, and the shank and foot, is to place additional anatomical markers on the segments during a static subject calibration process. Then the relationship between these static anatomical markers, which are removed before gait data collection, and the motion markers that remain on the patient during gait data collection may be calculated. It is assumed that this mathematical relationship remains constant during gait—that is, the instrumented body segments are assumed to be rigid. This process is illustrated in the Example Problem 4.11.

EXAMPLE PROBLEM 4.11

Given the following marker coordinate data that have been acquired while the patient stands quietly, also in meters,

lateral femoral condyle marker $\mathbf{LK} = -0.881\mathbf{i} - 0.858\mathbf{j} + 0.325\mathbf{k}$
medial femoral condyle marker $\mathbf{MK} = -0.855\mathbf{i} - 0.767\mathbf{j} + 0.318\mathbf{k}$

compute an anatomical coordinate system for the thigh.

Solution

A thigh plane is formed based on three anatomical markers or points: the hip center, the lateral femoral condyle marker **LK**, and the medial femoral condyle marker **MK**. The knee center location can then be estimated as the midpoint between **LK** and **MK**. With these points, the vector

Continued

from the knee center to the hip center represents the longitudinal axis of the coordinate system. A second coordinate axis is perpendicular to the thigh plane. The third coordinate axis is computed to be orthogonal to the first two.

The location of the knee center of rotation may be approximated as the midpoint between the medial and lateral femoral condyle markers,

$$\frac{\mathbf{LK}+\mathbf{MK}}{2}=\frac{(-0.881)+(-0.855)}{2}\mathbf{i}+\frac{(-0.858)+(-0.767)}{2}\mathbf{j}+\frac{(0.325)+(0.318)}{2}\mathbf{k}$$

yielding

$$\text{knee center location } \mathbf{K}=-0.868\mathbf{i}-0.812\mathbf{j}+0.321\mathbf{k}$$

The location of the center of the head of the femur, referred to as the hip center, is approximated based either on patient anthropometry and a statistical model of pelvic geometry [6] or from data collected while the subject executes a specific hip movement [8]. In this example, the hip center can be located at approximately

$$\text{hip center location } \mathbf{H}=-0.906\mathbf{i}-0.763\mathbf{j}+0.593\mathbf{k}$$

Now the anatomical coordinate system for the thigh may be computed as follows:

1. Subtract the vector **K** from **H**, giving

$$\mathbf{r}_4=-0.038\mathbf{i}+0.049\mathbf{j}+0.272\mathbf{k}$$

and its associated unit vector

$$\mathbf{e}_{\mathbf{r4}}=\mathbf{e}_{\mathbf{taz}}=-0.137\mathbf{i}+0.175\mathbf{j}+0.975\mathbf{k}$$

Unit vector $\mathbf{e}_{\mathbf{r4}}$ represents the longitudinal direction, or z-axis, of the thigh anatomical coordinate system $\mathbf{e}_{\mathbf{taz}}$.

2. As with the pelvis, a second vector in the thigh plane is required to compute the coordinate axis that is perpendicular to the plane. Consequently, subtract vector **LK** from **MK**:

$$\mathbf{r}_5=0.026\mathbf{i}+0.091\mathbf{j}-0.007\mathbf{k}$$

3. Form the vector cross product $\mathbf{r}_5\times\mathbf{e}_{\mathbf{taz}}$ to yield

$$\mathbf{r}_6=0.090\mathbf{i}-0.024\mathbf{j}+0.017\mathbf{k}$$

and its associated unit vector

$$\mathbf{e}_{\mathbf{r6}}=\mathbf{e}_{\mathbf{tax}}=0.949\mathbf{i}-0.258\mathbf{j}+0.180\mathbf{k}$$

Unit vector $\mathbf{e}_{\mathbf{r6}}$ represents the fore-aft direction, or x-axis, of the thigh anatomical coordinate system $\mathbf{e}_{\mathbf{tax}}$.

4. Again, the third coordinate axis is computed to be orthogonal to the first two. Determine the medial-lateral or y-axis of the thigh anatomical coordinate system, $\mathbf{e}_{\mathbf{tay}}$, from the cross product $\mathbf{e}_{\mathbf{taz}}\times\mathbf{e}_{\mathbf{tax}}$:

$$\mathbf{e}_{\mathbf{tay}}=0.284\mathbf{i}+0.950\mathbf{j}-0.131\mathbf{k}$$

For this example, the anatomical coordinate system for the thigh can be expressed as

$$\{\mathbf{e_{ta}}\} = \begin{bmatrix} \mathbf{e_{tax}} \\ \mathbf{e_{tay}} \\ \mathbf{e_{taz}} \end{bmatrix} = \begin{bmatrix} 0.949 & -0.258 & 0.180 \\ 0.284 & 0.950 & -0.131 \\ -0.137 & 0.175 & 0.975 \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix}$$

This defines an anatomical coordinate system fixed to the thigh, $\{\mathbf{e_{ta}}\}$, comprising axes $\mathbf{e_{tax}}$, $\mathbf{e_{tay}}$, and $\mathbf{e_{taz}}$. Its basis, however, includes an external marker (medial femoral condyle **MK**) that must be removed before the walking trials. Consequently, the location of the knee center cannot be computed as described in the preceding example. This dilemma is resolved by placing another marker on the surface of the thigh such that it also forms a plane with the hip center and lateral knee marker. These three reference points can then be used to compute a "technical" coordinate system for the thigh to which the knee center location may be mathematically referenced.

EXAMPLE PROBLEM 4.12

Continuing Example Problem 4.11, and given the coordinates of another marker placed on the thigh but not anatomically aligned,

$$\text{thigh wand marker } \mathbf{TW} = -0.890\mathbf{i} - 0.937\mathbf{j} + 0.478\mathbf{k}$$

compute a technical coordinate system for the thigh.

Solution

A technical coordinate system for the thigh can be computed as follows:

1. Compute the longitudinal direction, or z-axis, of the technical thigh coordinate system $\mathbf{e_{tt}}$. Start by subtracting vector **LK** from the hip center **H** to form

$$\mathbf{r}_7 = -0.025\mathbf{i} + 0.094\mathbf{j} + 0.268\mathbf{k}$$

and its associated unit vector

$$\mathbf{e}_{r7} = \mathbf{e_{ttz}} = -0.088\mathbf{i} + 0.330\mathbf{j} + 0.940\mathbf{k}$$

Unit vector $\mathbf{e_{r7}}$ represents the z-axis of the thigh technical coordinate system, $\mathbf{e_{ttz}}$.

2. To compute the axis that is perpendicular to the plane formed by **LK**, **H**, and **TW**, subtract vector **LK** from **TW** to compute

$$\mathbf{r}_8 = -0.009\mathbf{i} - 0.079\mathbf{j} + 0.153\mathbf{k}$$

3. Calculate the vector cross product $\mathbf{r}_7 \times \mathbf{r}_8$ to yield

$$\mathbf{r}_9 = 0.036\mathbf{i} + 0.001\mathbf{j} + 0.003\mathbf{k}$$

with its associated unit vector

$$\mathbf{e}_{r9} = \mathbf{e_{ttx}} = 0.996\mathbf{i} + 0.040\mathbf{j} + 0.079\mathbf{k}$$

Unit vector $\mathbf{e_{r9}}$ represents the fore-aft direction, or x-axis, of the thigh technical coordinate system $\mathbf{e_{ttx}}$.

Continued

4. The third coordinate axis is computed to be orthogonal to the first two axes. Compute the vector cross product $\mathbf{e}_{\mathbf{ttz}} \times \mathbf{e}_{\mathbf{ttx}}$ to determine the medial-lateral direction, or y-axis, of the thigh technical coordinate system:

$$\mathbf{e}_{\mathbf{tty}} = \mathbf{e}_{\mathbf{ttz}} \times \mathbf{e}_{\mathbf{ttx}} = -0.012\mathbf{i} + 0.943\mathbf{j} + 0.332\mathbf{k}$$

For this example, the technical coordinate system for the thigh can be expressed as

$$\{\mathbf{e}_{\mathbf{tt}}\} = \begin{bmatrix} \mathbf{e}_{\mathbf{ttx}} \\ \mathbf{e}_{\mathbf{tty}} \\ \mathbf{e}_{\mathbf{ttz}} \end{bmatrix} = \begin{bmatrix} 0.996 & 0.040 & 0.079 \\ -0.012 & 0.943 & 0.332 \\ -0.088 & 0.330 & 0.940 \end{bmatrix} \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix}$$

Note that this thigh technical coordinate system $\{\mathbf{e}_{\mathbf{tt}}\}$ computed during the standing subject calibration can also be computed from each camera frame of walking data. That is, its computation is based on markers (the lateral femoral condyle and thigh wand markers) and an anatomical landmark (the hip center) that are available for both the standing and walking trials. Consequently, the technical coordinate system $\{\mathbf{e}_{\mathbf{tt}}\}$ becomes the embedded reference coordinate system to which other entities can be related. The thigh anatomical coordinate system $\{\mathbf{e}_{\mathbf{ta}}\}$ can be related to the thigh technical coordinate system $\{\mathbf{e}_{\mathbf{tt}}\}$ by using either direction cosines or Euler angles, as described in Section 4.2.2. Also, the location of markers that must be removed after the standing subject calibration (e.g., the medial femoral condyle marker MK), or computed anatomical locations (e.g., the knee center) can be transformed into the technical coordinate system $\{\mathbf{e}_{\mathbf{tt}}\}$ and later retrieved for use in walking trial data reduction.

Segment and Joint Angles

Tracking the anatomical coordinate system for each segment allows for the determination of either the absolute angular orientation, or attitude, of each segment in space or the angular position of one segment relative to another. In the preceding example, the three pelvic angles that define the position of the pelvic anatomical coordinate system $\{\mathbf{e}_{\mathbf{pa}}\}$ relative to the laboratory (inertially fixed) coordinate system can be computed from the Euler angles, as described in Section 4.2.2 with Eqs. (4.32)–(4.34). Note that in these equations the laboratory coordinate system represents the proximal (unprimed) coordinate system, and the pelvic anatomical coordinate system $\{\mathbf{e}_{\mathbf{pa}}\}$ represents the distal (triple primed) coordinate system. Consequently, Eq. (4.32)

$$\theta_x = -\arcsin(\mathbf{k}''' \cdot \mathbf{j})$$

becomes

$$\begin{aligned} \theta_x &= -\arcsin(\mathbf{e}_{\mathbf{paz}} \cdot \mathbf{j}) \\ &= -\arcsin((0.188\mathbf{i} - 0.024\mathbf{j} + 0.982\mathbf{k}) \cdot \mathbf{j}) \\ &= -\arcsin(-0.024) \\ &= 1^\circ \text{of pelvic obliquity} \end{aligned}$$

Similarly, Eq. (4.33)

$$\theta_y = \arcsin\left(\frac{(\mathbf{k}''' \cdot \mathbf{i})}{\cos\theta_x}\right)$$

becomes

$$\begin{aligned}\theta_y &= \arcsin\left(\frac{(\mathbf{e_{paz}} \cdot \mathbf{i})}{\cos\theta_x}\right)\\ &= \arcsin\left(\frac{(0.188\mathbf{i} - 0.024\mathbf{j} + 0.982\mathbf{k}) \cdot \mathbf{i}}{\cos 1^\circ}\right)\\ &= \arcsin\left(\frac{0.188}{\cos 1^\circ}\right)\\ &= 11^\circ \text{of anterior pelvic tilt}\end{aligned}$$

and Eq. (4.34)

$$\theta_z = \arcsin\left(\frac{(\mathbf{i}''' \cdot \mathbf{j})}{\cos\theta_x}\right)$$

becomes

$$\begin{aligned}\theta_z &= \arcsin\left(\frac{(\mathbf{e_{pax}} \cdot \mathbf{j})}{\cos\theta_x}\right)\\ &= \arcsin\left(\frac{(0.974\mathbf{i} - 0.123\mathbf{j} - 0.190\mathbf{k}) \cdot \mathbf{j}}{\cos 1^\circ}\right)\\ &= \arcsin\left(\frac{-0.123}{\cos 1^\circ}\right)\\ &= -7^\circ \text{of pelvic rotation}\end{aligned}$$

This Euler angle computation may be repeated to solve for the three hip angles that define the position of the thigh anatomical coordinate system $\{\mathbf{e_{ta}}\}$ relative to the pelvic anatomical coordinate system $\{\mathbf{e_{pa}}\}$. For the hip angles, the proximal (unprimed) coordinate system is the pelvis and the distal (triple-primed) coordinate system is the thigh. Substituting the values of $\{\mathbf{e_{pa}}\}$ and $\{\mathbf{e_{ta}}\}$ from Example Problems 4.10 and 4.11 into Eq. (4.32) yields:

$$\begin{aligned}\theta_z &= -\arcsin(\mathbf{e_{taz}} \cdot \mathbf{e_{pay}})\\ &= -\arcsin((-0.137\mathbf{i} + 0.175\mathbf{j} + 0.975\mathbf{k}) \cdot (0.125\mathbf{i} + 0.992\mathbf{j} + 0.000\mathbf{k}))\\ &= \arcsin(0.156)\\ &= -9^\circ \text{of hip abduction-adduction}\end{aligned}$$

The negative sign is associated with hip adduction of the left thigh or hip abduction of the right thigh.

Further substitution of values of $\{\mathbf{e_{pa}}\}$ and $\{\mathbf{e_{ta}}\}$ into Eqs. (4.33) and (4.34) yields

$$\text{hip flexion-extension } \theta_y = 20^\circ$$
$$\text{hip internal-external rotation } \theta_z = -8^\circ$$

For hip internal-external rotation, the negative sign is associated with hip internal rotation of the left thigh or hip external rotation of the right thigh. A negative hip flexion-extension angle corresponds to hip extension, independent of side. This process may be repeated for other body segments such as the shank (lower leg), foot, trunk, arms, and head with the availability of properly defined anatomical coordinate systems.

4.6.3 Kinetic Data Analysis

The marker displacement or motion data provide an opportunity to appreciate segment and joint kinematics. Kinematic data can be combined with ground reaction data—that is, forces and torques and their points of application, which are referred to as the centers of pressure. Combined with estimates of segment mass and mass moments of inertia, the net joint reaction forces and moments may then be computed.

To illustrate the details of this computational process, consider the following determination of the reactions at the ankle (Figure 4.30) for an individual with mass of 25.2 kg. Data for one instant in the gait cycle are shown in the following table.

	Symbol	Units	x_{lab}	y_{lab}	z_{lab}
Ankle center location	$\mathbf{A}$	[m]	0.357	0.823	0.056
Toe marker location	$\mathbf{T}$	[m]	0.421	0.819	0.051
Center of pressure location	$\mathbf{CP}$	[m]	0.422	0.816	0.000
Ground reaction force vector	$\mathbf{F_g}$	[N]	3.94	−15.21	242.36
Ground reaction torque vector	$\mathbf{T_g}$	[N-m]	0.000	0.000	0.995
Foot anatomical coordinate system	$\mathbf{e_{fax}}$		0.977	–0.0624	−0.202
	$\mathbf{e_{fay}}$		0.0815	0.993	0.0877
	$\mathbf{e_{faz}}$		0.195	−0.102	0.975
Foot linear acceleration vector	$\mathbf{a_{foot}}$	[m/s^2]	2.09	−0.357	−0.266
Foot angular velocity vector	$\boldsymbol{\omega}_{\mathbf{foot}}$	[rad/s]	0.0420	2.22	−0.585
Foot angular acceleration vector	$\boldsymbol{\alpha}_{\mathbf{foot}}$	[rad/sec^2]	−0.937	8.85	−5.16
Ankle angular velocity vector	$\boldsymbol{\omega}_{\mathbf{ankle}}$	[rad/s]	−0.000759	1.47	0.0106

Anthropomorphic relationships presented in Table 4.1 are used to estimate the mass and mass moments of inertia of the foot, as well as the location of its center of gravity. The mass of the foot, m_{foot}, may be estimated to be 1.45 percent of the body mass, or 0.365 kg, and the location of the center of gravity is approximated as 50 percent of the foot length. The length

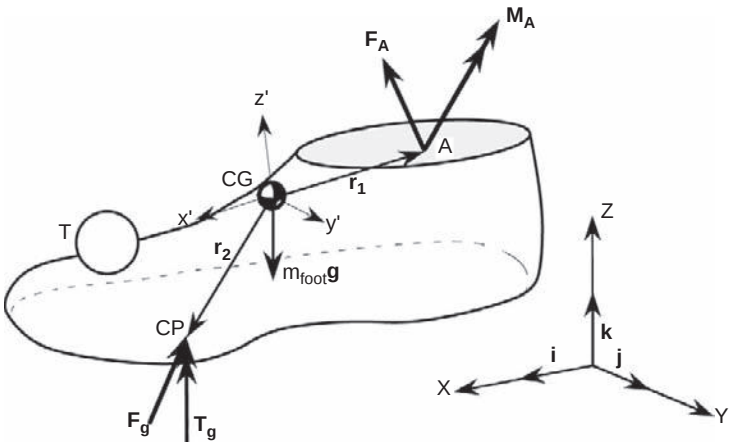


FIGURE 4.30 Ankle A and toe T marker data are combined with ground reaction force data $\mathbf{F_g}$ and segment mass and mass moment of inertia estimates to compute the net joint forces and moments.

of the foot ℓ_{foot} may be approximated as the distance between the ankle center and the toe marker, determined as follows:

$$\begin{aligned}\mathbf{T}-\mathbf{A} &= (0.421-0.357)\mathbf{i}+(0.819-0.823)\mathbf{j}+(0.051-0.056)\mathbf{k}\\ &= 0.064\mathbf{i}-0.004\mathbf{j}-0.005\mathbf{k}\\ \ell_{\text{foot}} &= |\mathbf{T}-\mathbf{A}|\\ &= \sqrt{(0.064)^2+(-0.004)^2+(-0.005)^2}=0.064\text{ m}\end{aligned}$$

Then the location of the center of gravity can be determined relative to the ankle center as

$$\mathbf{A}+\frac{\ell_{\text{foot}}}{2}\frac{(\mathbf{T}-\mathbf{A})}{|\mathbf{T}-\mathbf{A}|}=(0.357\mathbf{i}+0.823\mathbf{j}+0.056\mathbf{k})+\left[\frac{0.064}{2}\right]\left[\frac{0.064\mathbf{i}-0.004\mathbf{j}-0.005\mathbf{k}}{0.064}\right]$$

giving the location of the center of gravity:

$$\mathbf{CG}=0.389\mathbf{i}+0.821\mathbf{j}+0.054\mathbf{k}$$

which allows computation of position vectors $\mathbf{r_1}$ and $\mathbf{r_2}$ (see Figure 4.30). With a foot length of 0.064 m, a foot mass of 0.365 kg, and a proximal radius of gyration per segment length of 0.690, the mass moment of inertia relative to the ankle center may be estimated with Eq. (4.40) as

$$\begin{aligned}I_{\text{foot/ankle}} &= (0.365\text{ kg})[(0.690)(0.064\text{ m})]^2\\ &= 7.12\times10^{-4}\text{kg m}^2\end{aligned}$$

The centroidal mass moment of inertia, located at the foot's center of mass, may then be estimated using the parallel axis theorem (Eq. (4.41)):

$$I_{\text{foot/cm}}=I_{\text{foot/ankle}}-m_{\text{foot}}d^2$$

Note that the center of mass is equivalent to the center of gravity in a uniform gravitational field. In this case, d is the distance between the foot's center of mass and the ankle.

Table 4.1 shows the ratio of the foot center of mass location relative to its proximal end to be 0.5, so $d = 0.5(\ell_{\text{foot}}) = 0.032$ m. Therefore,

$$\begin{aligned} I_{\text{foot/cm}} &= (7.12 \times 10^{-4}\ \text{kg m}^2) - (0.365\ \text{kg})(0.032\ \text{m})^2 \\ &= 3.38 \times 10^{-4}\ \text{kg m}^2 \end{aligned}$$

$I_{\text{foot/cm}}$ represents the centroidal mass moment of inertia about the transverse principal axes of the foot (y' and z' in Figure 4.30). Consequently,

$$I_{y'y'} = 3.38 \times 10^{-4}\ \text{kg m}^2$$
$$I_{z'z'} = 3.38 \times 10^{-4}\ \text{kg m}^2$$

The foot is approximated as a cylinder with a length to radius ratio of 6. The ratio of transverse to longitudinal (x') mass moments of inertia can be shown to be approximately 6.5. Then the longitudinal mass moment of inertia (about x' in Figure 4.30) may be estimated as

$$I_{x'x'} = 5.20 \times 10^{-5}\ \text{kg m}^2$$

Having estimated the anthropomorphic values for the foot, the kinetic analysis may now begin. The unknown ankle reaction force, $\mathbf{F_A}$, is found by using Newton's Second Law, or $\sum \mathbf{F} = m\mathbf{a}$:

$$\mathbf{F_g} + \mathbf{F_A} - m_{\text{foot}}\, g\mathbf{k} = m_{\text{foot}}\, \mathbf{a_{foot}}$$

$$\begin{aligned} \mathbf{F_A} &= m_{\text{foot}}\, \mathbf{a_{foot}} - \mathbf{F}_{\text{g}} + m_{\text{foot}}\, g\mathbf{k} \\ &= (0.365\ \text{kg})[2.09\mathbf{i} - 0.357\mathbf{j} - 0.266\mathbf{k}]\ \text{m/s} \\ &\quad - (3.94\mathbf{i} - 15.21\mathbf{j} + 242.4\mathbf{k})\ \text{N} \\ &\quad + (0.365\ \text{kg})(9.81\ \text{m/s}^2)\mathbf{k} \\ &= -3.18\mathbf{i} + 15.08\mathbf{j} - 238.9\mathbf{k}\ \text{N} \end{aligned}$$

Euler's equations of motion (Eqs. (4.44)–(4.46)) are then applied to determine the unknown ankle moment reaction $\mathbf{M_A}$. Euler's equations are defined relative to the principal axes fixed to the segment—that is, x', y', and z' fixed to the foot. It is noted, however, that the data required for the solution presented previously—for example, $\boldsymbol{\omega}_{\mathbf{foot}}$ and $\boldsymbol{\alpha}_{\mathbf{foot}}$—are expressed relative to the laboratory coordinate system (x, y, z). Consequently, vectors required for the solution of Euler's equations must first be transformed into the foot coordinate system. In the preceding data set, the foot anatomical coordinate system was given as

$$\mathbf{e_{fax}} = 0.977\mathbf{i} - 0.0624\mathbf{j} - 0.202\mathbf{k}$$
$$\mathbf{e_{fay}} = 0.0815\mathbf{i} + 0.993\mathbf{j} + 0.0877\mathbf{k}$$
$$\mathbf{e_{faz}} = 0.195\mathbf{i} - 0.102\mathbf{j} + 0.975\mathbf{k}$$

where $\mathbf{e_{fax}}$, $\mathbf{e_{fay}}$, and $\mathbf{e_{faz}}$ correspond to x', y', and z', or $\mathbf{i}'$, $\mathbf{j}'$, and $\mathbf{k}'$. Recall from the discussion in Section 4.2.2 that coefficients in the expression for $\mathbf{e_{fax}}$ represent the cosines of the angles between x' and x, x' and y, and x' and z, respectively. Similarly, the coefficients in the expression for $\mathbf{e_{fay}}$ represent the cosines of the angles between y' and x, y' and y, and y' and z, and

the coefficients in the expression for $\mathbf{e_{faz}}$ represent the cosines of the angles between z' and x, z' and y, and z' and z. Consequently, these relationships can be transposed as

$$\begin{aligned}\mathbf{i} &= 0.977\mathbf{i}' + 0.0815\mathbf{j}' + 0.195\mathbf{k}'\\ \mathbf{j} &= -0.0624\mathbf{i}' + 0.993\mathbf{j}' - 0.102\mathbf{k}'\\ \mathbf{k} &= -0.202\mathbf{i}' + 0.0877\mathbf{j}' + 0.975\mathbf{k}'\end{aligned}$$

In this form, these relationships can be used to transform vectors expressed in terms of lab coordinates:

$$\mathbf{A} = A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k}$$

into foot coordinates:

$$\mathbf{A} = A_x\mathbf{i}' + A_y\mathbf{j}' + A_z\mathbf{k}'$$

To demonstrate this process, consider the foot angular velocity vector

$$\boldsymbol{\omega}_{\text{foot}} = 0.042\mathbf{i} + 2.22\mathbf{j} - 0.585\mathbf{k}\ \text{rad/s}$$

Substituting the relationships for the lab coordinate system in terms of the foot coordinate system, $\boldsymbol{\omega}_{\text{foot}}$ becomes

$$\begin{aligned}\boldsymbol{\omega}_{\text{foot}} &= 0.042(0.977\mathbf{i}' + 0.0815\mathbf{j}' + 0.195\mathbf{k}')\\ &\quad + 2.22(-0.0624\mathbf{i}' + 0.993\mathbf{j}' - 0.102\mathbf{k}')\\ &\quad - 0.585(-0.202\mathbf{i}' + 0.0877\mathbf{j}' + 0.975\mathbf{k}')\\ &= 0.0210\mathbf{i}' + 2.16\mathbf{j}' - 0.789\mathbf{k}'\ \text{rad/s}\end{aligned}$$

In a similar manner, the other vectors required for the computation are transformed into the foot coordinate system:

$$\begin{aligned}\mathbf{r}_1 &= -0.032\mathbf{i} + 0.002\mathbf{j} + 0.002\mathbf{k}\\ &= -0.032\mathbf{i}' - 0.004\mathbf{k}'\text{m}\\ \mathbf{r}_2 &= 0.033\mathbf{i} - 0.005\mathbf{j} - 0.054\mathbf{k}\\ &= 0.0435\mathbf{i}' - 0.007\mathbf{j}' - 0.0457\mathbf{k}'\ \text{m}\\ \mathbf{F_g} &= 3.94\mathbf{i} - 15.21\mathbf{j} + 242.36\mathbf{k}\\ &= -44.16\mathbf{i}' + 6.47\mathbf{j}' + 238.62\mathbf{k}'\ \text{N}\\ \mathbf{T_g} &= 0.995\mathbf{k}\\ &= -0.201\mathbf{i}' + 0.0873\mathbf{j}' + 0.970\mathbf{k}'\ \text{N m}\\ \mathbf{F_A} &= -3.18\mathbf{i} + 15.1\mathbf{j} - 239\mathbf{k}\\ &= 44.2\mathbf{i}' - 6.23\mathbf{j}' - 235\mathbf{k}'\text{N}\\ \boldsymbol{\omega}_{\mathbf{foot}} &= 0.0420\mathbf{i} + 2.22\mathbf{j} - 0.585\mathbf{k}\\ &= 0.021\mathbf{i}' + 2.16\mathbf{j}' - 0.789\mathbf{k}'\text{rad/s}\\ \boldsymbol{\alpha}_{\mathbf{foot}} &= -0.937\mathbf{i} + 8.85\mathbf{j} - 5.16\mathbf{k}\\ &= -0.425\mathbf{i}' + 8.26\mathbf{j}' - 6.116\mathbf{k}'\text{rad/s}^2\end{aligned}$$

Expanding Euler's equations of motion (Eqs. (4.44)–(4.46)),

$$\mathbf{M}_{\mathbf{Ax}'} + (\mathbf{r}_1 \times \mathbf{F}_\mathbf{A})_{x'} + (\mathbf{r}_2 \times \mathbf{F}_\mathbf{g})_{x'} + \mathbf{T}_{\mathbf{gx}'} = I_{x'x'}\alpha_{\mathbf{x}'} + (I_{z'z'} - I_{y'y'})\omega_{y'}\omega_{z'}$$
$$\mathbf{M}_{\mathbf{Ay}'} + (\mathbf{r}_1 \times \mathbf{F}_\mathbf{A})_{y'} + (\mathbf{r}_2 \times \mathbf{F}_\mathbf{g})_{y'} + \mathbf{T}_{\mathbf{gy}'} = I_{y'y'}\alpha_{\mathbf{y}'} + (I_{x'x'} - I_{z'z'})\omega_{z'}\omega_{x'}$$
$$\mathbf{M}_{\mathbf{Az}'} + (\mathbf{r}_1 \times \mathbf{F}_\mathbf{A})_{z'} + (\mathbf{r}_2 \times \mathbf{F}_\mathbf{g})_{z'} + \mathbf{T}_{\mathbf{gz}'} = I_{z'z'}\alpha_{\mathbf{z}'} + (I_{y'y'} - I_{x'x'})\omega_{x'}\omega_{y'}$$

where $(\mathbf{r}_1 \times \mathbf{F}_\mathbf{A})_{x'}$ represents the $\mathbf{x}'$ component of $\mathbf{r}_1 \times \mathbf{F}_\mathbf{A}$, $(\mathbf{r}_2 \times \mathbf{F}_\mathbf{g})_{x'}$ represents the $\mathbf{x}'$ component of $\mathbf{r}_2 \times \mathbf{F}_\mathbf{g}$, and so forth.

Substitution of the required values and arithmetic reduction yields

$$\mathbf{M}_{\mathbf{A}'} = 1.50\mathbf{i}' + 15.9\mathbf{j}' - 1.16\mathbf{k}' \text{ N m}$$

which can be transformed back into fixed lab coordinates,

$$\mathbf{M}_\mathbf{A} = 2.54\mathbf{i} + 15.9\mathbf{j} - 0.037\mathbf{k} \text{ N m}$$

By combining the ankle moment with the ankle angular velocity, the instantaneous ankle power may be computed as

$$\begin{aligned}\mathbf{M}_\mathbf{A} \cdot \boldsymbol{\omega}_\mathbf{ankle} &= (2.54\mathbf{i} + 15.9\mathbf{j} - 0.037\mathbf{k} \text{ N m}) \cdot (-0.000759\mathbf{i} + 1.47\mathbf{j} + 0.0106\mathbf{k} \text{ rad/s}) \\ &= 23.3 \text{ Watts}\end{aligned}$$

or

$$\begin{aligned}\mathbf{M}_{\mathbf{A}'} \cdot \boldsymbol{\omega}_{\mathbf{ankle}'} &= (1.50\mathbf{i}' + 15.9\mathbf{j}' - 1.16\mathbf{k}' \text{ N m}) \cdot (-0.0946\mathbf{i}' + 1.46\mathbf{j}' - 0.140\mathbf{k}' \text{ rad/s}) \\ &= 23.3 \text{ Watts}\end{aligned}$$

which is thought to represent a quantitative measure of the ankle's contribution to propulsion.

4.6.4 Clinical Gait Interpretation

The information and data provided for treatment decision making in clinical gait analysis include not only the quantitative variables described previously—3-D kinematics such as angular displacement of the torso, pelvis, hip, knee and ankle/foot, and 3-D kinetics, such as moments and power of the hip, knee, and ankle—but the following as well:

- Clinical examination measures
- Biplanar video recordings of the patient walking
- Stride and temporal gait data such as step length and walking speed
- Electromyographic (EMG) recordings of selected lower extremity muscles

Generally, the interpretation of gait data involves the identification of abnormalities, the determination of the causes of the apparent deviations, and the recommendation of treatment alternatives. As each additional piece of data is incorporated, a coherent picture of the patient's walking ability is developed by correlating corroborating data sets and resolving apparent contradictions in the information. Experience allows the team to distinguish a gait anomaly that presents the difficulty for the patient from a gait compensatory mechanism that aids the patient in circumventing the gait impediment to some degree.

To illustrate aspects of this process, consider the data presented in Figures 4.31–4.33, which were measured from a 9-year-old girl with cerebral palsy spastic diplegia. Cerebral

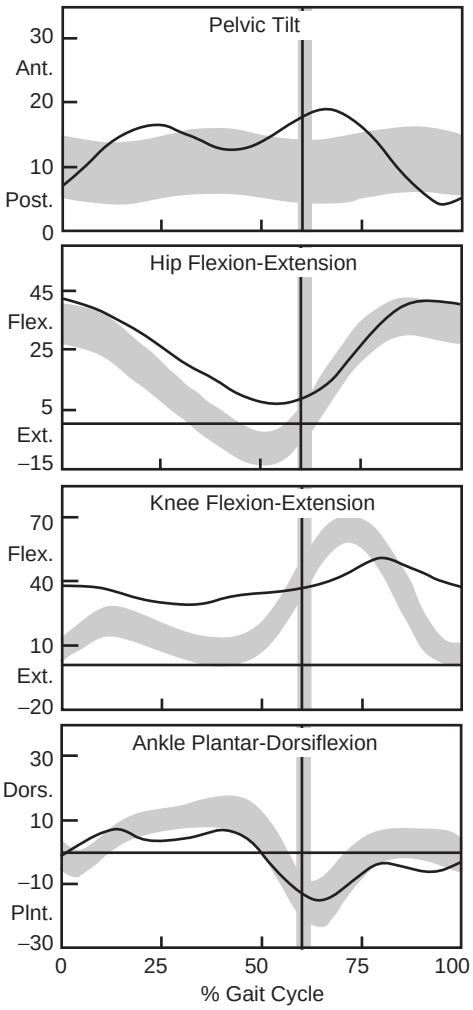


FIGURE 4.31 Sagittal plane kinematic data for the left side of a 9-year-old patient with cerebral palsy spastic diplegia (solid curves). Shaded bands indicate +/− one standard deviation about the performance of children with normal ambulation. Stance phase is 0–60 percent of the gait cycle, and swing phase is 60–100 percent, as indicated by the vertical solid lines.

palsy is a nonprogressive neuromuscular disorder that is caused by an injury to the brain during or shortly after birth. The neural motor cortex is most often affected. In the ambulatory patient, this results in reduced control of the muscles required for balance and locomotion, causing overactivity, inappropriately timed activity, and muscle spasticity. Treatment options include physical therapy, bracing (orthoses), spasmolytic medications such as botulinum toxin and Baclofen, and orthopedic surgery and neurosurgery.

The sagittal plane kinematics for the left side of this patient (Figure 4.31) indicate significant involvement of the hip and knee. Her knee is effectively "locked" in an excessively

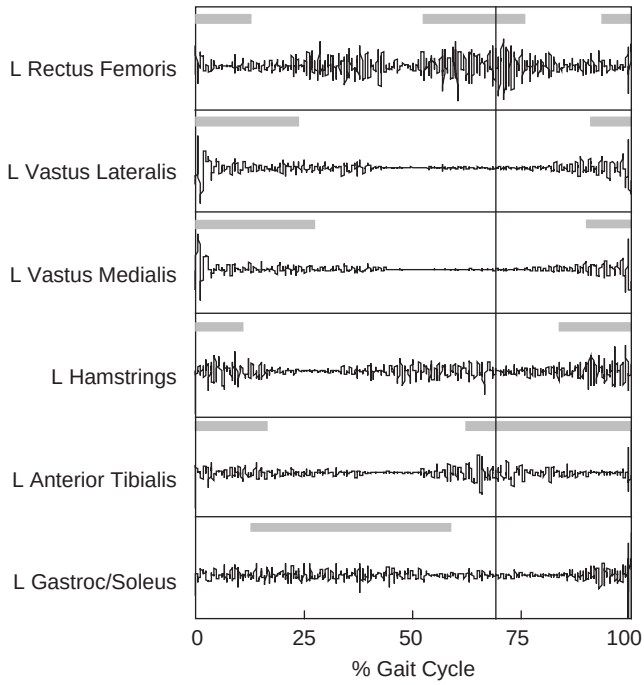


FIGURE 4.32 Electromyogram (EMG) data for the same cerebral palsy patient as in Figure 4.31. Plotted are EMG activity signals for each of six left lower extremity muscles, each plotted as functions of percent of gait cycle. Gray bars represent mean normal muscle activation timing.

flexed position throughout stance phase (0–60 percent of the gait cycle) when her foot is contacting the floor. Knee motion in swing phase (60–100 percent) is also limited, with the magnitude and timing of peak knee flexion in swing reduced and delayed. The range of motion of her hip during gait is less than normal, failing to reach full extension at the end of stance phase. The motion of her pelvis is significantly greater than normal, tilting anteriorly in early stance coincident with extension of the hip, and tilting posteriorly in swing coincident with flexion of the hip.

The deviations noted in these data illustrate neuromuscular problems commonly seen in this patient population. Inappropriate hamstring tightness, observed during the clinical examination, and inappropriate muscle activity during stance, seen in Figure 4.32, prevent the knee from properly extending. This flexed knee position also impedes normal extension of the hip in stance due to hip extensor weakness, also observed during the clinical examination. Hip extension is required in stance to allow the thigh to rotate under the advancing pelvis and upper body. To compensate for her reduced ability to extend the hip, she rotates her pelvis anteriorly in early stance to help move the thigh through its arc of motion. The biphasic pattern of the pelvic curve indicates that this is a bilateral issue to some degree.

The limited knee flexion in swing combines with the plantar flexed ankle position to result in foot clearance problems during swing phase. The inappropriate activity of the rectus femoris muscle (Figure 4.32) in midswing suggests that spasticity of that muscle, a

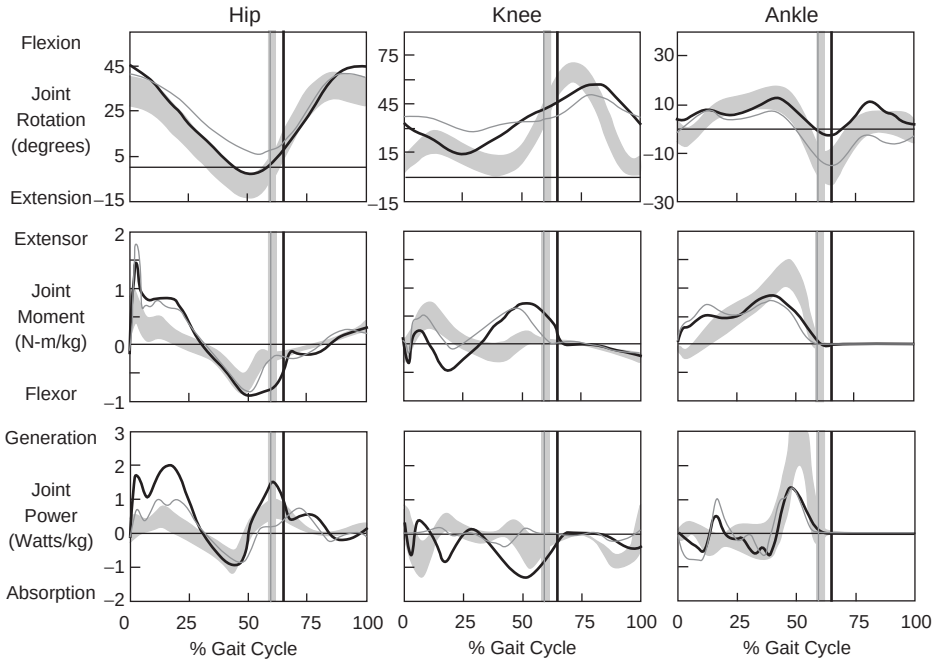


FIGURE 4.33 Sagittal joint kinetic data for the same cerebral palsy patient of Figure 4.31. Joint rotation, joint moment, and joint power are plotted as functions of percent of gait cycle for the hip, knee, and ankle. Dark and light solid curves denote right and left sides, respectively. Bands indicate +/− one standard deviation of the normal population.

knee extensor (also observed during clinical examination), impedes knee flexion. Moreover, the inappropriate activity of the ankle plantar flexor, primarily the gastrocnemius muscle, in late swing suggests that it is overpowering the pretibial muscles, primarily the anterior tibialis muscle, resulting in plantar flexion of the ankle or "foot drop."

The sagittal joint kinetics for this patient (Figure 4.33) demonstrate asymmetrical involvement of the right and left sides. Of special note is that her right knee and hip are compensating for some of the dysfunction observed on the left side. Specifically, the progressively increasing right knee flexion beginning at midstance (first row, center) and continuing into swing aids her contralateral limb in forward advancement during swing—that is, her pelvis can rock posteriorly along with a flexing hip to advance the thigh. One potentially adverse consequence of this adaptation is the elevated knee extensor moment in late stance that increases patella-femoral loading with indeterminate effects over time. The asymmetrical power production at the hip also illustrates clearly that the right lower extremity, in particular the muscles that cross the hip, provides the propulsion for gait with significant power generation early in stance to pull the body forward and elevate its center of gravity. Moreover, the impressive hip power generation, both with respect to magnitude and timing, at toe-off accelerates the stance limb into swing and facilitates knee flexion in spite of the elevated knee extensor moment magnitude. This is important to appreciate given the bilateral spastic response of the plantar flexor muscles, as evidenced by the premature ankle power

generation and the presentation of a spastic stretch reflex in the clinical examination. This girl uses her hip musculature, right more than left, to a much greater degree than her ankle plantar flexors to propel herself forward during gait.

This cursory case examination illustrates the process whereby differences from normal gait are recognized and the associated biomechanical etiology is explored. Some of the effects on gait of neuromuscular pathology in the sagittal plane have been considered in this discussion. Clinical gait analysis can also document and elucidate gait abnormalities associated with static bony rotational deformities. It also is useful in areas of clinical research by documenting treatment efficacy associated with bracing, surgery, and so forth. It should be noted, however, that although engineers and applied physicists have been involved in this work for well over a hundred years, there remains significant opportunity for improvement in the biomechanical protocols and analytical tools used in clinical gait analysis; in other words, there remains much to learn.

4.7 CARDIOVASCULAR DYNAMICS

One major organ system benefiting from the application of mechanics principles is cardiovascular system dynamics, or *hemodynamics*, the study of the motion of blood. From a functional point of view, the cardiovascular system is driven by a complex pump, the heart, that generates pressure resulting in the flow of a complex fluid, blood, through a complex network of complex pipes, the blood vessels. Cardiovascular dynamics focuses on the measurement and analysis of blood pressure, volume, and flow within the cardiovascular system. The complexity of this elegant system is such that mechanical models, typically formulated as mathematical equations, are relied on to understand and integrate experimental data, to isolate and identify physiological mechanisms, and to lead ultimately to new clinical measures of heart performance and health and guide clinical therapies.

As described in Chapter 3, the heart is a four-chambered pump connected to two main collections of blood vessels: the systemic and pulmonary circulations. This pump is electrically triggered and under neural and hormonal control. One-way valves control blood flow. Total human blood volume is approximately 5.2 liters. The left ventricle, the strongest chamber, pumps 5 liters per minute at rest, almost the body's entire blood volume. With each heartbeat, the left ventricle pumps 70 ml, with an average of 72 beats per minute. During exercise, left ventricular output may increase sixfold, and heart rate more than doubles. The total combined length of the circulatory system vessels is estimated at 100,000 km, a distance two and one half times around the earth. The left ventricle generates approximately 1.7 watts of mechanical power at rest, increasing threefold during heavy exercise. One curious constant is the total number of heartbeats in a lifetime, around one billion in mammals [31]. Larger animals have slower heart rates and live longer lives, and vice versa for small animals.

4.7.1 Blood Rheology

Blood is composed of fluid, called plasma, and suspended cells, including erythrocytes (red blood cells), leukocytes (white cells), and platelets. From a mechanical point of view, a fluid is distinguished from a solid as follows. Figure 4.34 shows a two-dimensional block of solid material (left panel) subjected to two opposite, parallel, transverse external forces, depicted by the solid arrows at the top and bottom surfaces. This applied shear force is resisted by

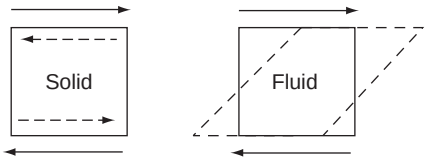


FIGURE 4.34 (Left) A solid material resists applied external shear stress (solid vectors) via internally generated reaction shear stress (dashed vectors). (Right) A fluid subjected to applied shear stress is unable to resist and instead flows (dashed lines).

the solid via internally generated reaction forces, depicted by the dashed arrows. When applied to a fluid (right panel), the fluid cannot resist the applied shear but rather flows.

The applied shear forces lead to shear stresses, force per area, and the measure of flow can be quantified by the resulting shear strain rate. In essence, the harder one pushes on a fluid (higher shear stress), the faster the fluid flows (higher shear strain rate). The relationship between shear stress, τ, and shear strain rate, $\dot{\gamma}$, is the fluid's viscosity, μ. Viscosity is sometimes written as η in biomedical applications. As shown in Figure 4.35, many fluids, including water, are characterized by a constant, linear viscosity and are called Newtonian. Others possess nonlinear shear stress-strain rate relations, and are non-Newtonian fluids. For example, fluids that behave more viscously as shear strain rate increases are called dilatant, or shear thickening. One example of dilatant behavior is Dow Corning 3179 dilatant compound, a silicone polymer commonly known as "Silly Putty." When pulled slowly, this fluid stretches (plastic deformation); when pulled quickly, it behaves as a solid and fractures. Fluids that appear less viscous with higher shear strain rates are called pseudoplastic, or shear thinning. For example, no-drip latex paint flows when applied with a brush or roller that provides shear stress, but it does not flow after application.

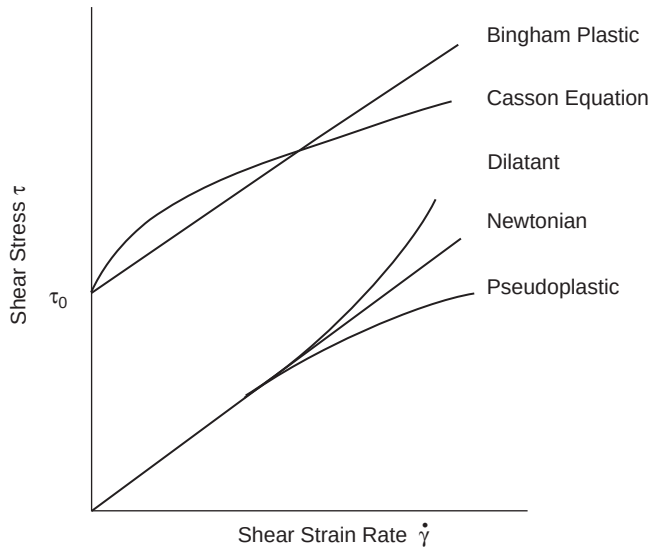


FIGURE 4.35 Newtonian fluids exhibit a constant viscosity$\mu = \tau/\dot{\gamma}$, arising from the linear relation between shear stress and shear strain rate. Non-Newtonian fluids are nonlinear. Blood is often characterized with a Casson equation but under many conditions may be described as Newtonian.

Biological fluids are typically non-Newtonian. Blood plasma is Newtonian and is very similar in physical properties to water. Whole blood behaves as a Bingham plastic, whereby a nonzero shear stress, denoted yield stress τ_0, is required before this fluid begins to flow. Blood is often characterized by a power law function, of the form

$$\tau = k\dot{\gamma}^n \tag{4.57}$$

where k and n are constants derived from a straight-line fit of $\ln \tau$ plotted as a function of $\ln \dot{\gamma}$, since

$$\ln \tau = \ln k + n \ln \dot{\gamma}$$

Another common description of blood's viscosity is the Casson equation:

$$\tau^{\frac{1}{2}} = \tau_0^{\frac{1}{2}} + k\dot{\gamma}^{\frac{1}{2}} \tag{4.58}$$

From a Casson plot, the yield stress τ_0 can be measured. Rheology, the study of deformation and flow of fluids, focuses on these often complex viscous properties of fluids. Textbooks with rheological data for biofluids include *Basic Transport Phenomena in Biomedical Engineering* by Fournier [9] and *Biofluid Mechanics* by Chandran and colleagues [4].

EXAMPLE PROBLEM 4.13

The following rheological data were measured on a blood sample:

Shear Strain Rate [s^{-1}]	Shear Stress [dyne/cm^2]
1.5	12.5
2.0	16.0
3.2	25.2
6.5	40.0
11.5	62.0
16.0	80.5
25.0	120
50.0	240
100	475

Fit the data to a power law function using a MATLAB m-file.

Solution

```
% Power Law Fit of Blood Data
%
% Store shear strain rate and stress data in arrays
alpha = [1.5,2,3.2,6.5,11.5,16,25,50,100];
T = [12.5,16,25.2,40,62,80.5,120,240,475];
% Take natural logs of both
x = log(alpha);
y = log(T);
```

```
% Use MATLAB's polyfit function to do linear curve fit
coeff = polyfit(x,y,1)
% Write curve fit coefficients as a new x-y function for plotting
x1=[0;0.01;5]
y1=polyval(coeff,x1)
% Plot the original data as 'o' points
plot(x,y,'o')
hold on
% Overlay a plot of the curve-fit line
plot(x1,y1)
grid on
title('Power Law Function')
xlabel('ln Strain Rate  [ln(1/s)]')
ylabel('ln Shear Stress  [ln dyne/cm2]')
%
```

The resulting plot appears in Figure 4.36.

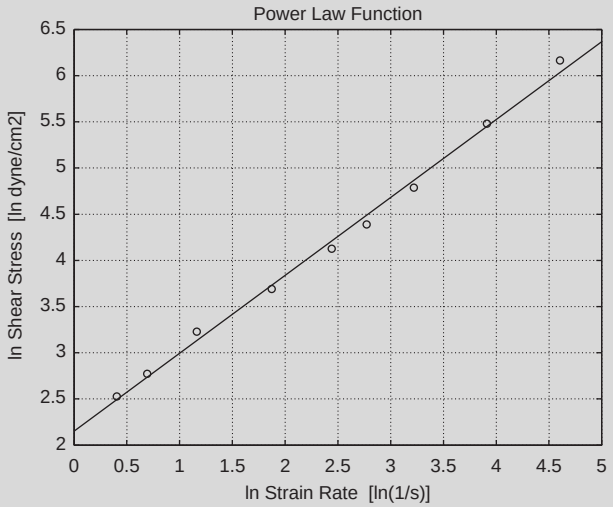


FIGURE 4.36 The power law curve-fit using MATLAB of the rheological blood data in Example Problem 4.13.

When subjected to very low shear rates, blood's apparent viscosity is higher than expected. This is due to the aggregation of red blood cells, called rouleaux. Such low shear rates are lower than those typically occurring in major blood vessels or in medical devices. In small tubes of less than 1 mm diameter, blood's apparent viscosity at high shear rates is smaller than in larger tubes. This Fahraeus-Lindquist effect arises from plasma–red blood cell dynamics. Beyond these two special cases, blood behaves as a Newtonian fluid and is widely accepted as such. We shall see that the assumption of Newtonian fluid greatly simplifies mechanical description of the circulation.

4.7.2 Arterial Vessels

Mechanical description of blood vessels has a long and somewhat complicated history. Much of the advanced mathematics and applied mechanics associated with this work is beyond the scope of this textbook. This section will therefore give an overview of some of the main developments and will present a simplified, reduced arterial system model for use in the following subsection. The reader is referred to the following textbooks for more in-depth coverage: *Circulatory System Dynamics* [21] by Noordergraaf, *Hemodynamics* [15] by Milnor, and *Biofluid Mechanics* [4] by Chandran and colleagues, and for basic fluid mechanics, *Fluid Mechanics* [30] by White.

Study of the mechanical properties of the heart as a pump requires the computation of pressures and flows arising from forces and motion of the underlying heart muscle. Consequently, general equations of motion in the cardiovascular system typically arise from the conservation of linear momentum. The Reynold's transport theorem, a conservation equation from fluid mechanics, applied to linear momentum yields the following general equation of motion for any fluid:

$$\rho\mathbf{g} - \nabla p + \nabla \cdot \boldsymbol{\tau}_{ij} = \rho \frac{d\mathbf{V}}{dt} \tag{4.59}$$

where ρ is fluid density (mass/volume), p is pressure, $\boldsymbol{\tau}_{ij}$ are viscous forces, and $\mathbf{V}$ is velocity. ∇ is the differential operator

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}$$

The general velocity $\mathbf{V}$ is a vector function of position and time and is written

$$\mathbf{V}(x,y,z,t) = u(x,y,z,t)\mathbf{i} + v(x,y,z,t)\mathbf{j} + w(x,y,z,t)\mathbf{k}$$

where u, v, and w are the local velocities in the x, y, and z directions, respectively.

Equation (4.59) comprises four terms: *gravitational*, *pressure*, and *viscous forces*, plus a time-varying term. Note that this is a *vector* equation and so can be expanded in x, y, and z components as the set of three equations:

$$\rho g_x - \frac{\partial p}{\partial x} + \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} = \rho\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z}\right) \tag{4.60}$$

$$\rho g_y - \frac{\partial p}{\partial y} + \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{zy}}{\partial z} = \rho\left(\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z}\right) \tag{4.61}$$

$$\rho g_z - \frac{\partial p}{\partial z} + \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} = \rho\left(\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z}\right) \tag{4.62}$$

This set of nonlinear, partial differential equations is general but not solvable; solution requires making simplifying assumptions. For example, if the fluid's viscous forces are neglected, Eq. (4.59) reduces to Euler's equation for inviscid flow. The latter, when integrated along a streamline, yields the famous Bernoulli equation relating pressure and flow. In application, Bernoulli's inviscid, and consequently frictionless, origin is sometimes forgotten.

If flow is steady, the right-hand term of Eq. (4.59) goes to zero. For incompressible fluids, including liquids, density ρ is constant, which greatly simplifies integration of the

gravitational and time-varying terms that contain ρ. Similarly, for Newtonian fluids, viscosity μ is constant. In summary, although we can write perfectly general equations of motion, the difficulty of solving these equations requires making reasonable simplifying assumptions.

Two reasonable assumptions for blood flow in major vessels are those of Newtonian and incompressible behavior. These assumptions reduce Eq. (4.59) to the Navier-Stokes equations:

$$\rho g_x - \frac{\partial p}{\partial x} + \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = \rho \frac{du}{dt} \tag{4.63}$$

$$\rho g_y - \frac{\partial p}{\partial y} + \mu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}\right) = \rho \frac{dv}{dt} \tag{4.64}$$

$$\rho g_z - \frac{\partial p}{\partial z} + \mu\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2}\right) = \rho \frac{dw}{dt} \tag{4.65}$$

Blood vessels are more easily described using a cylindrical coordinate system rather than a rectangular one. Hence, the coordinates x, y, and z may be transformed to radius r, angle θ, and longitudinal distance x. If we assume irrotational flow, $\theta = 0$ and two Navier-Stokes equations suffice:

$$-\frac{dP}{dx} = \rho\left(\frac{dw}{dt} + u\frac{dw}{dr} + w\frac{dw}{dx}\right) - \mu\left(\frac{d^2 w}{dr^2} + \frac{1}{r}\frac{dw}{dr} + \frac{d^2 w}{dx^2}\right) \tag{4.66}$$

$$-\frac{dP}{dr} = \rho\left(\frac{du}{dt} + u\frac{du}{dr} + w\frac{du}{dx}\right) - \mu\left(\frac{d^2 u}{dr^2} + \frac{1}{r}\frac{du}{dr} + \frac{d^2 u}{dx^2} - \frac{u}{r^2}\right) \tag{4.67}$$

where w is longitudinal velocity dx/dt, and u is radial velocity dr/dt. Most arterial models also use the continuity equation, arising from the conservation of mass:

$$\frac{du}{dr} + \frac{u}{r} + \frac{dw}{dx} = 0 \tag{4.68}$$

In essence, the net rate of mass storage in a system is equal to the net rate of mass influx minus the net rate of mass efflux.

Noordergraaf and his colleagues [20] rewrote the Navier-Stokes Eq. (4.66) as

$$-\frac{dP}{dx} = RQ + L\frac{dQ}{dt} \tag{4.69}$$

where P is pressure, Q is the volume rate of flow, R is an equivalent hydraulic resistance, and L is fluid inertance. The Navier-Stokes equations describe fluid mechanics within the blood vessels. Since arterial walls are elastic, equations of motion for the arterial wall are also required. The latter have evolved from linear elastic and linear viscoelastic, to complex viscoelastic (see [21]). The most general mechanical description of linear anisotropic arterial wall material requires 21 parameters (see [10]), most of which have never been measured. Noordergraaf and colleagues divided the arterial system into short segments and combined the fluid mechanical equation (Eq. (4.69)) with the continuity equation for each vessel segment. The arterial wall elasticity leads to a time-varying amount of blood stored in the

vessel as it bulges with each heartbeat. For a segment of artery, the continuity equation becomes

$$-\frac{dQ}{dx} = GP + C\frac{dP}{dt} \tag{4.70}$$

where G is leakage through the blood vessel wall. This pair of hydraulic equations—Eqs. (4.69) and (4.70)—was used to describe each of 125 segments of the arterial system and was the first model sufficiently detailed to explain arterial pressure and flow wave reflection. Arterial branching leads to reflected pressure and flow waves that interact in this pulsatile system. Physical R-L-C circuits were constructed and built into large transmission line networks with measured voltages and currents corresponding to hydraulic pressures and flows, respectively. If distributed arterial properties such as pulse wave reflection are not of interest, the arterial system load seen by the heart can be much reduced, as an electrical network may be reduced to an equivalent circuit.

The most widely used arterial load is the three-element model shown in Figure 4.37. The model appears as an electrical circuit due to its origin prior to the advent of the digital computer. Z_0 is the characteristic impedance of the aorta, in essence the aorta's flow resistance. C_s is transverse arterial compliance, the inverse of elastance, and describes stretch of the arterial system in the radial direction. R_s is the peripheral resistance, describing the systemic arteries' flow resistance downstream of the aorta. This simple network may be used to represent the systemic arterial load seen by the left ventricle. The following ordinary differential equation relates pressure at the left-hand side, $p(t)$, to flow, $Q(t)$:

$$C_s\frac{dp}{dt} + \frac{1}{R_s}p(t) = Q(t)\left(1 + \frac{Z_0}{R_s}\right) + Z_0C_s\frac{dQ}{dt} \tag{4.71}$$

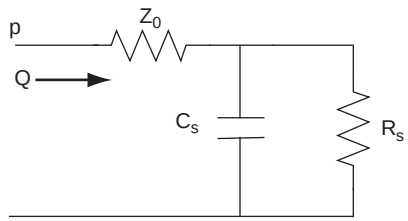


FIGURE 4.37 Equivalent systemic arterial load. Circuit elements are described in the text.

EXAMPLE PROBLEM 4.14

Using basic circuit theory, derive the differential equation (Eq. (4.71)) from Figure 4.37.

Solution

Define node 1 as shown in Figure 4.38. By Kirchoff's current law, the flow Q going into node 1 is equal to the sum of the flows Q_1 and Q_2 coming out of the node:

$$Q = Q_1 + Q_2$$

We can write Q_1 and Q_2 as

$$Q_1 = C_s \frac{dp_1}{dt}$$

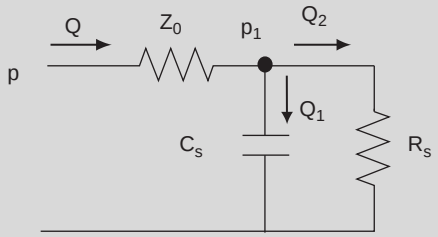


FIGURE 4.38 Nodal analysis of three-element arterial load.

$$Q_2 = \frac{p_1}{R_s}$$

so

$$Q = C_s \frac{dp_1}{dt} + \frac{p_1}{R_s}$$

From Ohm's law,

$$p - p_1 = QZ_0$$

Solving the last expression for p_1 and substituting back into the flow expression:

$$Q = C_s \frac{d}{dt}[p - QZ_0] + \frac{1}{R_s}[p - QZ_0]$$

$$= C_s \frac{dp}{dt} - Z_0 C_s \frac{dQ}{dt} + \frac{1}{R_s} p - \frac{Z_0}{R_s} Q$$

Grouping terms for Q on the left and p on the right gives Eq. (4.71).

4.7.3 Heart Mechanics

Mechanical performance of the heart, more specifically the left ventricle, is typically characterized by estimates of ventricular elastance. The heart is an elastic bag that stiffens and relaxes with each heartbeat. Elastance is a measure of stiffness, classically defined as the differential relation between pressure and volume:

$$E_v = \frac{dp_v}{dV_v} \tag{4.72}$$

Here, p_v and V_v denote ventricular pressure and volume, respectively. For any instant in time, ventricular elastance E_v is the differential change in pressure with respect to volume. Mathematically, this relation is clear. Measurement of E_v is much less clear.

In the 1970s Kennish and colleagues tried to estimate the differential relation of Eq. (4.72) using the ratio of finite changes in ventricular pressure and volume:

$$E_v = \frac{\Delta p_v}{\Delta V_v} \tag{4.73}$$

This approach leads to physically impossible results. For example, before the aortic valve opens, the left ventricle is generating increasing pressure while there is not yet any change in volume. The ratio in Eq. (4.73) gives an infinite elastance when the denominator is zero. Suga and Sagawa [28] used the ratio of pressure to volume itself, rather than differential or discrete changes, to estimate elastance:

$$E_v(t) = \frac{p_v(t)}{V_v(t) - V_d} \tag{4.74}$$

In this equation, V_d is a dead volume that remains constant. All the other terms are allowed to be varying with time. Ventricular elastance measured in this way leads to elastance curves as depicted in Figure 4.39. These curves show wide variation, as suggested by the large error bars. The distinctive asymmetric shape leads to a major contradiction. A simple experiment involves clamping the aorta, thereby preventing the left ventricle from ejecting blood, denoted an *isovolumic* beat. Equation (4.74) shows that under isovolumic conditions (V_v is constant) ventricular pressure p_v must have the same shape as elastance $E_v(t)$. However, experiments show that isovolumic pressure curves are symmetric, unlike Figure 4.39. A further complication is the requirement of ejecting beats for measuring $E_v(t)$, which requires not only the heart (a ventricle) but also a circulation (blood vessels). Hence, time-varying elastance curves such as Figure 4.39 are measures of both a particular heart, the source, combined with a particular circulation, its load. Experiments show that elastance curves measured in this way are subject to vascular changes, as well as the desired ventricular properties. As such, this approach cannot uniquely separate out ventricular from vascular properties. Consequently, a new measure of the heart's mechanical properties is required.

The problems just described—inconsistent isovolumic and ejecting behavior and combined heart-blood vessel properties—led to the development of a new mechanical description of the

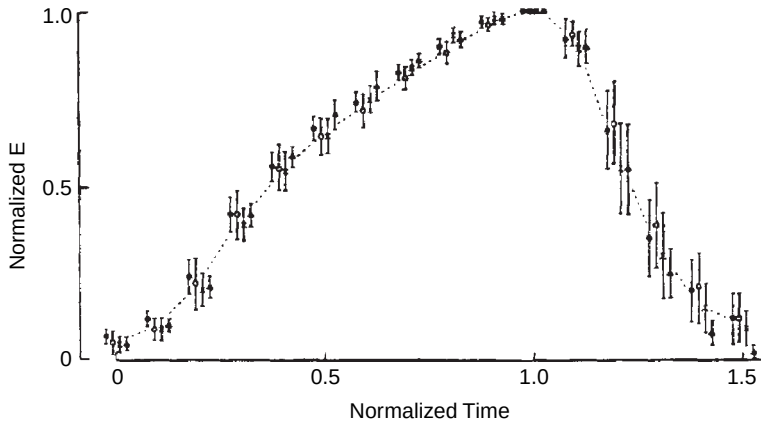


FIGURE 4.39 Time-varying ventricular elastance curves measured using the definition in Eq. (4.74). Measured elastance curves exhibit distinctive asymmetry. *Adapted from [28].*

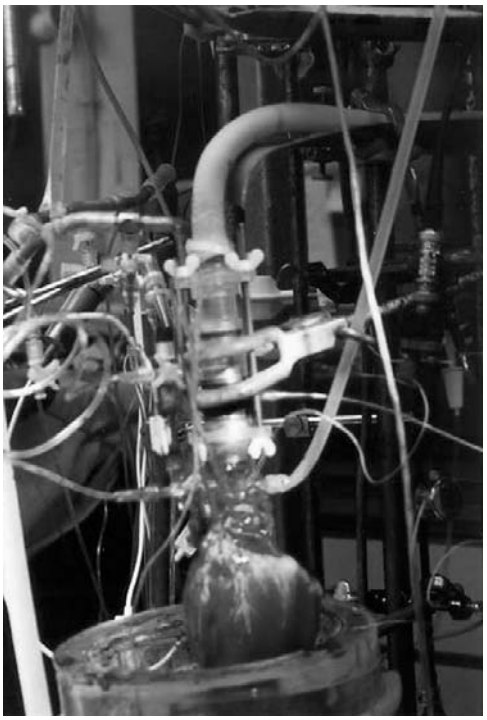


FIGURE 4.40 Isolated canine left ventricle used to develop a new biomechanical model of the heart. *Photo courtesy of Dr. Jan Mulier, Leuven, Belgium.*

left ventricle [17, 24, 23]. This model should be simple and versatile and should have direct physiological significance, in contrast with simulations, which merely mimic physiological behavior. The model was developed using isolated canine heart experiments, as depicted in Figure 4.40. The left ventricle was filled with an initial volume of blood, subjected to different loading conditions, stimulated, and allowed to beat. Ventricular pressure, and in some experiments ventricular outflow, was then measured and recorded.

Experiments began with measurement of isovolumic ventricular pressure. For each experiment the isolated left ventricle was filled with an initial end-diastolic volume and the aorta was clamped to prevent outflow of blood. The ventricle was stimulated, and generated ventricular pressure was measured and recorded. The ventricle was then filled to a new end-diastolic volume and the experiment was repeated. As in the famous experiments of Otto Frank (c. 1895), isovolumic pressure is directly related to filling. Figure 4.41 shows a set of isovolumic pressure curves measured on a normal canine left ventricle.

These isovolumic pressure curves were then described by the following equation. Ventricular pressure p_v is a function of time t and ventricular volume V_v according to

$$p_v = a(V_v - b)^2 + (cV_v - d)\left[\frac{(1 - e^{-(\frac{t}{\tau_c})^\alpha})e^{-(\frac{t-t_b}{\tau_r})^\alpha}}{(1 - e^{-(\frac{t_p}{\tau_c})^\alpha})e^{-(\frac{t_p-t_b}{\tau_r})^\alpha}}\right] \tag{4.75}$$

or written more compactly,

$$p_v(t, V_v) = a(V_v - b)^2 + (cV_v - d)f(t) \tag{4.76}$$

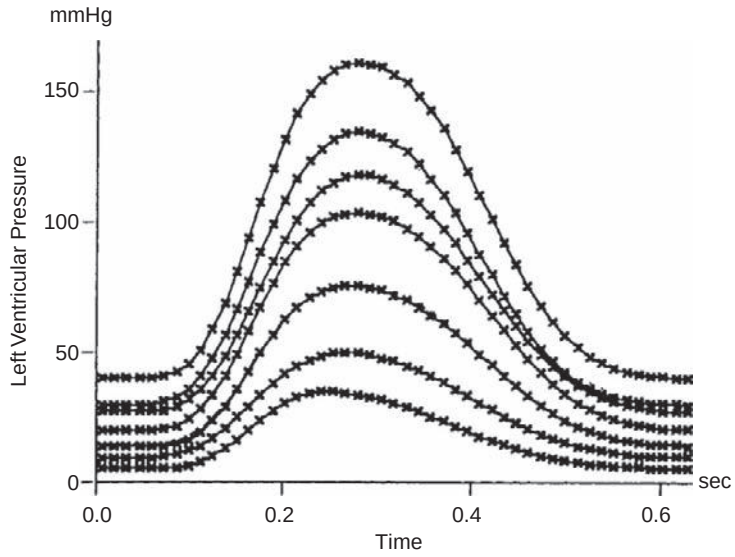


FIGURE 4.41 Isovolumic ventricular pressure curves. For each curve, the left ventricle is filled with a fixed initial volume, the heart is stimulated, and generated ventricular pressure is measured with respect to time. Other curves arise from different fixed initial volumes.

where $f(t)$ is the activation function in square brackets in Eq. (4.75). The constants a, b, c, d, t_p, τ_c, τ_r, and α were derived from the isolated canine ventricle experiments. Physiologically, Eq. (4.76) says that the ventricle is a time- and volume-dependent pressure generator. The term to the left of the plus sign, including constants a and b, describes the ventricle's passive elastic properties. The term to the right, including c and d, describes its active elastic properties, arising from the active generation of force in the underlying heart muscle. Representative model quantities measured from canine experiments are given in Table 4.3. This model was adapted to describe the human left ventricle using quantities in the right-hand column [24, 25].

TABLE 4.3 Ventricle Model Quantities Measured from Animal Experiments and Adapted for the Human Analytical Model

Quantity	Dog (Measured)	Human (Adapted)
a	0.003 [mmHg/ml^2]	0.0007
b	1.0 [ml]	20.0
c	3.0 [mmHg/ml]	2.5
d	20.0 [mmHg]	80.0
τ_c	0.164 [s]	0.264
t_p	0.271 [s]	0.371
τ_r	0.199 [s]	0.299
t_b	0.233 [s]	0.258
α	2.88	2.88

EXAMPLE PROBLEM 4.15

Solve Eq. (4.75) and plot ventricular pressure $p_v(t)$ for one human heartbeat. Use initial ventricular volume of 150 ml and the parameter values in Table 4.3.

Solution

The following MATLAB m-file will perform the required computation and plot the results, shown in Figure 4.42.

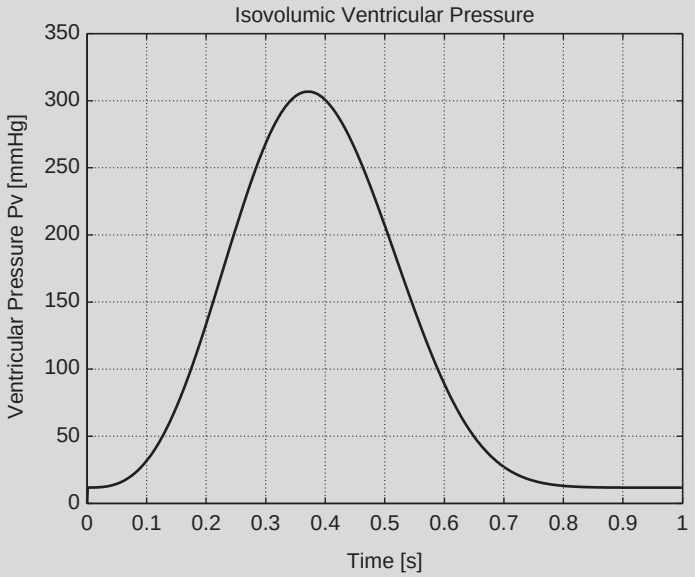


FIGURE 4.42 Isovolumic ventricular pressure computed for a human heartbeat.

```
% ventricle.m
%
% MATLAB m-file to compute isovolumic pressure using ventricle model
%
% Initial conditions:
%
delt = 0.001; % The iteration time step delta t
a = 7e-4;
b = 20.;
c = 2.5;
d = 80.;
tc = 0.264;
tp = 0.371;
tr = 0.299;
tb = 0.258;
alpha = 2.88;
```

Continued

```
Vv0 = 150; % Initial (end-diastolic) ventricular volume
%
% Compute intermediate term denom
% to simplify computations:
%
   denom = ((1.-exp(-(tp/tc)^alpha))*exp(-((tp-tb)/tr)^alpha));
%
% Compute for initial time t=0 (MATLAB does not allow 0 index)
%
   t(1) = 0.;
   Vv(1) = Vv0;
   edp = a*((Vv0 -b))^2;
   pdp = c*Vv0 - d;
   pp = pdp/denom;
   t1 = 0.; % Time step for first exponential
   t2 = 0.; % Time step for second exponential
   e1 = exp(-(t1/tc)^alpha);
   e2 = exp(-(t2/tr)^alpha);
   pv0 = edp + pp*((1.-e1)*e2);
%
% Main computation loop:
%
   for j=2:1000
   t(j) = t(j-1) + delt;
   Vv(j) = Vv(j-1);
%
   edp = a*((Vv(j) -b))^2;
   pdp = c*Vv(j) - d;
   pp = pdp/denom;
   t1 = t(j);
% Second exponential begins at t > tb
   t2 = t(j) - tb;
   if (t2 < 0.) ;
   t2 = 0.;
   end
   e1 = exp(-(t1/tc)^alpha);
   e2 = exp(-(t2/tr)^alpha);
   pv(j) = edp + pp*((1.-e1)*e2);
   end
%
plot(t,pv)
grid on
title('Isovolumic Ventricular Pressure')
xlabel('Time [s]')
ylabel('Ventricular Pressure Pv [mmHg]')
```

4.7.4 Cardiovascular Mechanics

This concise generalized pressure model of the left ventricle (Eq. (4.75)) was coupled to the reduced arterial load model of Figure 4.37 and allowed to eject blood. Model parameter values for a normal arterial load are given in Table 4.4. Figure 4.43 shows results for a normal canine left ventricle ejecting into a normal arterial system. The solid curves (left ordinate) describe ventricular pressure p_v and root aortic pressure as functions of time. Clinically, arterial pressure is reported as two numbers—for example, 110/60. This corresponds to the maximum and minimum root arterial pulse pressures—in this case, about 120/65 mmHg. The dashed curve (right ordinate) shows ventricular outflow. The ventricle was filled with an end-diastolic volume of 45 ml, and it ejected 30 ml (stroke volume), giving an ejection fraction of 66 percent, which is about normal for this size animal.

The same ventricle may be coupled to a pathological arterial system—for example, one with doubled peripheral resistance R_s. This change is equivalent to narrowed blood vessels.

TABLE 4.4 Representative Systemic Arterial Model Element Values

Element	Symbol	Control Value
Characteristic aorta impedance	Z_0	0.1 mmHg-s/ml
Systemic arterial compliance	C_s	1.5 ml/mmHg
Peripheral arterial resistance	R_s	1.0 mmHg-s/ml

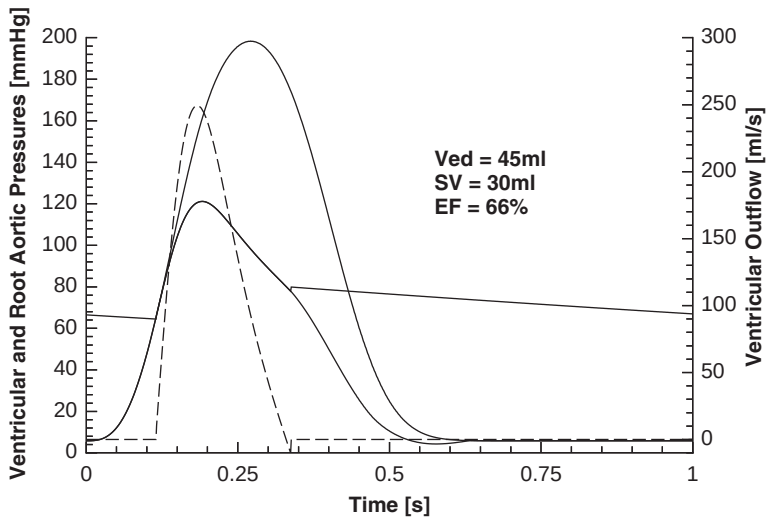


FIGURE 4.43 Ventricular and root aortic pressures (solid curves, left ordinate) and ventricular outflow (dashed curve, right ordinate) computed using the model of Eq. (4.75) for a normal canine left ventricle pumping into a normal arterial circulation. The topmost solid curve corresponds to a clamped aorta (isovolumic). The ventricle has initial volume of 45 ml and pumps out 30 ml, for an ejection fraction of 66 percent, which is about normal.

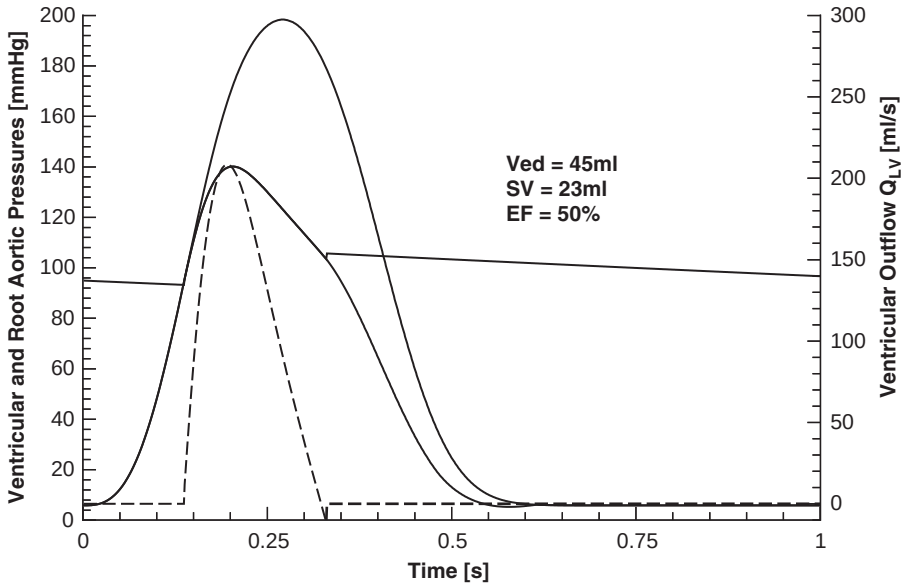


FIGURE 4.44 The same normal canine ventricle of Figure 4.43 now pumping into an arterial system with doubled peripheral (flow) resistance. As expected, increased resistance, corresponding to narrowed vessels, leads to increased arterial pulse pressure. Stroke volume is reduced from 66 to to 50 percent.

As expected, increased peripheral resistance raises arterial blood pressure to 140/95 mmHg and impedes the ventricle's ability to eject blood (Figure 4.44). The ejection fraction decreases to 50 percent in this experiment. Other experiments, such as altered arterial stiffness, may be performed. The model's flexibility allows description of heart pathology as well as changes in blood vessels. This one ventricular equation with one set of measured parameters is able to describe the wide range of hemodynamics observed experimentally [24, 23].

The previous expressions for ventricular elastance defined in Eqs. (4.73) and (4.74) have the same units as elastance defined classically as Eq. (4.72), but are mathematically not the same. Since ventricular pressure is defined as an analytical function (Eq. (4.75)), ventricular elastance, E_v, defined in the classical sense, may now be calculated as $\partial p_v / \partial V_v$:

$$E_v(t, V_v) = 2a(V_v - b) + c\left[\frac{(1 - e^{-(\frac{t}{\tau_c})^\alpha})e^{-(\frac{t-t_b}{\tau_r})^\alpha}}{(1 - e^{-(\frac{t_p}{\tau_c})^\alpha})e^{-(\frac{t_p-t_b}{\tau_r})^\alpha}}\right] \tag{4.77}$$

or

$$E_v(t, V_v) = 2a(V_v - b) + cf(t) \tag{4.78}$$

Figure 4.45 shows ventricular elastance curves computed using this new analytical definition of elastance (Eq. (4.77)). Elastance was computed for a wide range of ventricular and arterial states, including normal and pathological ventricles, normal and pathological arterial systems, and isovolumic and ejecting beats. These elastance curves are relatively

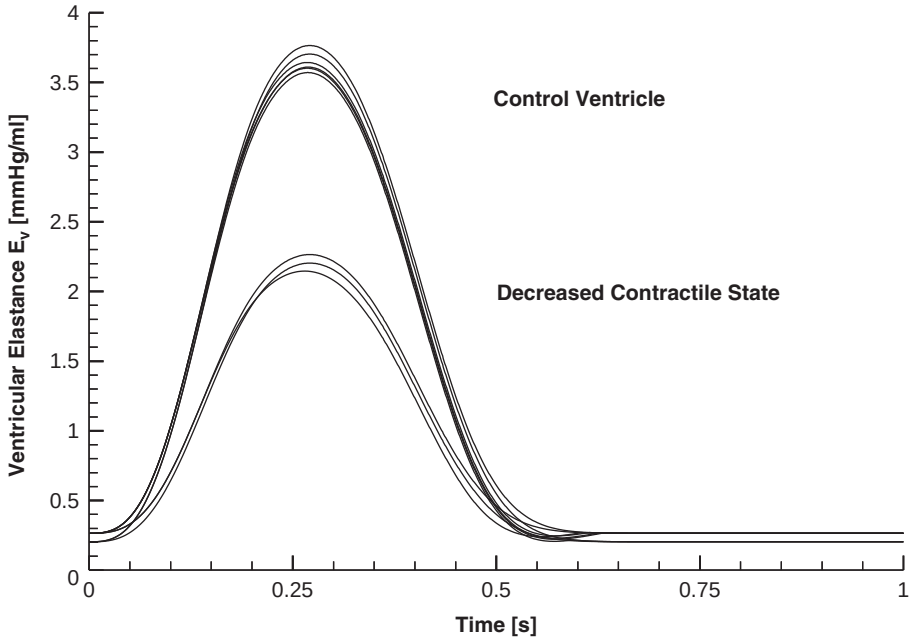


FIGURE 4.45 Ventricular elastance curves computed using the new analytical function of Eq. (4.77). Elastance curves computed in this way are representative of the ventricle's contractile state—that is, its ability to pump blood.

invariant and cluster in two groups: either normal or weakened ventricle contractile state. Consequently, this new measure of elastance may now effectively assess the health of the heart alone, separate from blood vessel pathology.

Chapter 3 gives a brief overview of the circulatory system, a mass and heat transfer system that circulates blood throughout the body. Figure 3.18 shows the four chambers of the heart, the major blood vessels, and valves between the two. From a mechanical point of view, the contracting heart chambers generate pressures that propel blood into the downstream blood vessels. This process is depicted in detail for the most important chamber, the left ventricle, in Figure 4.46. Plotted are representative waveforms for an isolated canine left ventricle, with ventricular pressure and root aortic pressure (top), ventricular volume (middle) and ventricular outflow (bottom), all as functions of time. The numbers 1–4 at the top of the figure correspond to four major phases of the contraction cycle, marked by dashed lines. At time 1, filling is complete, the mitral valve closes, and the ventricle begins to contract isovolumically (no change in volume). At time 2, ventricular pressure exceeds root aortic pressure, the aortic valve opens, and blood ejection begins. Heart valves are passive, and the outflow of blood results simply from the pressure difference across the valve. When ventricular pressure falls below aortic pressure, the aortic valve closes (time 3) and outflow ends. The initial volume (time 1) is denoted end-diastolic volume, EDV, and the volume at the end of ventricular ejection (time 3) is end-systolic volume, ESV. At time 4, ventricular pressure falls below left atrial pressure (not shown), the mitral valve opens, and filling begins in preparation for the next heartbeat. The difference between

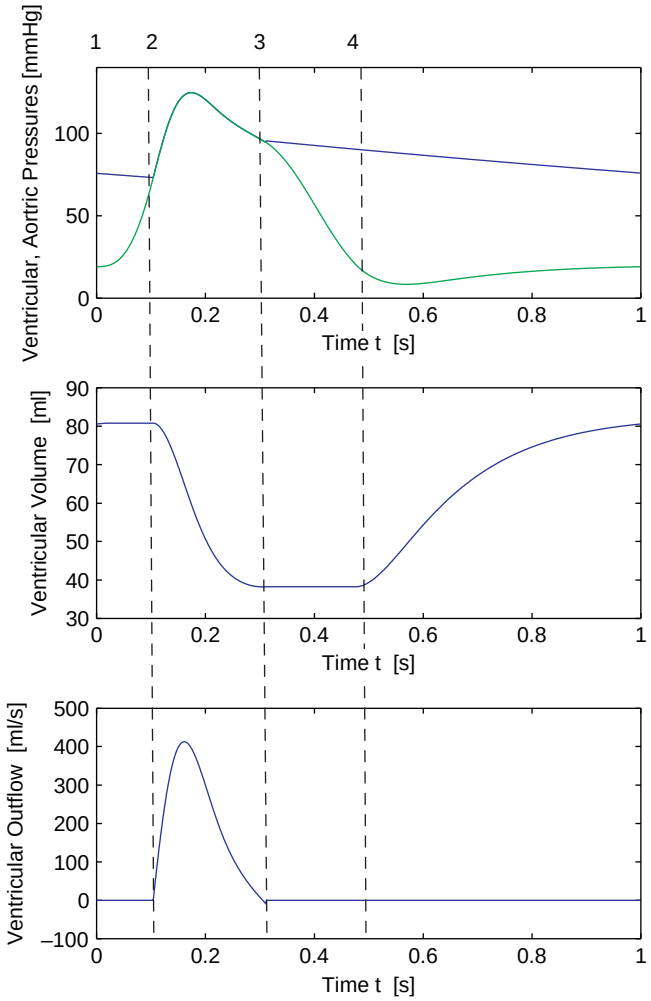


FIGURE 4.46 Representative left ventricular and root aortic pressures (top), ventricular volume (middle), and ventricular outflow (bottom) for an isolated canine heart. The numbers 1–4 at top correspond to distinct phases of the cardiac cycle, described in the text.

end-diastolic and end-systolic volumes, EDV-ESV, is denoted stroke volume, SV, and is the amount of blood pumped in one heartbeat. The ratio SV/EDV is the ejection fraction, EF, and is approximately 50 percent in this example. Clinical blood pressure corresponds to the maximum and minimum arterial pressures—in this example, around 125/75 mmHg.

Clinicians often study heart performance via pressure-volume work loops, plotting ventricular pressure as a function of ventricular volume, as in Figure 4.47. The four phases of the heart cycle just described are plotted for the same heart as in Figure 4.46, with isovolumic contraction between points 1 and 2, ejection from 2 to 3, isovolumic relaxation between 3 and 4, and

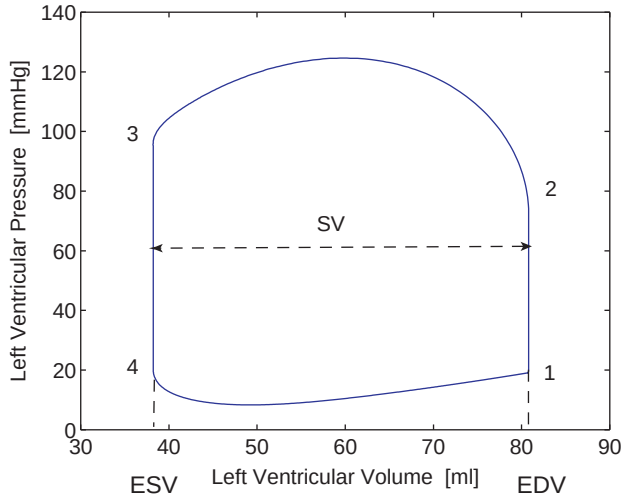


FIGURE 4.47 Pressure-volume work loop corresponding to the left ventricular data in Figure 4.46.

refilling from 4 to 1. Also shown are volumes EDV, ESV, and SV. The area bounded by the work loop corresponds to mechanical work performed by the left ventricle.

Figure 4.48 shows work loops calculated from the generalized pressure model of the left ventricle (Eq. (4.75)) pumping into the three-element arterial load model (Figure 4.37) using human parameter values for both the ventricle and arterial load. The left ventricle was filled from a constant pressure reservoir. The control work loop (center) may be compared to those corresponding to a 10 percent increase and 10 percent decrease in filling pressure.

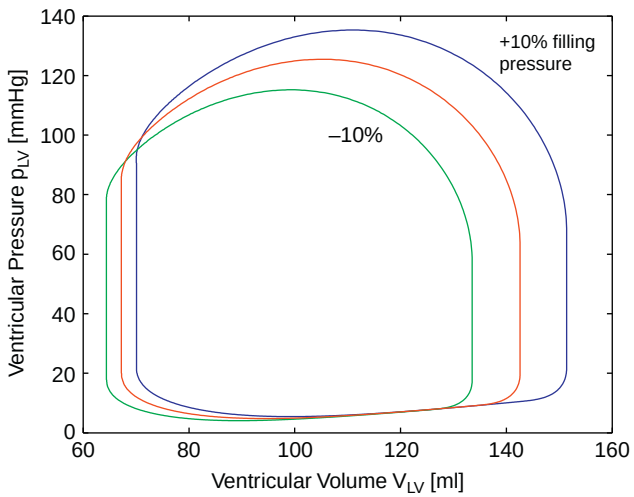


FIGURE 4.48 Pressure-volume work loops computed for the human left ventricle (Eq. (4.75)) and arterial load (Figure 4.37) for control (middle curve) and varied preload (filling) conditions.

These changes directly affect the amount of filling of the ventricle, denoted *preload*. Stroke volume increases with increased preload, and vice versa. As the ventricle is filled more, the end-diastolic volume shifts to the right and the work loop area increases due to the Frank-Starling mechanism. In essence, increased filling stretches the constituent heart muscle, which allows the muscle to generate more force via the force-length relation, the direct relation between muscle length and maximum force of contraction (up to a point). There is also a small shift of end-systolic volume to the right with increased preload. As the heart ejects more blood (SV), the load against which the heart must work is increased.

Figure 4.49 shows work loops for the same ventricle and arterial load for direct variations in *afterload*, the load against which the ventricle pumps. Afterload changes were achieved by varying the value of arterial peripheral resistance R_S. Stroke volume decreases as afterload is increased due to the underlying force-velocity relation of muscle. Higher peripheral resistance forces muscle to operate at a lower velocity of shortening, so less time is available for the ventricle to eject blood. Figure 4.49 shows a small shift of end-diastolic volume to the right with increased afterload. Since stroke volume is reduced, there is remaining blood available for the next heartbeat. This effect is more pronounced in the natural system, which can be examined by extending the model as follows.

The left ventricle model of Eq. (4.75) was used to describe each of the four chambers of the human heart, depicted in Figure 4.50 [25, 23]. This complete model of the circulatory system displays a remarkable range of cardiovascular physiology with a small set of equations and parameters. Changes in blood vessel properties may be studied alone or in combination with altered heart properties. Other system parameters such as atrial performance, as well as other experiments, may be examined. The modular form of this model allows its expansion for more detailed studies of particular sites in the circulatory system.

Figure 4.51 shows work loops computed for the left ventricle in the complete circulatory system model for three different preloads, achieved by increasing and decreasing total

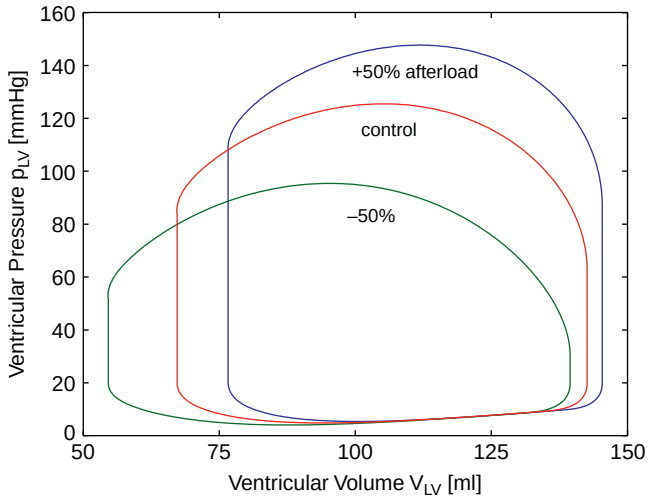
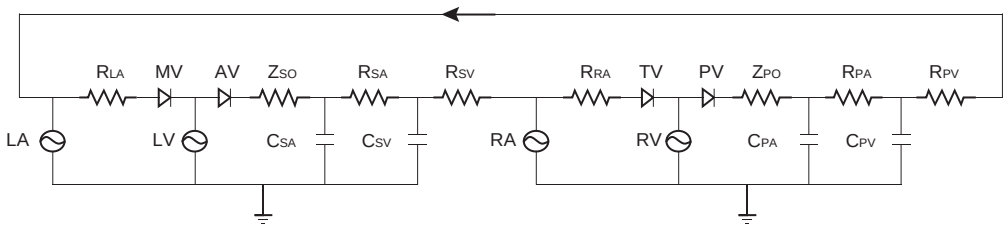


FIGURE 4.49 Pressure-volume work loops computed for the human left ventricle (Eq. (4.75)) and arterial load (Figure 4.37) for control and varied afterload, achieved by varying peripheral resistance R_S.



Systemic Circulation

LA = Left Atrium
R_{LA} = Left Atrial Resistance
MV = Mitral Valve
LV = Left Ventricle
AV = Aortic Valve
Z_{SO} = Systemic Characteristic Impedance
C_{SA} = Systemic Arterial Compliance
R_{SA} = Systemic Peripheral Resistance
C_{SV} = Systemic Venous Compliance
R_{SV} = Systemic Venous Resistance

Pulmonary Circulation

RA = Right Atrium
R_{RA} = Right Atrial Resistance
TV = Tricuspid Valve
RV = Right Ventricle
PV = Pulmonic Valve
Z_{PO} = Pulmonic Characteristic Impedance
C_{PA} = Pulmonic Arterial Compliance
R_{PA} = Pulmonic Peripheral Resistance
C_{PV} = Pulmonic Venous Compliance
R_{PV} = Pulmonic Venous Resistance

FIGURE 4.50 Application of the canine left ventricle model to hemodynamic description of the complete human cardiovascular system. *Adapted from [25].*

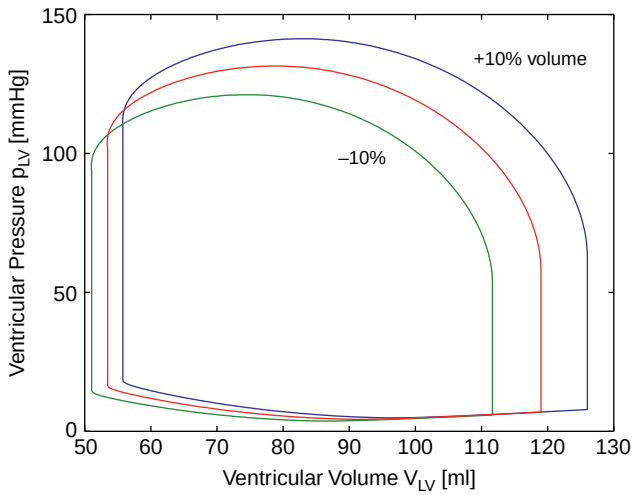


FIGURE 4.51 Left ventricular pressure-volume work loops computed for the complete human circulation model (Figure 4.50) for control (middle curve) and varied preload, achieved by varying total blood volume (5 liters) ±10 percent.

blood volume, 5 liters, by 10 percent. As for the isolated left ventricle model, stroke volume increases with increased preload via the Frank-Starling mechanism, and this increase is moderated by increased end-systolic volume due to increased afterload. Preload may be increased in the natural system by increased central venous pressure, resulting from decreased venous compliance caused by sympathetic venoconstriction, or by augmented

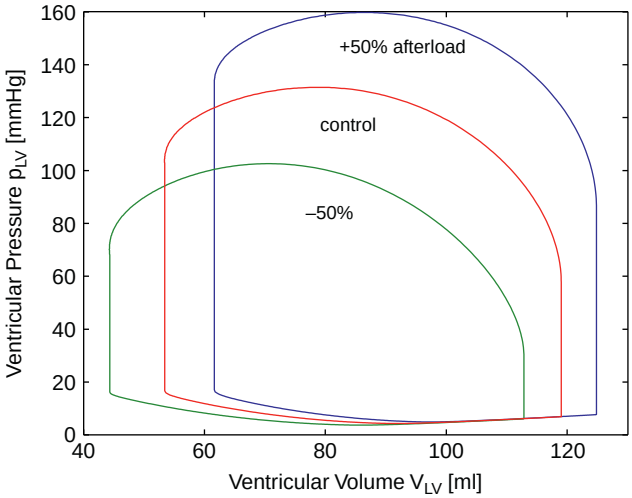


FIGURE 4.52 Left ventricular pressure-volume work loops computed for the complete human circulation model (Figure 4.50) for control and varied afterload, achieved by changing both systemic (R_{SA}) and pulmonic (R_{PA}) peripheral resistances ±50 percent.

venous return due, for example, to gravity via head-down tilt. Preload may also change with total blood volume, which may, for example, decrease as the direct consequence of hemorrhage, or change with renal regulation.

Figure 4.52 shows work loops computed for the left ventricle in the complete circulation model for three different values of afterload, achieved by varying both systemic peripheral resistance, R_{SA}, and pulmonic peripheral resistance, R_{PA}. Stroke volume is inversely related to peripheral resistance, but the Frank-Starling mechanism partially compensates. For increased afterload, smaller SV results in increased filling for the subsequent beat, and this increased EDV moderates the reduction in SV. This compensatory mechanism is more pronounced in the full circulation model than for the isolated left ventricle (Figure 4.49). Afterload commonly increases in the natural system via increased aortic pressure with increased systemic vascular resistance. The latter occurs, for example, when arterial vessel diameter is reduced associated with chronic hypertension. Afterload also increases with aortic valve stenosis, the narrowing of the valve orifice.

The interdependence of preload and afterload is manifested in treatment of heart failure with vasodilator drugs. These drugs decrease afterload, allowing the ventricle to eject blood more rapidly via muscle's force-velocity relation, which increases stroke volume. As SV increases, less blood remains to fill the ventricle for the next beat, but this decrease in EDV is less than the reduction of ESV, resulting in a net increase in stroke volume.

The heart's contractile state may be changed by varying the parameter c in the model (Eq. (4.75)). Figure 4.53 shows left ventricular work loops for such variations in inotropy, executed by changing the contractile parameter c for each of the four heart chambers. Increased inotropy causes an increase in stroke volume, with a decrease in end-systolic volume due to the more strongly contracting heart. End-diastolic volume decreases a small amount

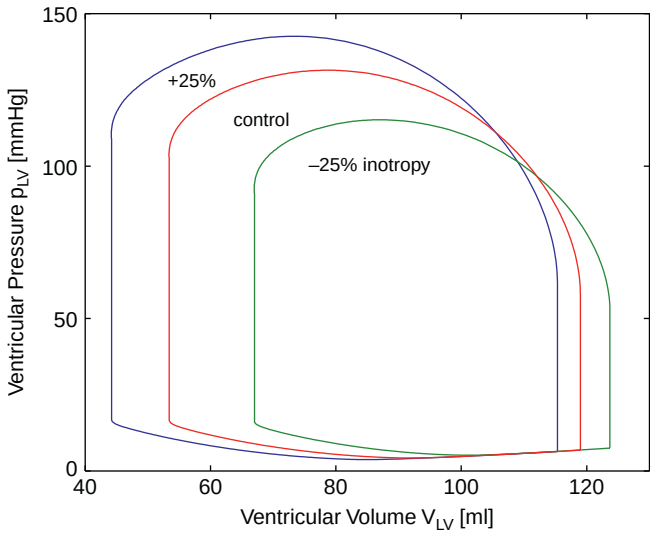


FIGURE 4.53 Left ventricular pressure-volume work loops computed for the complete human circulation model (Figure 4.50) for control and varied inotropy, achieved by changing the contractile parameter c in Eq. (4.75) for each heart chamber ±25 percent.

since the more forcefully ejecting heart leaves less volume available for the subsequent heartbeat. Inotropy is regulated by the autonomic nerves, augmented by sympathetic adrenergic effects of circulating catecholamines, such as epinephrine, the so-called "fight or flight" response to stress. Inotropic drugs used clinically to stimulate the heart in acute and chronic heart failure include digoxin and the beta-adrenoceptor agonists dopamine, dobutamine, epinephrine, and isoproterenol.

The complete circulatory system model may be used to study interactions between left (systemic) and right (pulmonary) circulations. Figure 4.54 shows left and right ventricle work loops for the normal heart ejecting into the normal (control) circulatory system, depicted by solid curves. The right ventricle work loop is smaller, as expected, than the left. Figure 4.54 also shows the same two work loops for a weakened left ventricle (dashed curves). As expected, this left ventricle work loop is diminished in size and shifts to the right on the volume axis. Since the weaker ventricle ejects less blood, more remains to fill the heart more for the subsequent beat (EDV of 192 instead of 119 ml). This increased filling partially compensates for the weakened ventricle via Starling's law (increased pressure for increased filling).

Table 4.5 shows examples of congestive heart failure, resulting from decreases in c for the left ventricle and for the right ventricle. Decreasing left ventricular contractile state to one-third of the control value lowers the left ventricular ejection fraction from 55 to 27 percent, and root aortic pulse pressure decreases from 131/58 to 102/47 mmHg. Left ventricular stroke volume decreases less, from 66 to 53 ml, since it is compensated for by the increased left end-diastolic volume (192 ml) via Starling's law. Decreasing left ventricular contractile state is equivalent to left congestive heart failure. Consequently, pulmonary venous volume increases from 1,540 ml to 2,057 ml (not shown), indicating pulmonary congestion for this case.

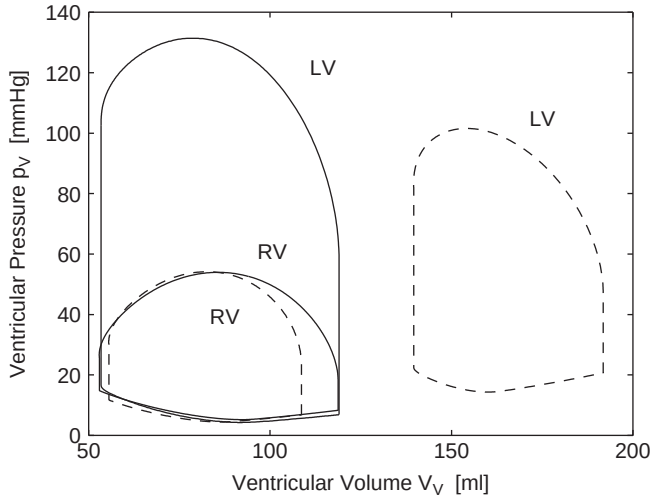


FIGURE 4.54 Computed work loops for the left and right ventricles under control conditions (solid curves) and for the case of a weakened left ventricle (dashed curves).

Similar changes are noted when the right ventricle's contractile state is halved. The right ventricular ejection fraction drops from 55 to 37 percent, root pulmonary artery pulse pressure decreases from 54/18 to 40/15 mmHg, and right stroke volume decreases from 66 to 56 ml, with an increased end-diastolic volume of 153 ml, from 119 ml. Conversely, c can be increased in any heart chamber to depict administration of an inotropic drug. Although not plotted, pressures, flows, and volumes are available at any circuit site, all as functions of time.

In summary, the left ventricle may be described as a dynamic pressure generator. A small number of experimentally derived parameters are sufficient to describe the wide range of observed cardiovascular dynamics. This approach links experiment and theory, leading to new ideas and experiments. It also links underlying muscle dynamics to heart performance. Work is under way to devise a new measure of cardiovascular health using this model.

TABLE 4.5 Cardiovascular Performance for a Normal Heart and for Weakened Left and Right Ventricles

	SV [ml]		EDV [ml]		EF [%]		p_{AO} [mmHg]	p_{PU} [mmHg]
	LV	RV	LV	RV	LV	RV		
Control	66	66	119	119	55	55	131/58	54/18
Weak LV	53	53	192	109	27	49	102/47	54/24
Weak RV	56	56	106	153	53	37	117/54	40/15

Note: SV denotes stroke volume, EDV denotes end-diastolic volume, EF is ejection fraction, and p_{AO} and p_{PU} are root aorta and root pulmonary artery pressures, respectively, for the left (LV) and right (RV) ventricles. Note that SV left and right are equal under all conditions.

The field of biomechanics applies physical principles to living systems using the language of mathematics. Hemodynamics studies the human cardiovascular system, which comprises a complex pump moving complex fluid around an extensive network of complex pipes. In developing hemodynamic principles, experiments and analysis go hand-in-hand, ensuring the validity of principles with experiments and with analysis clarifying, modifying, and often preceding experiments. In this fashion, interpretations of cardiovascular health are further defined.

4.8 EXERCISES

1. Write and evaluate all the vector expressions of Eqs. (4.1) through (4.14) using MATLAB.
2. The measurement error associated with the pointer marker data presented in Example Problem 4.2 is ± 0.5 mm in all three coordinate directions. Given the geometry of the pointer in the example problem, what is the measurement error associated with the pointer tip, point T? If you wanted to minimize the measurement error at point T, how would you design the pointer with respect to the distances between markers A and B, and between marker B and the pointer tip T?
3. Repeat Example Problem 4.3 using a z-x-y rotation sequence.
4. Write the free-body diagrams for each of the three orientations of the humerus in Figure 3.36. For a particular load and fixed position, write and solve the equations of static equilibrium.
5. The force plate in Figure 4.11 is 70 cm wide in the x-direction and 80 cm long in the y-direction. At a particular instant of the gait cycle each transducer reads $F_1 = 150$ N, $F_2 = 180$ N, $F_3 = 220$ N, and $F_4 = 210$ N. Compute the resultant force and its location.
6. Solve Example Problem 4.8 for forearm orientations angled θ from the horizontal position. Let θ vary from 0° to 70° down from the horizontal in 5° increments. Using MATLAB, plot the required biceps muscle force F_B for static equilibrium as a function of θ. By how much does this force vary over this range?
7. Repeat Problem 6, this time plotting forces F_A, F_B, and F_C over the same range of angles θ.
8. Considering the previous problem, explain why Nautilus weight machines at the gym use asymmetric pulleys.
9. Solve Example Problem 4.7 using the moment of inertia of the thigh with respect to the knee.
10. For your own body, compute the mass moment of inertia of the body segments: Forearm, Total Arm, Thigh, Foot, and Trunk in Table 4.1 with respect to their centers of mass.
11. Repeat Example Problem 4.9 using a cobalt alloy rod with circular cross-sectional diameter of 10 mm.
12. Write the Simulink models of the three-element Kelvin viscoelastic description and perform the creep and stress relaxation tests, the results of which appear in Figures 4.26 and 4.27.
13. Use the three-element Kelvin model to describe the stress relaxation of a biomaterial of your choice. Using a stress response curve from the literature, find the model spring constants K_1 and K_2, and the viscous damping coefficient β.
14. Write and solve the kinematic equations defining an anatomically referenced coordinate system for the pelvis, $\{\mathbf{e_{pa}}\}$, using MATLAB.
15. Using the kinetic data of Section 4.6.3, compute the instantaneous ankle power of the 25.2 kg patient using MATLAB.

Continued

16. Pelvic obliquity is commonly associated with an angular displacement of the pelvis as seen from the front of the patient and is estimated as the angle formed between the vector from right to left anterior superior iliac spine (ASIS) and the horizon. Given the pelvic position data below, calculate the clinician's estimate of pelvic obliquity. Note: The clinician would be looking down the x-axis of the global coordinate system at the pelvic data in the y-z plane.

Right ASIS	=	$0.400\mathbf{i} + 0.400\mathbf{j} + 0.580\mathbf{k}$ m
Left ASIS	=	$0.390\mathbf{i} + 0.435\mathbf{j} + 0.600\mathbf{k}$ m
PSIS	=	$0.100\mathbf{i} + 0.418\mathbf{j} + 0.820\mathbf{k}$ m

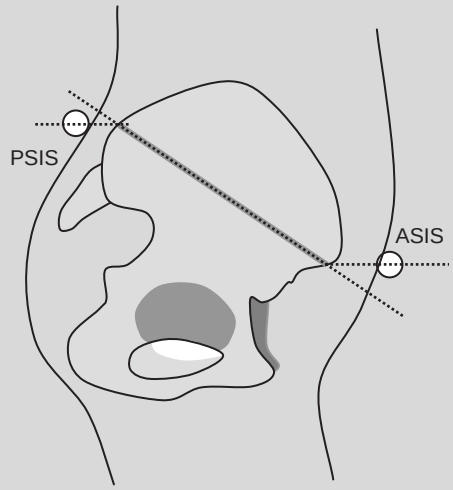
Which Euler angle algorithm, one based on a y-x-z rotation sequence or one based on a z-x-y rotation sequence, would provide a value of pelvic tilt that would best match the clinician's estimate in this case?

17. Given the pelvis and thigh anatomical coordinate systems defined in Example Problems 4.10 and 4.11, compute the pelvis tilt, obliquity, and rotation angles using a z-x-y rotation sequence.

18. Repeat Problem 17 (using a z-x-y rotation sequence) solving for hip flexion/extension, hip abduction/adduction, and internal/external hip rotation. Hint: $\{\mathbf{e_{ta}}\}$ is the triple-primed coordinate system, and $\{\mathbf{e_{pa}}\}$ is the unprimed system in this case.

19. Using MATLAB, determine the effect that a 10 mm perturbation in each coordinate direction would have on ankle power amplitude (Section 4.6.3). Hint: Increase the foot anatomical coordinate system $\{\mathbf{e_{fa}}\}$ by 0.010 m in each direction.

20. Reflective markers placed over the anterior superior iliac spine (ASIS) and the posterior superior iliac spine (PSIS) define the pelvic anatomical coordinate system. An easy mistake to make in the placement of these markers on an overweight or obese subject is to place the ASIS marker(s) at the horizontal level of the ASIS, but not in the pelvic plane, as shown in the figure below. The same type of error can be made with the placement of the PSIS. If the actual pelvic tilt for a patient is 34° and reflective markers are placed in the following pelvic locations relative to the fixed laboratory coordinate system, what is the magnitude of the error in pelvic tilt, if any?



21. Fit the blood rheological data of Example Problem 4.13 to a Casson model and find the yield stress τ_0 for blood.

22. Solve Eq. (4.71) for pressure $p(t)$ when the aortic valve is closed. Using the parameter values in Table 4.4, plot p as a function of time for one heartbeat (t = 0 – 1 sec).

23. Compute isovolumic ventricular pressure $p_v(t)$ for the canine heart with initial volumes $V_v = 30, 40, 50, 60, 70$ ml. Overlay these plots as in Figure 4.41.

24. Write a MATLAB m-file to compute ventricular elastance using Eq. (4.77). Compute and plot $E_v(t)$ for the parameter values in Example Problem 4.15.

25. For the three pressure-volume work loops in Figure 4.52, measure end-diastolic volume EDV, end-systolic volume ESV, stroke volume SV, and ejection fraction EF. Estimate the total mechanical power done by the ventricle in units of watts.

References

[1] R. Baker, Pelvic angles: a mathematically rigorous definition which is consistent with a conventional clinical understanding of the terms, Gait Posture 13 (2001) 1–6.

[2] A.H. Burstein, T.M. Wright, Fundamentals of Orthopaedic Biomechanics, Williams & Wilkins, Baltimore, MD, 1994.

[3] A. Cappozzo, Gait analysis methodology, Hum. Mov. Sci. 3 (1984) 27–50.

[4] K.B. Chandran, S.E. Rittgers, A.P. Yoganathan, Biofluid Mechanics: The Human Circulation, CRC Taylor & Francis Group, Boca Raton, FL, 2007.

[5] R.B.D. Davis III, Musculoskeletal biomechanics: Fundamental measurements and analysis, Chpt. 6 in: J.D. Bronzino (Ed.), Biomedical Engineering and Instrumentation, PWS Engineering, Boston, MA, 1986.

[6] R.B. Davis, S. Õunpuu, D.J. Tyburski, J.R. Gage, A gait analysis data collection and reduction technique, Hum. Mov. Sci. 10 (1991) 575–587.

[7] R. Dugas, A History of Mechanics, Dover Publications, New York, NY, 1988, reprinted from a 1955 text.

[8] R.M. Ehrig, W.R. Taylor, G.N. Duda, M.O. Heller, A survey of formal methods for determining functional joint axes, J. Biomech. 40 (2007) 2150–2157.

[9] R.L. Fournier, Basic Transport Phenomena in Biomedical Engineering, Taylor & Francis, Philadelphia, PA, 1999.

[10] Y.C. Fung, A First Course in Continuum Mechanics, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1977.

[11] Y.C. Fung, Biomechanics: Mechanical properties of living tissues, second ed., Springer-Verlag, New York, NY, 1993.

[12] D.T. Greenwood, Principles of Dynamics, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1988.

[13] A.F. Huxley, Muscle structure and theories of contraction, Prog. Biophys. 7 (1957) 255–318.

[14] A. Kennish, E. Yellin, R.W. Frater, Dynamic stiffness profiles in the left ventricle, J. Appl. Physiol. 39 (1975) 665.

[15] W.R. Milnor, Hemodynamics, second ed., Williams and Wilkins, Baltimore, MD, 1989.

[16] V.C. Mow, W.C. Hayes, Basic Orthopaedic Biomechanics, second ed., Lippencott-Rave, Philadelphia, PA, 1997.

[17] J.P. Mulier, Ventricular pressure as a function of volume and flow, Ph.D. dissertation, Univ. of Leuven, Belgium, 1994.

[18] D.M. Needham, Machina Carnis: The Biochemistry of Muscular Contraction in its Historical Development, Cambridge University Press, Cambridge, U.K., 1971.

[19] B.M. Nigg, W. Herzog, Biomechanics of the Musculo-Skeletal System, third ed., John-Wiley, New York, NY, 2007.

[20] A. Noordergraaf, Hemodynamics, Chpt. 5 in: H.P. Schwan (Ed.), Biological Engineering, McGraw-Hill, New York, NY, 1969.

[21] A. Noordergraaf, Circulatory System Dynamics, Academic Press, New York, NY, 1978.

[22] J.L. Palladino, A. Noordergraaf, Muscle contraction mechanics from ultrastructural dynamics, Chpt. 3 in: G.M. Drzewiecki, J.K.J. Li (Eds.), Analysis and Assessment of Cardiovascular Function, Springer-Verlag, New York, NY, 1998.
[23] J.L. Palladino, A. Noordergraaf, Functional requirements of a mathematical model of the heart, in: Proc. IEEE Eng. Med. Biol. Conf., Minneapolis, MN, 2009, pp. 4491–4494.
[24] J.L. Palladino, J.P. Mulier, A. Noordergraaf, Closed-loop circulation model based on the Frank mechanism, Surv. Math. Ind 7 (1997) 177–186.
[25] J.L. Palladino, L.C. Ribeiro, A. Noordergraaf, Human circulatory system model based on Frank's mechanism, in: J.T. Ottesen, M. Danielsen (Eds.), Mathematical Modelling in Medicine, IOS Press, Amsterdam, Netherlands, 2000, pp. 29–39.
[26] R.J. Roark, Formulas for Stress and Strain, sixth ed., McGraw-Hill, New York, NY, 1989.
[27] C. Singer, E.A. Underwood, A Short History of Medicine, second ed., Oxford Univ. Press, New York, NY, 1962.
[28] H. Suga, K. Sagawa, Instantaneous pressure-volume relationship under various enddiastolic volume, Circ. Res. 35 (1974) 117–126.
[29] E. Weber, B.R. Wagner (Eds.), Handwörterbuch der Physiologie, vol. 3, Vieweg, Braunschweig, 1846.
[30] F.M. White, Fluid Mechanics, sixth ed., McGraw-Hill, New York, NY, 2008.
[31] S. Vogel, Vital Circuits: On Pumps, Pipes, and the Workings of Circulatory Systems, Oxford Univ. Press, New York, NY, 1992.

Suggested Readings

P. Allard, I.A.F. Stokes, J.P. Blanchi (Eds.), Three-Dimensional Analysis of Human Movement, Human Kinetics, Champagne, IL, 1995.
R. Davis, P. DeLuca, Clinical Gait Analysis: Current Methods and Future Directions, in: G. Harris, P. Smith (Eds.), Human Motion Analysis: Current Applications and Future Directions, IEEE Press, Piscataway, NJ, 1996, pp. 17–42.
J.R. Gage, M.H. Schwartz, S.E. Koop, T.F. Novacheck, The Identification and Treatment of Gait Problems in Cerebral Palsy, MacKeith Press, London, U.K., 2009.
W.F. Ganong, Review of Medical Physiology, twenty-second ed., McGraw-Hill, New York, NY, 2005.
J.L. Meriam, L.G. Kraige, Engineering Mechanics, sixth ed., John Wiley, New York, NY, 2008.
W.R. Milnor, Cardiovascular Physiology, Oxford Univ. Press, New York, NY, 1990.
A.D. McCulloch, Cardiac Biomechanics, Chpt. 8 in: D.R. Peterson, J.D. Bronzino (Eds.), Biomechanics Principles and Applications, CRC Taylor & Francis Group, Boca Raton, FL, 2008.
W.W. Nichols, M.F. O'Rourke, McDonald's Blood Flow in Arteries: Theoretical, Experimental and Clinical Principles, third ed., Edward Arnold, London, U.K., 1990.
J. Rose, J.G. Gamble, Human Walking, Lippincott Williams & Wilkins, Philadelphia, PA, 2006.
N. Westerhof, A. Noordergraaf, Arterial Viscoelasticity: A Generalized Model, J. Biomech. 3 (1970) 357–379.
D.A. Winter, Biomechanics and Motor Control of Human Movement, fourth ed., John Wiley, New York, NY, 2009.

CHAPTER

5

Biomaterials

Liisa T. Kuhn, PhD

With contributions from Katharine Merritt, PhD, and Stanley Brown, EngD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Understand the complexity of natural tissue structure that biomaterials scientists seek to replace with biomaterials.
- Describe several different types of biological responses to implanted materials.
- Understand the benefits and differences among the various classes of biomaterials used in medicine.
- Design bio-inspired medical device features to enhance or modify cellular interactions.
- Explain a variety of methods to fabricate scaffolds for tissue engineering.
- Understand the rationale for selecting particular chemistries and structures for several different medical product applications.
- Know where to find the appropriate testing protocols to demonstrate medical product safety.

5.1 MATERIALS IN MEDICINE: FROM PROSTHETICS TO REGENERATION

Throughout the ages, materials used in medicine (biomaterials) have made an enormous impact on the treatment of injury and disease of the human body. Biomaterials use increased rapidly in the late 1800s, particularly after the advent of aseptic surgical technique by Dr. Joseph Lister in the 1860s. The first metal devices to fix bone fractures were used as early as the late eighteenth to nineteenth centuries, the first total hip replacement prosthesis was implanted in 1938, and in the 1950s and 1960s, polymers were introduced for cornea replacements and as blood vessel replacements. Today, biomaterials are used throughout the body (Figure 5.1). Estimates of the numbers of biomedical devices incorporating biomaterials used in the United States in 2006 include the following:

Total hip joint replacements: 579,271
Knee joint replacements: 1,349,641
Shoulder joint replacements: 49,000
Dental implants: 1,040,172
Coronary stents: 1,489,980
Coronary catheters: 1,648,235

Millions of lives have been saved due to biomaterials, and the quality of life for millions more is improved every year due to biomaterials. The field remains a rich area for research and invention because no one material is suitable for all biomaterial applications, and new applications are continually being developed as medicine advances. In addition, there are still many unanswered questions regarding the biological response to biomaterials and

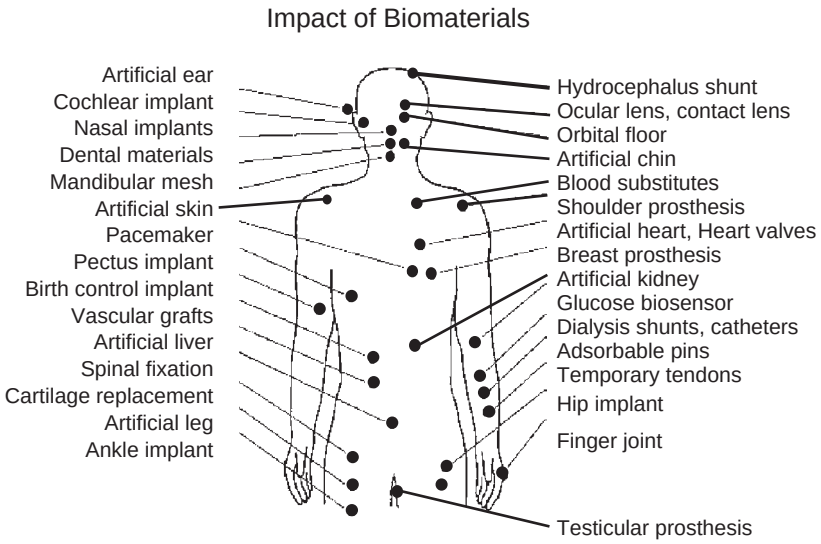


FIGURE 5.1 Biomaterials have made an enormous impact on the treatment of injury and disease and are used throughout the body.

the optimal role of biomaterials in tissue regeneration that continue to motivate biomaterials research and new product development.

Over most of history, minimal understanding of the biological mechanisms of tissues meant that the biomedical engineering approach was to completely replace the damaged body part with a prosthetic—a simple, nonbiologically active piece of hardware. As our understanding of developmental biology, disease, and healthy tissue structure and function improved, the concept of attempting to repair damaged tissues emerged. More recently, with the advent of stem cell research, the medical field believes it will be possible to regenerate damaged or diseased tissues by cell-based tissue engineering approaches (see Chapter 6). The notion of a biomaterial has evolved over time in step with changing medical concepts. Williams in 1987 defined a biomaterial as "a nonviable material used in a medical device, intended to interact with biological systems." This definition still holds true today and encompasses the earliest use of biomaterials for replacing form (e.g., wooden leg, glass eye), as well as the current use of biomaterials in regenerative medical devices such as a biodegradable scaffold used to deliver cells for tissue engineering. While the definition has remained the same, there have been dramatic changes in understanding of the level of interaction of biomaterials with the biological system (in this case, the human body). The expectations for biomaterial function have advanced from remaining relatively inert in the body, to being "bioactive" and not blocking regeneration, to providing biological cues that initiate and guide regeneration. Now there are biomaterials that can initiate a biological response after implantation such as cell adhesion, proliferation, or more excitingly, the differentiation of a stem cell that may one day lead to regeneration of a whole organ.

Due to the complexity of cell and tissue reactions to biomaterials, it has proven advantageous to look to nature for guidance on biomaterials design, selection, synthesis, and fabrication. This approach is known as biomimetics. Within the discipline of biomaterials, biomimetics involves imitating aspects of natural materials or living tissues such as their chemistry, microstructure, or fabrication method. This does not always lead to the desired outcome, since many of the functionalities of natural tissues are as yet unknown. Furthermore, the desirable or optimal properties of a biomaterial vary enormously, depending on where they will be used in the body. Therefore, in addition to presenting general strategies for guiding tissue repair by varying the chemistry, structure, and properties of biomaterials, this chapter includes application-specific biomaterials solutions for several of the major organ systems in the body and for drug delivery applications. This chapter also includes a section on the regulatory approval process and the testing required that play an essential role in establishing and ensuring the safety and efficacy of medical products.

5.2 BIOMATERIALS: TYPES, PROPERTIES, AND THEIR APPLICATIONS

There is a wide choice of possible biomaterials to use for any given biomedical application. The engineer must begin by selecting which general class of material to use. The four basic classes or types of materials are metals, ceramics/glasses, polymers, and composites, which are mixtures of any of the first three types of materials. Natural materials such as animal heart valves are made of proteins that have a repeating polymeric-type structure

and thus fall under the polymer category. Every type of biomaterial can be categorized as belonging to one of these four main classes. It is useful to know the types or classes of materials and the basic properties they possess based on their molecular structure when designing a new medical device. The current uses of the various types of biomaterials for medical devices are shown in Table 5.1, and their mechanical properties are shown in Table 5.2.

TABLE 5.1 Materials and Their Medical Uses

Class of Material	Current Uses
Metal	
Stainless steel	Joint replacements, bone fracture fixation, heart valves, electrodes
Titanium and titanium alloys	Joint replacements, dental bridges and dental implants, coronary stents
Cobalt-chrome alloys	Joint replacements, bone fracture fixation
Gold	Dental fillings and crowns, electrodes
Silver	Pacemaker wires, suture materials, dental amalgams
Platinum	Electrodes, neural stimulation devices
Ceramics	
Aluminum oxides	Hip implants, dental implants, cochlear replacement
Zirconia	Hip implants
Calcium phosphate	Bone graft substitutes, surface coatings on total joint replacements, cell scaffolds
Calcium sulfate	Bone graft substitutes
Carbon	Heart valve coatings, orthopedic implants
Glass	Bone graft substitutes, fillers for dental materials
Polymers	
Nylon	Surgical sutures, gastrointestinal segments, tracheal tubes
Silicone rubber	Finger joints, artificial skin, breast implants, intraocular lenses, catheters
Polyester	Resorbable sutures, fracture fixation, cell scaffolds, skin wound coverings, drug delivery devices
Polyethylene (PE)	Hip and knee implants, artificial tendons and ligaments, synthetic vascular grafts, dentures, and facial implants
Polymethylmethacrylate (PMMA)	Bone cement, intraocular lenses
Polyvinylchloride (PVC)	Tubing, facial prostheses
Natural Materials	
Collagen and gelatin	Cosmetic surgery, wound dressings, tissue engineering cell scaffold
Cellulose	Drug delivery

TABLE 5.1 Materials and Their Medical Uses—Cont'd

Class of Material	Current Uses
Chitin	Wound dressings, cell scaffold, drug delivery
Ceramics or demineralized ceramics	Bone graft substitute
Alginate	Drug delivery, cell encapsulation
Hyaluronic acid	Postoperative adhesion prevention, ophthalmic and orthopedic lubricant, drug delivery, cell scaffold

TABLE 5.2 Mechanical Properties of Materials with Literature Values or Minimum Values from Standards

	Yield MPa	UTS MPa	Deform %	Modulus GPa
Metals				
High-strength carbon steel	1,600	2,000	7	206
F138[1], annealed	170	480	40	200
F138, cold worked	690	860	12	200
F138, wire	—	1,035	15	200
F75[2], cast	450	655	8	200
F799[3], forged	827	1,172	12	200
F136[4] Ti64	795	860	10	105
Gold		2–300	30	97
Aluminum, 2024-T4	303	414	35	73
Polymers				
PEEK		93	50	3.6
PMMA Cast		45–75	1.3	2–3
Acetal (POM)		65	40	3.1
UHMWPE		30	200	0.5
Silicone rubber		7	800	0.03
Ceramics				
Alumina		400	0.1	380
Zirconia, Mg partially stabilized		634		200
Zirconia, Yttria stabilized		900		200
Carbons and composites				
LTI pyrolytic carbon + 5–12% Si		600	2.0	30
PAN AS4 fiber		3,980	1.65	240

Continued

TABLE 5.2 Mechanical Properties of Materials with Literature Values or Minimum Values from Standards—Cont'd

	Yield MPa	UTS MPa	Deform %	Modulus GPa
PEEK, 61% C fiber, long		2,130	1.4	125
PEEK, 61% C fiber, +− 45		300	17.2	47
PEEK, 30% C fiber, chopped		208	1.3	17
Biologic tissues				
Hydroxyapatite (HA) mineral		100	0.001	114–130
Bone (cortical)		80–150	1.5	18–20
Collagen		50		1.2

[1] *F138, wrought stainless steel: 17–19 Cr, 13–15.5 Ni, 2–3 Mo, <2 Mn, <0.08 or <0.03 C.*
[2] *F75, cast cobalt-chromium-molybdenum alloy: 27–30 Cr, <1.0 Ni, 5–7 Mo, <1 Mn.*
[3] *F799, wrought Co-Cr-Mo alloy: 26–30 Cr, <1.0 Ni, 5–7 Mo, <1.0 Mn, <1.5 Fe, <1.5 C.*
[4] *F136 Titanium 6Al-4V alloy: 5.5–5.5 Al, 3.5–4.5 V, <0.015 N, < 0.13 O, <0.08 C.*

5.2.1 Metals

Metallic biomaterials represent the most highly used class of biomaterials. Metals have high strength and resistance to fracture and are designed to resist corrosion. The main metallic biomaterials in use today can be categorized into three groups: iron-base alloys (stainless steels), cobalt-base alloys, and titanium-base alloys. Many orthopedic devices are made in part of metal, such as hip and knee joint replacements (Figures 5.2 and 5.3) due to its high strength and ability to resist failure even after many cycles of loading. The implants provide relief from pain and restore function to joints in which the natural cartilage has been worn down or damaged. Plates and screws that hold fractured bone together during healing also are made of metal and are shown in Figure 5.4. Sometimes the metallic plates and screws are retrieved after successful healing, but in other cases they are left in place. Dental root prosthetic implants are also made of metal (Figure 5.5). Other examples of metals used in medical devices and their mechanical properties are shown in Tables 5.1 and 5.2.

Materials selection for a medical device is complicated. The selection depends on a number of factors, including the mechanical loading requirements, chemical and structural properties of the material itself, and the biological requirements. The longstanding use of metals for knee and hip joints, bone plates, and spinal fusion devices is due to the high mechanical strength requirements of these applications and proven biocompatibility in these settings. The advantages of metals over other materials such as ceramics and polymers are that they are strong, tough, and ductile (or deformable, particularly as compared to ceramics). The metal atoms are arranged in a highly ordered crystalline manner, yet have nondirectional metallic bonding that allows for the propagation of energy-absorbing dislocations of some of the atoms rather than abrupt, catastrophic cracking. Disadvantages include susceptibility to corrosion, again due to the nature of the metallic bond (free electrons). In fact, the steels that were used in the early 1900s for hip implants corroded rapidly in the body and caused adverse effects on the healing process. This has led to the

FIGURE 5.2 A typical total hip joint replacement is made primarily of metal. The ball of the femoral hip stem fits into a pelvic acetabular cup that is lined with ultra high molecular weight polyethylene (UHMWPE) for friction-free motion. *Photograph of the PROFEMUR® Z minimally invasive hip stem with modular necks courtesy of Wright Medical Technology, Inc.*

preferred selection of alloys of titanium or cobalt-chrome for hip, knee, and dental implants. Other typical properties of metallic materials include a high density and much greater stiffness than most natural materials they replace, which lead to undesirable stress shielding. Stress shielding prevents the tissue from being exposed to normal levels of mechanical loading and has been observed after implantation of metal joint replacements to lead to loss of adjacent bone, which requires loading to maintain its mass. Certain metals known as shape memory alloys (e.g., nitinol) can be bent or deformed and still return to their original shape when the stress is released. These metals have found application in eye glasses and coronary artery stents that can be inserted through a catheter while collapsed and then spring into a cylindrical shape once they are pushed beyond the confines of the catheter.

Metallic devices are typically made by investment casting, computer-aided design and machining (CAD/CAM), grinding, or powder metallurgy techniques. The specific steps involved in the fabrication of a medical device will depend on factors such as final geometry of the implant, the forming and machining properties of the metal, and the costs of alternative fabrication methods.

FIGURE 5.3 A metallic artificial knee joint with an ultra high molecular weight polyethylene bearing surface. *Photograph of the ADVANCE® medial-pivot knee system courtesy of Wright Medical Technology, Inc.*

FIGURE 5.4 (a) Metal plates and screws are used to hold fractured bone segments together during healing. Depending on the extent of injury, the plates and screws or rods may be removed when the bone is fully repaired.

Continued

FIGURE 5.4, cont'd (b) Through the use of x-rays, an implanted metal plate with screws can be visualized in this patient's foot and hand. *Photograph of the HALLU®-FIX MTP Fusion System (registered mark of NEWDEAL) is courtesy of Wright Medical Technology, Inc. X-ray courtesy of Wright Medical Technology, Inc.*

FIGURE 5.5 As an alternative to dentures, patients can have metallic dental root prosthetics implanted to replace each missing tooth. The implant is then topped with a porcelain crown. One advantage of dental implants over dentures is that the implant transmits mechanical forces into the jawbone and stimulates it, resulting in less bone recession over time. *Photograph courtesy of Dr. Martin Freilich of the University of Connecticut Health Center.*

EXAMPLE PROBLEM 5.1

A stent is a device that is implanted into an occluded artery to permit increased blood flow. A permanent, yet flexible, device is needed for use as a vascular stent. What material meets that need? In addition to information contained in this chapter, search the web for information on current materials selections using keywords such as "stents" and "metals." The National Institutes of Health PUBMED website catalogs scientific publications within the biological, biomedical, and medical sciences (http://www.ncbi.nlm.nih.gov/entrez). Corporate web pages can provide additional information. Guidant and Boston Scientific are two companies that currently produce coronary stents. The United States Patent and Trademark Office provides another very useful web page for researching uses of materials in surgical and medical devices (www.uspto.gov).

Solution

The preferred materials for stents are metals such as platinum and titanium because they are relatively inert and can be shaped into the helical woven device that can be collapsed and fed through a small tube until the insertion point in the blood vessel. Shape memory alloys such as nitinol also have been used. Stents made of nitinol are self-expanding and "remember" their manufactured shape when they are deployed in the body. They are particularly good for curved or tapered vessels. Metal materials have the strength required for this application.

5.2.2 Ceramics and Glasses

The advantages of the class of materials known as ceramics, which includes glasses, are that they are very biocompatible (particularly with bone), are inert, have low wear rates, are resistant to microbial attack, and are strong in compression. Some disadvantages include brittleness, the potential to fail catastrophically, and being difficult to machine. These properties arise from the atomic structure of ceramics. Unlike metal, in which atoms are loosely bound and able to move, ceramics are composed of typically two different types of atoms that are ionically and/or covalently bound into compound forms. This atomic immobility means that most ceramics do not conduct heat or electricity. A ceramic that does not have a crystalline structure and is amorphous is referred to as a glass. Glasses are often silica-based. Silica is a network-forming oxide that can be heated to its melting point and, unlike most ceramics, is more easily manufactured. Two very obvious properties of ceramics that are different from metals are melting point and brittleness. Ceramics have very high melting points, generally above 1,000°C, and are brittle. Examples of ceramics used in medical devices are shown in Table 5.1. A photograph of a ceramic femoral head of a hip implant is shown in Figure 5.6, and an example of a granular calcium phosphate bone graft substitute is shown in Figure 5.7.

Certain compositions of ceramics, glasses, glass-ceramics, and composites have been shown to stimulate direct bone bonding, which is important in securing orthopedic medical devices such as replacement hips and knees and spinal fusion devices. These types of materials are known as bioactive ceramics. Studies on retrieved implants have shown that a biologically active calcium phosphate forms on the biomaterial surface upon implantation in the body. Since the calcium phosphate that forms is much like that found in our bones,

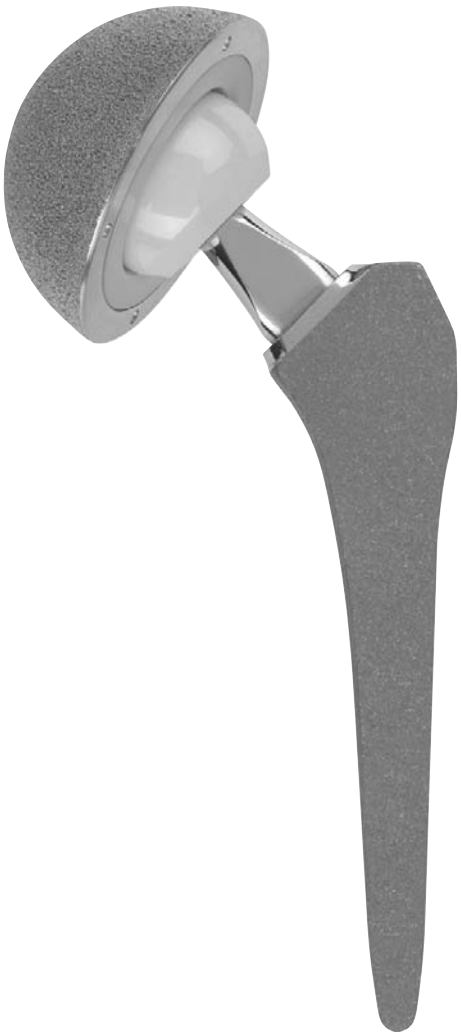


FIGURE 5.6 In this artificial hip joint, the polymer-bearing surface and some of the metallic components have been replaced by ceramics to improve the durability of the joint replacement. This design features a ceramic femoral head and an acetabular cup. *Photograph of the LINEAGE® ceramic-ceramic acetabular cup system is courtesy of Wright Medical Technology, Inc.*

bone cells are able to form an intimate attachment to the biomaterial surface after this bone mineral-like layer has formed. The same results are attained by implanting a material that already has a bone-like calcium phosphate surface, such as the granules shown in Figure 5.7. Further studies have shown that the bone-like calcium phosphate actually provides a direct cue to the cells because of the way the cell attaches to the calcium phosphate.

Several different atomic structures (or phases) of calcium phosphate have been used in medical applications, including hydroxyapatite, carbonated apatite, di-calcium phosphate

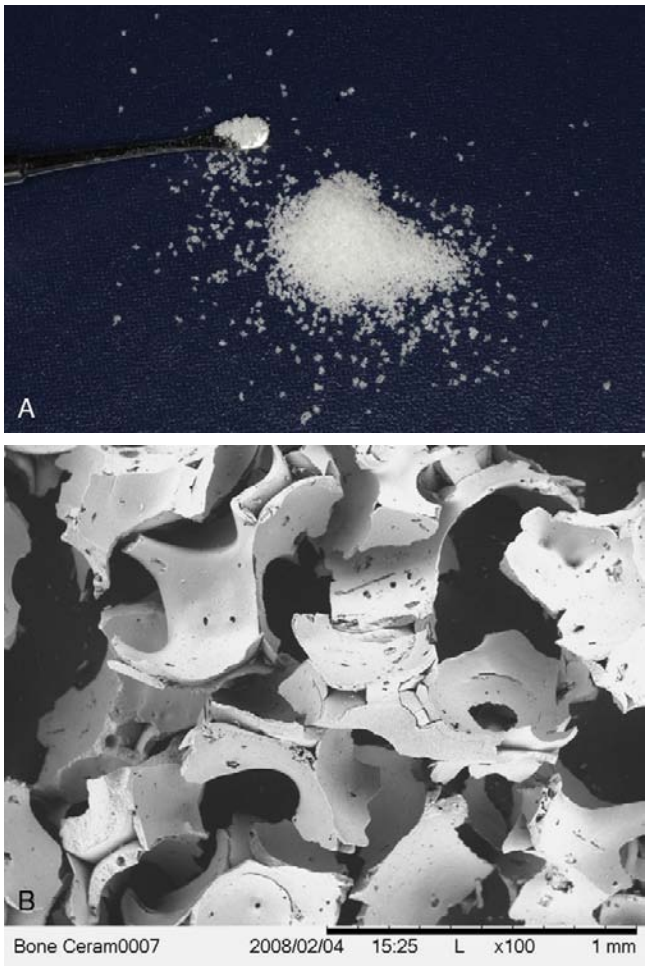


FIGURE 5.7 If there is an insufficient amount of the patient's own bone or donor bone available to fill a bone defect, synthetic bone graft substitutes made of calcium phosphate or calcium sulfate may be used. (a) These are biphasic calcium phosphate granules made of hydroxyapatite and tri-calcium phosphate with an optimized porosity and particle shape used for repair of mandibular bone defects (Straumann® BoneCeramic, Institut Straumann AG). (b) The high porosity visible in this scanning electron micrograph allows maximum space for new blood vessel ingrowth and bone cell influx.

or brushite, beta-tricalcium phosphate or tetracalcium phosphate, and amorphous calcium phosphate. The stability of a given calcium phosphate medical device depends on the crystal phase, the crystal size and perfection, the temperature used during processing, the density, and the in-use environment. At physiological temperature and pH, hydroxyapatite is the stable phase, and it generally takes a long time to resorb via physiochemical dissolution. However, bone cells and other cells called macrophages can initiate cell-mediated resorption of calcium phosphates by changing the local pH to acidic. Nonhydroxyapatite phases of calcium phosphate or other calcium-based biomaterials such as calcium carbonate or calcium sulfate can simply dissolve in the body and do not require cell-mediated resorption.

Due to the high melting point of most ceramics, which prevents them from being cast or extruded, ceramic components are typically made from powdered stock. The powders are formed by wet synthesis methods or by pulverizing raw materials. The ceramic powder is either added to a liquid with binders to form a slurry that is cast in a mold or is dry pressed to form "green ware." The green ware must be finally sintered or fired to densify the powders and remove the porosity between the powder particles. In weight-bearing applications, the porosity must be nearly totally removed, or the residual porosity acts as microcracks within the material and weakens it. In other applications, such as bone graft substitutes or tissue engineering scaffolds, it is desirable to have large pores (<250–750 μm) like those in trabecular or cancellous bone so cells can infiltrate the material and grow new vital tissue. In this case, pores are typically created by using second phases, such as polymer beads, that maintain pore space during the early processing steps and are then burned out during the final sintering stage. More detailed descriptions of how porous scaffolds are formed are included later in this chapter.

EXAMPLE PROBLEM 5.2

What material is preferred for the acetabular cup of a hip implant? What design parameters are utilized during the selection process? Use scientific, corporate, and patent websites to locate information on this topic using keywords such as "ceramic" and "hip replacement."

Solution

Acetabular cups are currently made with a metal support structure and a polyethylene cup; however, problems with wear debris from the soft polyethylene have led to new products with ceramic acetabular cups and femoral balls (alumina or zirconia). The cup must resist wear and deformation and be a low-friction surface because it is in contact with the ball component of the artificial joint. Ceramic materials generate less wear debris during use than the traditional metal on plastic design. Thus, in theory, the risk that ceramic total hips will fail is low compared to the traditional metal on plastic design, and they are now recommended for younger patients. Since the ceramic components are fragile relative to the metallic components and not well tolerated by osteoporotic bone, both types of hip replacements are utilized today.

5.2.3 Polymers

Polymers are well suited for biomedical applications because of their diverse properties and because they resemble natural materials such as protein-based extracellular matrices. For example, by slightly varying the chemical bonding and structure, polymers can be flexible or rigid, or can have low strength or high strength. Depending on the surface modification, they are resistant to protein attachment or encourage protein attachment. Polymers can be biodegradable or permanent, and they can be fabricated into complex shapes by many methods. Some disadvantages of polymers are that they tend to have lower strengths than metals or ceramics, and they tend to deform with time. Certain types of polymers may deteriorate during sterilization and may degrade in the body catastrophically or by release of toxic by-products. Thus, characterization before and after implantation is a standard practice during the development of a new medical device made of a polymer.

Examples of polymers used in medical devices and their mechanical properties are listed in Tables 5.1 and 5.2.

The large macromolecules of commercially useful polymers are synthesized by combining many (poly) smaller molecules (mers) in a process termed *polymerization*. Polymerization may proceed by addition (or chain reaction) polymerization, in which monomer units are attached one at a time and then terminated, or by condensation (or step reaction) polymerization, in which several monomer chains are combined and a by-product of the reaction, such as water, is generated. Additives such as fillers, plasticizers, stabilizers, and colorants typically are used in polymer synthesis to enhance the mechanical, chemical, and physical properties. It is now known that some of the additives used in manufacturing plastic water bottles have hormone-disrupting activity or are carcinogens, and this is an example why polymer additives used for biomedical applications must be carefully screened prior to being implanted in a human.

Polymers can be classified as thermoplastic or thermosetting. A thermoplastic polymer has a linear or branched structure. As a solid it is like a bowl of spaghetti in that the chains can slide over one other. With heating, the chains can slide more easily, and the polymer melts or flows. Thus, thermoplastic polymers can be heated, melted, molded, and recycled. Differences in properties can be achieved with the addition of different ligands. PVC is more rigid than PE because the chlorine atoms are larger and tend to prevent the sliding of one molecule over another. Polymethylmethacrylate (PMMA), as shown in Table 5.2, is stronger, stiffer, and much more brittle than UHMWPE. In this case, two of the four hydrogen atoms are replaced, one with a methyl group (CH_3) and the other with an acrylic group ($COOCH_3$). These large side-groups make sliding much more difficult, hence the increase in strength and modulus. They also make it difficult for the molecules to orient in an orderly, crystalline pattern. As a result of this amorphous structure, PMMA (Plexiglas® or Lucite®) is optically transparent.

Thermoplastic polymers can also degrade in the body by undergoing a hydrolysis reaction in which water causes a breakdown of the polymer chains. Poly(D,L-lactide-co-glycolide) or PLGA is an example of a thermoplastic polyester that is incorporated in a variety of Food and Drug Administration (FDA) approved biomedical devices such as surgical sutures and orthopedic fixation screws and plates. Vicryl sutures are made of a PLGA copolymer that is 90:10 glycolic and lactic acid, respectively. Polyglycolic acid resorbs faster than polylactic acid, and by varying the ratio within the copolymer, the biomaterials scientist can adjust the composition to achieve the desired degradation rate. As PLGA degrades, chemical by-products that are naturally present in the body are released, but they do cause a decrease in local pH that can be harmful to new tissue formation.

In contrast, a thermosetting polymer is composed of chains that are cross-linked. They do not melt with heating but degrade. The term *thermoset* implies that there is a chemical reaction, often involving heat, that results in setting in a three-dimensional cross-linked structure. A common example is "5-minute epoxy." When the two parts are mixed, the catalyst causes setting and cross-linking of the epoxy. Once set, it cannot be heated and reused. The amount of cross-linking affects the mechanical properties. A few cross-links are used to make rubber gloves. Adding more sulfur and cross-linking produces a car tire. Even more cross-links are added to make the hard casing of a car battery.

Hydrogels are water-swollen, cross-linked polymeric structures that have received significant attention because of their many applications in biomedical applications. Hydrogels are prepared by cross-linking the polymer chains while in an aqueous solution by irradiation or chemical methods to create a water-filled polymer network. The most widely used hydrogel is cross-linked polyhydroxyethylmethacrylate (PHEMA). The PHEMA structure has a water content similar to living tissue, has resistance to degradation, is not absorbed by the body, withstands heat sterilization without damage, and can be prepared in a variety of shapes and forms. Applications of hydrogels include contact lenses, drug delivery vehicles, wound healing adhesives, sexual organ reconstruction materials, artificial kidney membranes, and vocal cord replacement materials. The amount of cross-linking within hydrogels is a critical parameter that affects swelling ratios and mechanical strength.

Quite a variety of techniques are employed in forming polymer medical devices. The technique depends on several factors such as whether the material is thermosetting or thermoplastic, and, if thermoplastic, the temperature at which it softens. Thermosetting polymers must be prepared as a liquid linear polymer and then cured in a mold. They cannot be molded after this step. Thermoplastic polymers can be molded repeatedly (by compression, injection, extrusion, etc.), cast, and formed into fibers or films by extrusion followed by drawing or rolling to improve properties such as strength. Later in this chapter many methods will be described for forming three-dimensional porous structures from resorbable polymers that are suitable for tissue engineering applications. The ability to form complex shapes and control the structure at many length scales, as well as the fact they are degradable, make polymers the number one biomaterial type of choice for tissue engineering.

EXAMPLE PROBLEM 5.3

What material is preferred to produce a blood bag? A dialysis bag? What design parameters are involved?

Solution

PVC has been used for blood bags since the 1950s. Since PVC is naturally brittle, phthalate plasticizers are used to make it flexible. These leach out over time from the plastic bag and into the liquid that they contain. When fed in large quantities to rats, the plasticizers can cause cancer; therefore, other plastics are being investigated. Dialysis bags are made of low-density polyethylene (LDPE). In these examples, materials selection has been governed by the fact that the material must be flexible, chemically stable, and relatively inert.

5.2.4 Composites

Composite materials consist of two or more distinct parts. Although a pure material may have distinct structural subunits such as grains or molecules, the term *composite* is reserved for materials consisting of two of more chemically distinct constituents that are separated by a distinct interface. Examples of composites used in biomedical applications include carbon fiber–reinforced polyethylene and hydroxyapatite particle–reinforced polylactic acid

polymers for bone-healing applications. The discontinuous phase is typically harder and stronger than the continuous phase and is called the reinforcement.

Composites are made by mixing two components and molding, compacting, or chemically reacting them together. If the reinforcing primary phase is made of fibers, they are typically coated or impregnated with the second phase so the composite can be heated and pressurized to densify the assembly. A chemical reaction may be utilized to form composites in which a second phase precipitates or forms in place upon reaction. There may be a filament winding process of the first phase around a core if high-strength hollow cylinders are being formed, followed by densification of the construct.

Composites are well suited for devices that require a combination of properties not possible from one single material such as total joint replacements, dental fillings, and bone plates. The advantages of composites are that the properties can be tailored to fit nearly any application; however, it is difficult to make a composite with an ideal structure. There are typically problems with dispersion of the second phase or weak interfacial bonds between the two phases, which leads to less than ideal mechanical properties and thus poor product performance. However, in many cases the actual performance is still much better than any single component biomaterial, so composites are becoming more widely used in biomedical applications.

EXAMPLE PROBLEM 5.4

What materials are preferred for reconstructive dental applications? What are some advantages of the composite structure over a monolithic structure?

Solution

Mercury amalgams made of mercury, silver, and tin are the most commonly used filling materials. Aesthetically pleasing tooth-colored filling materials also can be used and are made of filled resins (e.g., large molecule bifunctional methyacrylates (BisGMA) filled with micro and nanoparticulate silica). PMMA is the predominant material used for complete and partial dentures. Chrome-based alloys are used for the framework of removable dentures. Crowns and bridges are made of a cast metal frame veneered with tooth-colored porcelain. All-ceramic systems are available as well. Recently, composites made of light curing resins reinforced with glass fibers have been developed for dental bridges. The metallic post typically used to provide structural support for crowns is now being replaced with this type of glass fiber reinforced composite, primarily for better aesthetic results; the polymer post does not show through the porcelain crown like the metallic post does.

5.2.5 Natural Materials

Natural materials are synthesized by an organism or plant and are typically more chemically and structurally complicated than synthetic materials. Examples of natural biomaterials currently used in medical devices are listed in Table 5.1. Proteins and polysaccharides are nature's form of polymers and thus can actually be classified into the polymer category. Proteins, the primary group of molecules responsible for making life

possible, are built of long chains of only 20 different amino acids that are held together by peptide bonds. Proteins have a myriad of functions in the human body. They can function as enzymes that catalyze thousands of important chemical reactions that are essential to life. Cell signaling molecules responsible for cell migration and proliferation are made of proteins. Proteins are the building blocks of the supporting extracellular matrix of many tissues. Changes in the levels of proteins or the structure of proteins lead to altered function and are responsible for many diseases. Implanting a natural product made of proteins is usually desired over synthetic polymers because of this ability of the proteins to communicate with cells.

The directional bonds within proteins give rise to the high mechanical properties of natural polymers. For example, the ultimate tensile strength of silk is higher than that of drawn nylon, one of the strongest synthetic polymers. Furthermore, the elastic modulus of silk is nearly 13 times that of the elastic modulus of nylon. As with synthetic polymers, the amount of cross-linking greatly affects the mechanical properties. For example, elastin found in our skin is made of coiled proteins with few cross-links, which makes it much more flexible than Type I collagen found in our bones that is made of rod-like molecules assembled into a repeating crystalline structure and are highly cross-linked.

There are also natural ceramic materials used in biomedical applications. Natural ceramics are typically calcium-based, such as purified calcium phosphate bone crystals or calcium carbonate coral and are both used in orthopedic applications as bone substitutes. Intact bone is a composite of both a natural ceramic and a polymer, and this makes bone much tougher (resistant to fracture) than synthetic ceramics due to their multilayered structure that prevents crack propagation. Small ceramic crystals are precisely arranged and aligned and are separated by thin sheets of organic matrix material that provides an interface. A crack in the material is forced to follow this tortuous organic matrix path.

Natural materials exhibit a lower incidence of toxicity and inflammation as compared to synthetic materials, particularly if they are patient-derived autografts; however, it is often expensive to produce or isolate natural materials. There is also variability between lots of natural materials when they are obtained from a variety of patients (allografts) or plant or animal sources, which makes it difficult to maintain consistency and sometimes prevents widespread commercial use. If the natural tissue comes from a different species than human, such as bovine or porcine sources, it is called a xenograft. The isolation or purification steps typically involve the use of solvents to extract the desired component from the rest of the tissue or the use of solvents to remove the undesired components such as cells from the tissue and leave the desired natural material intact. Collagen can be prepared by either method. If it is labeled as soluble collagen, it has been removed by pepsin enzymatic treatment from natural tissues such as porcine skin. Fibrillar collagen is prepared from natural tissue, such as tendon, by salt and lipid, and acid extraction steps to remove the noncollagenous proteins and molecules, leaving the collagen fibers intact.

Biopolymers also may be produced by bacteria. Production of polyhydroxybutyrate (PHB) is carried out through a fermentation procedure. The bacteria produce the polymer in granules within their cytoplasm when they are fed a precise combination of glucose and propionic acid. The cells are then disrupted, and the granules are washed and collected by centrifugation and then dried. This polymer has properties similar to polypropylene and polyethylene, but it degrades into natural components found in the body. Because of the

desirable environmental characteristics of biopolymers, they are rapidly finding use in several nonmedical niche markets such as biodegradable monofilament fishing nets. The more rapid and widespread introduction of biopolymers has been hindered by their high price (up to ten times) compared to petroleum-based polymers.

Biopolymers also can be produced by chemically polymerizing naturally occurring monomers. Although these polymers are not produced by biological systems, the fact that they are derived from basic biological building blocks makes them biocompatible, nontoxic, and biodegradable. Lactic acid–based polymers (e.g. polyesters such as poly-L-lactic acid or poly-L-glycolic acid) have been used widely for many years for medical devices ranging from biodegradable sutures to tissue engineering scaffolds. Lactic acid is found in blood and muscle tissue and is produced commercially by microbial fermentation of sugars such as glucose or hexose. Poly-L-lactides are frequently used in combination with poly-L-glycolic acid.

EXAMPLE PROBLEM 5.5

What materials are preferred by surgeons for repairing large surgical or traumatic defects in bone? What factors influence this decision? Search the Internet for companies that make bone repair products. Go to the website of the American Academy of Orthopaedic Surgeons and the American Society of Plastic Surgeons for answers to these questions.

Solution

Autograft bone is the first choice for surgeons for the repair of bony defects. The tissues are vital and contain living cells and growth factors that are required for bone regeneration. Allograft bone or demineralized allograft bone matrix is the second choice. Demineralized bone has advantages over as-harvested allograft bone because it is flexible and can conform to the defect site, resorbs more rapidly (within months as compared to years for nondemineralized allograft bone), and releases the bone inductive proteins known as bone morphogenetic growth factors originally discovered by Dr. M. R. Urist in the 1970s.

5.3 LESSONS FROM NATURE ON BIOMATERIAL DESIGN AND SELECTION

Biomedical engineers are asked to design medical devices or systems that repair, monitor, or assist the functions of the human body. Approaches that mimic or replicate nature's techniques, known as biomimetics, are often at the heart of a successful medical device or therapy. There is an incredible complexity to natural tissues and organs that is still far beyond the capacity of scientists to replicate. Furthermore, the precise function of every aspect of the tissues or organs is not known. For these reasons, it is very difficult to theoretically design medical devices, and the field has progressed through a fair amount of trial and error, using materials not synthesized specifically for a biomedical application. Nonetheless, there are several general concepts that have emerged that provide design strategies and guidance for a biomaterials scientist involved in tissue/organ repair and regeneration, and this includes biomimetics.

Imitating the structure of the tissue even at a coarse macroscopic level, such as a semipermeable gel as a temporary skin or using porosity within bone grafts the size of cancellous bone pores, has been shown to improve biological outcomes. Second, it is important to define what native tissue function is most needed, such as mechanical/structural function for bone and cartilage, or chemical function such as islet cells of the pancreas, or it may be a barrier function such as skin. It is difficult to make a fully synthetic organ that can accomplish every native function, so it's important to identify the primary goal and design the material to achieve that goal. Chemists have much to learn from the processing techniques utilized by nature. Nature uses a bottom-up approach to assemble a human being: all materials are made from a relatively simple set of 20 amino acids and four nucleic acids. It is the elegant way they are assembled and organized into larger structures that endows the body with its amazing functions. This section attempts to describe some of the key elements of natural tissue assembly and fabrication.

5.3.1 The Extracellular Matrix: Nature's Biomaterial Scaffold

Every tissue in the body has an extracellular matrix (ECM) that provides a physical and chemical support structure for cells. From a biomaterials perspective, it is a complex nanocomposite made of a woven polymeric material (fibrous proteins) with reinforcing struts (other fibrous proteins) and gels to resist compaction (water-swollen proteoglycans). A full description of the protein and polysaccharide components of the ECM and their functions can be found in Chapter 7. At the simplest level, the ECM provides a surface for cells to adhere to and to live within and to migrate across. The ECM has attachment sites for cell signaling proteins and sequesters them until encountered by the cells. It is a three-dimensional biomaterial scaffold that interconnects all the cells physically and mechanically. Cells attach to the ECM via specific cell surface receptors that govern proliferation, differentiation, and protein expression. Interactions of the cells with the ECM thus play a crucial role during tissue regeneration.

The field of biomaterials seeks to mimic the structure of the ECM to guide cell-biomaterial interactions toward regeneration and healing. The ECM is made by spontaneous folding of proteins produced by cells, typically different collagenous proteins (there are more than ten) and large glycoproteins (e.g., fibronectin, laminin, ostepontin). It is dynamic and is constantly being modified by the cells. The spontaneous folding of peptides is known as self-assembly. The collagen molecule self-assembles from three alpha chains into a triple helix. Within the triple helix, glycine must be present as every third amino acid, and proline and hydroxyproline are required to form and stabilize the triple helix. The collagen molecules then assemble together, a few molecules at a time, with a quarter overlap to form a staggered linear array. The linear aggregates then laterally associate into bundles (Figure 5.8). Hole zones are left open between the collagen molecule terminal groups for subsequent mineralization (as described further in the example on bone biomineralization later in this chapter). The molecular assembly of DNA is somewhat similar to that of collagen: Two preexisting complementary DNA chains combine to form a double helix and then the helices assemble on their own guided by positive and negative charges interacting with each other. Self-assembly is one of the key design strategies of tissue formation. Biomaterials scientists are now utilizing this principle to fabricate complicated, yet highly ordered, microstructures at the nanometer-length scale.

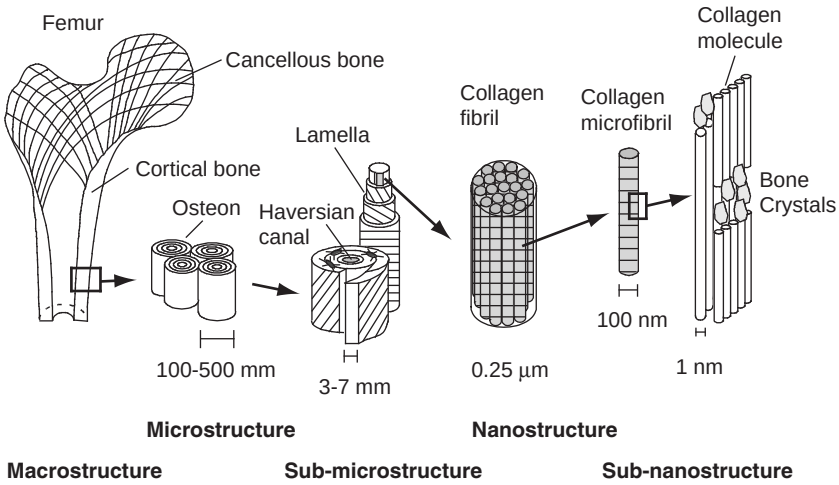


FIGURE 5.8 The hierarchical structure of bone. There are at least five levels of microarchitecture in bone: (a) collagen molecules, (b) microfibrils with periodic gaps between the ends of the molecules that provide nucleation sites for bone mineral, (c) fibrils composed of the microfibrils, which form (d) the layers of lamellar bone shown within a cylindrical osteon, that are the building blocks of the cortical bone of (e) the femur.

5.3.2 Hierarchical Design

Similar to a nested set of eggs or Russian dolls, in which you open up one just to find another smaller one inside, natural tissues have nested structures. This is known as hierarchical design. The same structural motif is repeated at multiple-length scales, and this endows the tissue with efficiency of assembly, properties, and function. During cell-mediated tissue construction, the smallest units self-assemble first, then these units self-assemble to form larger units, and finally the larger units self-assemble. This is how a functional tissue or organ is built. Cells typically become embedded or trapped within the layers of extracellular matrix as the tissue is built. The cells survive and act to maintain the tissue around them. Lamellar bone has similar layered fibrillar structures at the nano, micro, and macro levels. At the smallest level there are the alpha chains that have assembled in a triple helix to form the Type I collagen molecule, the collagen molecules then assemble into microfibrils, microfibrils assemble into larger fibers that assemble by alignment into sheets, and the sheets (lamellae) are layered like plywood in a crisscrossed orientation around blood vessels to form osteons, which have a tubular or large fiber appearance (see Figure 5.8).

Skeletal muscle also has a hierarchical structure. The actin and myosin filaments organize into linear constructs known as sarcomeres, which are bundled into myofibrils, which are then bundled together to form a muscle fiber (the muscle cell with hundreds of nuclei), and multiple muscle fibers form the muscle fascicles we call muscle. (See Chapter 3 for a diagram of this.)

The biological structure of chromosomes is also governed by hierarchical design. Nature efficiently compacts a 7-cm-long strand of DNA until it is 10,000 times smaller by twisting and coiling at multiple-length scales. At the nanometer level, DNA is a double helix of

single-stranded DNA. Subsections of the twisted DNA strand (146 base pairs each) are then subcoiled around protein (histone) cores to form chromatosomes that cause the strand to resemble a beaded necklace. The linked chromatosomes are then coiled to generate a shorter, thicker fiber. The thicker fiber is then coiled upon itself to further shorten its length to form the chromatid, two of which are linked together to form a chromosome.

Hierarchical design leads to a very efficient structure with maximum capability per ounce of material. In bone, the multiple layers each have an interface between them that leads to fracture resistance. In muscle, the many single fibers aligned and pulling together demonstrate inordinate strength. The amount of information contained in our DNA is staggering. Hierarchical design is thus another important concept that should be mimicked in biomaterials design.

5.3.3 Biomineralization

Bones, antlers, teeth, coral, eggshells, and seashells are all examples of biomineralized tissues. Although they serve different functions and have considerably different external shapes, from the biomaterials perspective, they are all composed of numerous small, isolated calcium phosphate or calcium carbonate crystals that are held together by a protein matrix. The nucleation and growth of the mineral crystals are highly regulated by the organic protein component that is secreted by cells and self-assembles to provide a template for the mineral growth and nucleation. Three general processing principles for biomineralized tissues have been identified that have significant implications for material scientists and engineers:

1. Biomineralization occurs only within specific subunit compartments of microenvironments, which implies stimulation of crystal production at certain functional sites and inhibition or prevention of the process at other sites.
2. A specific mineral phase is produced with a defined crystal size (frequently in the nanometer range), shape, and orientation.
3. Macroscopic shape forming is accomplished by packaging many incremental units together, which results in unique net-shaped composites with hierarchical microstructures that impart exceptional material properties. Point number 3 is another example of how nature uses a bottom-up approach to assemble a complex tissue. Another powerful feature of biomineralization is that, in most systems, remodeling of the original mineral structure occurs as needed to optimize strength, accommodate organism growth, maintain mineral ion equilibrium, and effect repairs. The initial biomineralized structure that is formed after injury typically has a looser structure that restores some level of function, but is not optimal. This is followed by a remodeling phase, so the biomaterials scientist may not need to recreate the final tissue and could instead allow the cells to form the final, most optimal, tissue over time.

In bone, controlled mineral nucleation and growth are accomplished within the micro-compartments formed by the collagen matrix. The Type I collagen molecules secreted by osteoblasts self-assemble into microfibrils with a specific tertiary structure having a 67-nm periodicity and 40-nm gaps or holes between the ends of the molecules (Figure 5.8). The holes localize a microenvironment containing free mineral ions and bound side chain

groups from phosphoproteins attached to the collagen. The molecular periodicity of the functional groups serves to nucleate the mineral phase heterogeneously. The nucleation of the thin, plate-like apatite crystals of bone occurs within the discrete spaces within the collagen fibrils, thereby limiting the possible primary growth of the mineral crystals and forcing them to be discrete and discontinuous. Only one phase of calcium phosphate is nucleated during normal, nonpathological mineralization processes (carbonated apatite), and the minerals grow with a specific crystalline orientation. In biomimetic ceramics, the calcium phosphate stimulates bone healing best when used in the form of nanometer-sized crystals rather than large crystals that are fused together by the application of high temperatures.

5.4 TISSUE–BIOMATERIAL INTERACTIONS

5.4.1 Interactions with Blood and Proteins

The implantation of a biomaterial often creates a wound, and bleeding generally ensues. Blood thus typically makes first contact with the implanted biomaterial. Blood is a mixture of water, various kinds of cells and cell fragments (platelets), salts, and proteins (plasma). Proteins play an important role in determining the biological activity of the tissue–implant interface. Biomaterials can promote cell/tissue attachment and activity by allowing selective protein adsorption or can inhibit tissue interactions by repelling protein. Most importantly, changes in the microenvironment that can occur after biomaterial implantation, such as changes to pH and ionic strength, can alter the conformation of a nearby protein and thus its function. Proteins also can experience structural alterations during interaction with the solid surfaces of biomaterials and lose some of their biological activity. Albumin is the most common protein in blood, followed by the protective immune system proteins known as immunoglobulins, which are involved in the recognition and adhesion processes of cells. However, because exchange between absorbed proteins occurs, the final layer of absorbed protein may be fibrinogen, which although less abundant, may have a greater affinity for the biomaterial surface. The important message is that proteins guide cell attachment, proliferation, and differentiation, so it is important to know which proteins are attached to the biomaterial after implantation. The specific proteins that may have been deliberately and carefully placed on the biomaterial before implantation, may no longer be available to the cells after blood contact due to additional adsorption of blood proteins.

Blood coagulation is directed by attachment of the protein clotting factor XII, which is found in blood, to the foreign biomaterial surface. After attachment of this factor, platelets from the blood can and will adhere to the biomaterial, which leads to fibrin clot formation. A cascading chain of cellular reactions that is governed by the initial protein attachment begins. Blood contact provides the cells and cytokines that participate in the biological interaction with the biomaterial. Therefore, every biomaterial that contacts blood and absorbs protein will elicit biological responses from the body. In the case of artificial blood vessels, formation of a blood clot must be avoided to prevent thrombosis, which is the blocking of flow through a blood vessel by a blood clot. An effective and widely used method for preventing surface-induced thrombosis of polymer materials intended for contact with blood is by modifying their surface with heparin, through the ionic or covalent addition of this anticoagulant.

5.4.2 The Wound Healing Response after Biomaterial Implantation

The implantation of a biomaterial creates a disruption of the anatomic continuity of tissue and, as such, creates a wound. The body has a highly developed wound-healing response that is immediately triggered by the wounding during biomaterial implantation. Much is known about the normal cellular events that transpire after the initiation of a wound (see Chapter 7), and this knowledge provides a foundation to understand and anticipate tissue–biomaterial interactions. From the perspective of tissue–biomaterial interactions, four overlapping phases will always occur (Figure 5.9):

1. *Hemostasis:* Platelet cells control bleeding through coagulation by adhering to the proteins attached to the biomaterial surface and by releasing clot-forming proteins. The clot that is formed acts as a provisional matrix for the initiation of repair tissue and fills the gaps around the implanted biomaterial.
2. *Inflammation:* Clot formation induces the production of cell-signaling molecules (cytokines) that induce the recruitment of inflammatory cells from a nearby bloodstream. These cells (neutrophils, monocytes, lymphocytes, and macrophages) arrive and attempt to digest tissue debris and the biomaterial by a process known as phagocytosis. The growth factors released at the wound site by the inflammatory cells initiate mitosis (cell replication) of sedentary connective tissue cells at the wound margin.
3. *Proliferation/initial repair:* As a result of all the growth factor signaling by the inflammatory cells, there is a proliferation and population of the biomaterial with cells that can recreate the lost or damaged tissue. A nondegrading biomaterial located in the center of the wound typically becomes encapsulated with tight fibrous tissue. The fibrous capsule isolates the material from the biological environment and protects the host. The extent of the inflammatory foreign body response governs the thickness of the fibrous capsule. The chemical characteristics, the shape and physical properties of the biomaterial implant, and the rate of release, accumulation, and bioactivity of released chemicals and corrosion products from implanted materials all also affect the thickness of the fibrous capsule. If the implant is permanent and does not biodegrade, then a small capsule remains throughout

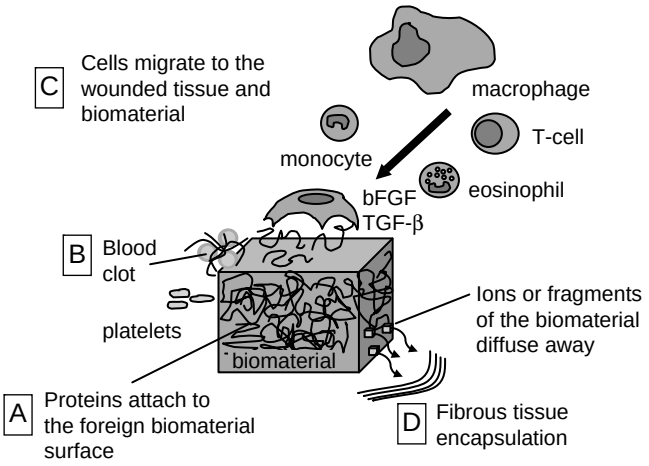


FIGURE 5.9 Normal tissue–biomaterial interactions involve the four overlapping and interdependent phases of wound healing: hemostasis, inflammation, proliferation/repair, and tissue remodeling. (a) Protein attachment to the biomaterial surface guides cellular interactions. (b) Hemostasis is accomplished by clot formation. (c) Cells found in blood and other inflammatory cells attempt to process the foreign biomaterial and repair adjacent material. (d) The host protects itself from the foreign biomaterial through encapsulation with fibrous tissue.

the life of the implant, except in bone, where there is direct bone apposition on calcium phosphate surfaces without an intervening fibrous tissue layer.

4. *Remodeling:* The rapidly formed neotissue will be remodeled by cells into functional tissue more similar to the original tissue, although typically a scar remains.

5.4.3 Metallic Corrosion

There are a number of mechanisms by which metals can corrode, and corrosion resistance is one of the most important properties of metals used for biomedical implants. The mechanisms of most significance to implant applications in the aqueous saline solutions of the human body are galvanic (or mixed metal) corrosion, crevice corrosion, and fretting corrosion.

Galvanic (mixed metal) corrosion results when two dissimilar metals in electrical contact are immersed in an electrolyte. There are four essential components that must exist for a galvanic reaction to occur: an anode, a cathode, an electrolyte, and an external electrical conductor. The in vivo environment contains electrolytes. A patient with two total hip replacements made of different alloys is not subject to mixed metal corrosion, since there is no electrical connection. However, a hip replacement made of two alloys or a fracture plate of one metal fixed with screws of another metal may be susceptible to mixed metal corrosion.

When two dissimilar metals are connected in an electrochemical cell, one will act as an anode, and the other will be the cathode. Metal oxidation will occur at the anode, as shown in Eq. (5.1). Metal oxidation may produce metal ions, and these ions can migrate away from the metal surface as free ions in solution or form metal oxide or other chemical compounds with other species. Various elements (e.g., nickel, chromium, etc.) are not well tolerated by tissue and have possible adverse reactions, including allergic and inflammatory reactions. The reaction at the cathode will depend on the pH of the environment (Eqs. (5.2) and (5.3)) and will lead to immunity from corrosion, passivation, or active corrosion. The direction of the reaction can be determined by examining the electromotive force (EMF) series, a short listing of which is shown in Table 5.3. These potentials represent half-cell potentials of metals in equilibrium with 1 molar solution of their ionic species. The potential for hydrogen is defined as zero. As shown, the standard potential for iron is -0.44 V. If iron is connected to copper with an EMF of $+0.34$ V, the potential difference is 0.78 V. Since iron is the anode, iron oxidation will occur according to the reaction shown in Eq. (5.1). The reaction at the copper cathode will depend on the pH of the solutions, as shown in Eqs. (5.2) and (5.3).

$$\text{anodic reaction } \mathrm{Fe} \rightarrow \mathrm{Fe}^{+2} + 2\mathrm{e}^{-} \tag{5.1}$$

$$\text{cathodic reaction (acidic solution) } 2\mathrm{H}^{+} + 2\mathrm{e}^{-} \rightarrow \mathrm{H}_2 \tag{5.2}$$

$$\text{cathodic reaction (neutral or basic) } \mathrm{O}_2 + 2\mathrm{H}_2\mathrm{O} + 4\mathrm{e}^{-} \rightarrow 4\mathrm{OH}^{-} \tag{5.3}$$

Because the free energy per mole of any dissolved species depends on its concentration, the free energy change and electrode potential of any cell depends on the composition of the electrolyte. Thus, the direction and rate of the reactions also depends on the concentration of the solutions. Increasing the concentration of Fe^{+2} in the environment will shift the

TABLE 5.3 Electromotive Force Series: Standard Reduction Potentials (E^0 V) in Aqueous Solution at 25°C

$K = K^+ + e^-$	–2.93 Active (more anodic)
$Na = Na^+ + e^-$	–2.71
$Al = Al^{+3} + 3e^-$	–1.66
$Ti = Ti^{+2} + 2e^-$	–1.63
$Zn = Zn^{+2} + 2e^-$	–0.76
$Cr = Cr^{+3} + 3e^-$	–0.74
$Fe = Fe^{+2} + 2e^-$	–0.44
$Co = Co^{+2} + 2e^-$	–0.28
$Ni = Ni^{+2} + 2e^-$	–0.25
$Sn = Sn^{+2} + 2e^-$	–0.14
$H_2 = 2H^+ + 2e^-$	0.000
$Cu = Cu^{+2} + 2e^-$	+0.34
$Ag = Ag^+ + e^-$	+0.80
$Pt = Pt^{+2} + 2e^-$	+1.20
$Au = Au^{+3} + 3e^-$	+1.50 Noble (more cathodic)

potential in the positive or noble direction. As the Fe^{+2} concentration increases, the potential difference between the iron and copper will become less as the iron becomes more cathodic. Similarly, the concentration of oxygen at the cathode will affect the EMF of the cell. Increasing O_2 will make it more noble, while decreasing O_2 will make it more anodic. In fact, crevice corrosion is initiated by changes in oxygen concentration, as is discussed in a following paragraph.

Galvanic cells occur not only with different alloys but also with differences within an alloy. Carbides, grain boundaries, and different phases within an alloy also present differences in EMF and thus the possibility for localized galvanic cells. Cold working also increases the free energy of metal and thus its susceptibility to corrosion. Bending a plate or pounding on a nail head causes localized cold working and makes that area anodic to the rest of the piece.

Galvanic corrosion can also be utilized to prevent corrosion by cathodically polarizing the part to be protected. Steel ships are protected from rusting by the attachment of blocks of zinc. The zinc blocks ("zincs") serve as a sacrificial anode and protect the steel hull. Metal pumps and other metallic components on ships are also protected with zincs. A power supply can be attached to a part, such as in a steel underground pipeline, to make the pipe cathodic to a replaceable anode. This protects the pipeline.

Electrode size also has an effect on galvanic reaction rates. The classic example is the difference between galvanized and tin-plated iron. As Table 5.3 shows, zinc is anodic to iron. Thus, galvanization results in coating the iron with an anodic material. When the zinc

is scratched and the iron is exposed, the small size of the iron cathode limits the reaction, and there is minimal corrosion. In contrast, tin is cathodic to iron. When a tin plate is scratched, the small iron anode is coupled with a large cathode. Anodic corrosion and removal of iron from the scratch result in an increased area of the exposed iron and thus an increase in corrosion rate. This self-accelerating corrosion can be prevented by coating the tin cathode with a nonconductive material such as paint or varnish.

Crevice corrosion can occur in a confined space that is exposed to a chloride solution. The space can be in the form of a gasket-type connection between a metal and a nonmetal or between two pieces of metal bolted or clamped together. Crevice corrosion involves a number of steps that lead to the development of a concentration cell, and it may take six months to two years to develop. Crevice corrosion has been observed in some implanted devices where metals were in contact, such as in some total hip replacement devices, screws and plates used in fracture fixation, and some orthodontic appliances.

The initial stage is uniform corrosion within the crevice and on the surfaces outside the crevice. Anodic and cathodic reactions occur everywhere, with metal oxidation at the anode and reduction of oxygen and OH- production at the cathode. After a time, the oxygen within the crevice becomes depleted because of the restricted convection of the large oxygen molecule. The cathodic oxygen reduction reaction ceases within the crevice, but the oxidation of the metal within the crevice continues. Metal oxidation within the crevice releases electrons that are conducted through the metal and consumed by the reduction reaction on the free surfaces outside the crevice. This creates an excess positive charge within the crevice that is balanced by an influx of negatively charged chloride ions. Metal chlorides hydrolyze in water and dissociate into an insoluble metal hydroxide and a free acid (Eq. (5.7)). This results in an ever-increasing acid concentration in the crevice and a self-accelerating reaction.

$$M^+Cl^- + H_2O \rightarrow MOH + H^+Cl^- \tag{5.4}$$

The surgical alloys in use today all owe their corrosion resistance to the formation of stable, passive oxide films, a process called passivation. Titanium, which appears as an active metal on the EMF series in Table 5.3, forms a tenacious oxide that prevents further corrosion. Stainless steels and cobalt alloys form chromium oxide films. As indicated in Table 5.3, they are active, but in the environment where this oxide film is formed, they become passive or noble.

To be self-passivating, stainless steels must contain at least 12 percent chromium. However, carbon has a strong affinity for chromium, and chromium carbides form with the average stoichiometry of $Cr_{23}C_{5.}$ The formation of a carbide results from the migration of chromium atoms from the bulk stainless steel alloy into the carbide. The result is that the carbide has high chromium content, while the alloy surrounding the carbide is depleted in chromium. If the chromium content is depleted and drops below 12 percent Cr, then there is insufficient Cr for effective repassivation, and the stainless steel becomes susceptible to corrosion. As a safety factor, surgical stainless contains 17 to 19 percent chromium, and the carbon content in surgical alloys is kept low at less than 0.08 percent or more than 0.03 percent.

The problem of carbide formation is especially important with welded stainless steel parts. If steel is heated to the "sensitizing range" of 425°C to 870°C, the chromium can

diffuse in the solid and form carbides. At temperatures above 870°C, the carbon is soluble in the atomic lattice. Below 425°C, the mobility is too low for carbide formation. If the peak temperature in the metal away from the weld is in the sensitizing range, carbides can form. This is known as weld decay or corrosion of the sensitized metal on each side of the weld. By heat treating after welding, the carbides can be redissolved, and the metal quickly quenched to avoid reformation.

With the oxide film intact, surgical alloys are passive and noble. If the film is damaged, as with scratching or fretting, the exposed metal is actively corroding. Reformation or repassivation results in restoration of the passive condition. Fretting corrosion involves continuous disruption of the film and the associated oxidation of exposed metal. Devices that undergo crevice corrosion are also examples in which fretting corrosion has accelerated crevice corrosion.

How often does corrosion of metallic implants happen in the patient? The most detailed information on *in vivo* corrosion of biomaterials comes from retrieval studies in which the implant is removed from the patient. In one study in which 148 total modular hip implants made with mixed –(Ti-6Al-4V/Co-Cr) and similar-(Co-Cr/Co-Cr) alloy femoral stem and head combinations were examined, moderate to severe corrosion was observed in 16 percent of necks and 35 percent of heads in the mixed-metal cases and in 14 percent of necks and 23 percent of heads for similar-alloy cases. This corrosion process is an example of mechanically assisted crevice corrosion attack.

In another study, Type 316 stainless steel multicomponent devices were removed from the human body, and 91 percent were found to have undergone visible corrosion, with crevice corrosion being the dominant form. So to conclude, corrosion does occur to metallic implants in the body, and this damages the metallic biomedical devices. Prior to implantation, these devices had been screened in simple laboratory tests under ideal conditions at neutral pH that did not replicate the complex environments that arise as a result of inflammation and wound healing cellular activity, so the *in vivo* corrosion was not predicted. Since metal ion release and other corrosion products formed by pitting and crevice corrosion cause adverse reactions in the patient, this is an area that requires further research.

EXAMPLE PROBLEM 5.6

Old cars in the northern United States are often rusted on the bottom of their doors and on their trunk lids, and the tailgates of old pickup trucks are also often rusted. Name and discuss five reasons for this. Would bolting on a zinc block help?

Solution

(1) The edges and bottoms of doors and lids are formed by bending the metal back on itself. This causes cold working at the bend, which makes it anodic to the rest of the metal. (2) The crimps are then spot-welded closed. This creates an area of different microstructure, which leads to a galvanic situation. (3) The roads are salted in the winter, and the saltwater spray gets caught in the crimp. That is the electrolyte. Sand may also get in the crimped space and help maintain a moist environment. (4) Car manufacturers put a decorative strip of chromium-plated steel along

Continued

the bottom of the lid. The chromium is cathodic to the steel and provides another source of galvanic corrosion. (5) There are potholes in the roads. This causes bouncing of the lid against the frame. This can chip the paint and expose the unprotected metal, or it may be a cause of fretting corrosion.

Bolting on a zinc block would not help except for corrosion of metal in the same electrolyte pool. A piece of zinc in the crevice would help slow the corrosion of the crevice. If the car fell into the ocean, then the zinc block would protect the whole car.

EXAMPLE PROBLEM 5.7

Your grandmother has a stainless steel total hip. Now she needs the other hip replaced, and the doctor wants to use one made of a cobalt chromium alloy. Is that a problem for corrosion?

Solution

No. One of the four essential elements for galvanic corrosion is missing. There is an anode: the stainless steel. There is a cathode: the cobalt alloy. There is an electrolyte: the salt water of the body. However, there is no electrical connection, so there is no problem. If she were to fall and fracture her pelvis, and the break was repaired with an external fixator, then there might be an electrical connection and a problem. However, these alloys are so corrosion resistant and similar electrochemically that there is probably no need to worry.

5.4.4 Bacterial Colonization and Medical Biofilms

Just as microbial biofilms form on devices and hardware used in aquatic ecosystems, wastewater treatment systems, and marine vessels and structures, they also have been observed on most, if not all, medical devices implanted into humans. This includes prosthetic heart valves, orthopedic implants, intravascular catheters, artificial hearts, cardiac pacemakers, urinary catheters, and intrauterine contraceptive devices. Regardless of the technological sophistication of the biomedical implant—for example, a catheter versus a stem cell–containing biomaterial scaffold—all medical devices or tissue engineering constructs are susceptible to microbial colonization and infection.

As already mentioned, immediately upon implantation and contact with blood, the surface of the biomedical implant is coated or "fouled" with a film consisting of proteins and glycoproteins, such as fibronectin, vitronectin, fibrinogen, albumin, and immunoglobulins. This film enables attachment of cells *and also* bacteria through receptors on the surface of the colonizing bacteria. Within minutes of attachment, the bacteria cells start producing bacteria-signaling biomolecules that lead to an orchestrated activity of the bacterial colony, called quorum sensing. The attached bacteria also secrete a mostly insoluble gelatinous polymer substance, thereby forming a three-dimensional cell:polymer matrix known as a biofilm. The bacteria thrive in this protected environment. Biofilms are found throughout the healthy body in places like the lung and intestinal mucus layer, and fortunately, most of the bacteria are not dangerous to the host.

While antibiotics can be effective against freely floating bacteria, microorganisms growing in biofilms are significantly less susceptible to antibiotics and host defenses than their counterparts not growing in a biofilm. Antimicrobial resistance, however, is on the rise, even for freely floating bacteria. Whereas resistant bacteria were previously common only in intensive care units, more recently such organisms (e.g., *Staphylococcus epidermidis*) are found in extended-care facilities and home health care sites. The primary defense the body has against infection from bacteria is the innate immunity provided by immune cells such as neutrophils, macrophages, and dendritic cells. However, the biofilm, which in biomaterials terms is a polysaccharide alginate matrix, protects the entry of the toxic molecules that the immune cells secrete. For example *Pseudomonas aeruginosa* biofilm protects the bacteria from interferon gamma-mediated leukocyte killing. Bacteria such as *Staphylococcus aureus* have developed an ability to thwart neutrophils and macrophages by chemically blocking their migration and making it difficult for the immune cells to phagocytose (swallow) them.

Biomaterials' researchers have attempted to produce anti-infective devices or implants by (1) mechanical design alternatives for indwelling catheters (liquid:air breaks, skin cuffs, antibiotic fills); (2) tethered, covalently attached anti-infective agents, bound directly to the surface of the material (silver coatings, tethered quaternary ammonium, synthetic antibiotics); or (3) the release of soluble toxic agents (chlorhexidine, antibiotics) into the adjacent surroundings. These current approaches have done little to stop the epidemic rise in bacterial infections and may have contributed to the rise of antibiotic-resistant bacteria.

Mechanical design alternatives have had only marginal success. Tethered anti-infective agents are only toxic to the initial wave of incoming bacteria and provide little residual effects once layers of dead cells accumulate. Regardless of the type of drug-release method used, release of a toxic agent from a biomaterial of a soluble anti-infective agent will inevitably stop once the entrapped agent is depleted. Further, delivery of sublethal dosages of antibiotics can lead to accelerated biofilm formation and induced virulence factor expression.

Increasing scientific research in biofilm formation has led to the possibility of several new approaches, including enhancing the phagocytosis of the bacteria by the immune cells by delivering artificial molecules that accelerate and increase the binding of the neutrophils to the bacteria; using antibody therapy, which causes the biofilm to detach; and increasing the ability of the dendritic immune cells to attack bacteria by applying a vaccine to the implanted biomaterial.

5.4.5 Biomaterial Degradation and Resorption

Biomaterials may be permanent or degradable. The degradation process may be chemically driven or accomplished by cells. Bioresorbable implants are designed to degrade gradually over time in the biological environment and be replaced with natural tissues. The goal is to meet the requirements of strength and cell support while the regeneration of tissues is occurring. Small changes in biomaterial chemistry and structure may greatly alter the resorption rate, allowing for materials to be tailored for various applications or leading to unexpected product failure. Collagen and the lactic acid and/or glycolic acid polymers (PLLA and PGA or copolymer PLGA) are the most commonly used for resorbable applications. PLLA and PGA degrade through a process of hydrolytic degradation of the polyester bond. At low molecular weights, the implant can disintegrate and produce small fragments

that elicit an immune response from macrophages. PLLA and PGA degrade in a time period of six months to several years, depending on initial molecular weight and crystallinity. Copolymers of the two typically degrade into fragments in a few months. The lactic acid and glycolic acid fragments are eventually metabolized into carbon dioxide and water. Tricalcium phosphate ceramics degrade through a surface dissolution process into calcium and phosphate salts, which are also present naturally in the body.

Biomaterial degradation may lead to chronic nonhealing wounds that are arrested at one of the normal phases of wound healing. This may happen if a biomaterial degrades too quickly and releases particulate matter that extends the inflammation stage. Persistent inflammation leads to the formation of giant multinucleated cells that continue to attempt to remove the offending material through secretion of acids and enzymes that also destroy healthy tissue. They are the trademark of a foreign body response and may necessitate surgical removal of the implanted device. If the healing passes through to the fibrous capsule formation stage, there may still be complications. For example, a drug delivery implant may eventually no longer function due to impaired drug release by the fibrous encapsulation in response to the degrading drug delivery implant. Fibrous encapsulation of tissue engineering constructs is considered a failure as well, because the host tissue is blocked from integrating with the engineered tissue.

EXAMPLE PROBLEM 5.8

Is it possible to successfully pass through all four wound healing phases only to have the biomaterial degrade and lead to wound healing reversal? Explain your answer.

Solution

Yes. This has happened in some patients with total joint replacements. Total joint replacements such as artificial hips typically consist of two metallic components that meet at a polymeric bearing surface (typically UHMWPE). During bending of the joint, wear debris is produced as the metal surface rubs against the softer polymeric surface. This leads to the recruitment of macrophages that identify the particles as foreign and attempt to remove them. Since the synthetic particles cannot be degraded by cell enzymes, inflammation continues indefinitely. The excessive production of inflammatory cytokines leads to resorption of the newly healed adjacent bone that supports the implant, resulting in implant loosening. Fortunately, there have been improvements in the processing of polyethylene so wear debris is no longer generated at the high rate observed in some of the earliest used hip replacements. Ceramics on ceramic hip joints have been developed that have better wear properties and are not susceptible to corrosion; however, ceramic hip replacements are badly tolerated by elderly osteoporotic bone because the material is very hard.

5.4.6 Immunogenicity

Immunogenicity is the tendency for an object to stimulate the immune response. Common examples of immunogens are proteins in food that lead to allergies or inflammation and pollen from grass or trees. Related to biomaterials, bacteria, small particulates of biomaterials, and cellular debris left behind after processing of allografts are also immunogens.

Basically, our immune system protects us through a combination of physical barriers, such as skin, chemical barriers, such as enzymes and antibodies, and cellular barriers, such as targeted cytotoxic T lymphocytes (T cells). When a biomaterial is implanted in the body, the immune system–associated proteins found in blood immediately attach to the surface, thereby directing the subsequent cell behavior toward the biomaterial. Once again, it must be emphasized that the surface chemistry and structure of the biomaterial play important roles in determining the extent and type of protein attachment and thus the tissue–biomaterial interaction. Proteins of all types will be competing for attachment sites on the biomaterial surface. Depending on the conformation of the attached proteins, a variety of messages may be sent to the nearby cells. Methods for modifying the biomaterial surface to control tissue–biomaterial interactions are discussed later in this chapter.

When allogenic (human) graft biomaterials are implanted in another human, acute rejection can occur if the major histocompatibility complex (MHC) proteins on the cells in the graft are of different types than the donor's MHCs. MHCs are a class of cell-surface molecules that provide information as to what has been identified as foreign in the past to the cytotoxic T cells. With nonmatching MHC groups, the T cells receive two sets of instructions as to what is foreign, and this causes an extremely vigorous immune response. Tissue typing can reduce this type of rejection, although the patient usually still requires long-term medication to suppress some of the activity of the immune system. Rejection can also occur against cell–biomaterial scaffolds in tissue engineering applications. The implanted cells may be recognized as foreign and be damaged directly by the attacking immune cells such as macrophages or be starved to death by the lack of nutrients passing through a thick fibrous capsule created through the inflammatory process to protect the host. Therefore, in some tissue engineering applications, the implanted cells are protected from the immune system by enclosing the cells in selectively permeable biomaterials (e.g., islet cells that produce insulin in alginate hydrogels).

Corrosion of metallic implants releases metal ions that can cause metal sensitivity or allergic reactions in some individuals. Allergic reactions can lead to slow or inadequate bone fusion or skin dermatitis. Both of these conditions usually require removal of the implant. Once again, this demonstrates that biomaterials are not inert.

Biomaterials are sometimes deliberately designed to enhance the immune system's response. For example, vaccines are typically given with a particulate biomaterial known as an adjuvant for enhanced and longer-lasting immunity. Vaccine adjuvants have their own immunogenic properties, resulting in a stronger local stimulus to the immune system. Biomaterial adjuvants can be simple particles that adsorb the weak immunogen, increasing the effective size of the weak immunogen and enhancing phagocytosis of the particle by macrophages. Biomaterial adjuvants also work like controlled-release vehicles by prolonging the local retention of a weak immunogen and increasing the chance of a local immune response. Vaccine adjuvant selection represents a compromise between a requirement for adjuvanicity and an acceptable low level of side effects. The FDA has approved only three materials for human use (all of which are mineral salts): aluminum phosphate, aluminum hydroxide, and calcium phosphate. Aluminum compounds are often incorrectly identified in the scientific literature as alum. Alum is potassium aluminum sulfate that is used as the starting solution to precipitate antigens with either aluminum phosphate or aluminum hydroxide. Other biomaterial adjuvants used in research include oil emulsions, lipopolysaccharide products from bacteria (LPS), and their synthetic derivatives (liposomes).

5.5 BIOMATERIALS PROCESSING TECHNIQUES FOR GUIDING TISSUE REPAIR AND REGENERATION

5.5.1 Methods to Modify Surface Chemistry (1-D)

The interaction of cells and tissues with biomaterial surfaces is critically important to promote new tissue deposition and for healthy integration with the surrounding extracellular matrix. In nature, the surface chemistry of every cell and the extracellular matrix is carefully controlled to obtain the desired response. Cell–to–extracellular matrix interactions are dominated by the interaction of cell-adhesion proteins (ligands) bound to the extracellular matrix and cell surface receptors (primarily integrins). In medical device applications, it is important to understand that the biomaterial appears as an extracellular matrix to the cells and sends signals to the cells interacting with it through similar mechanisms (Figure 5.10). Thus, the biomaterial surface plays a very important role in determining tissue–biomaterial interactions, and this concept has governed the development of many new surface modification techniques.

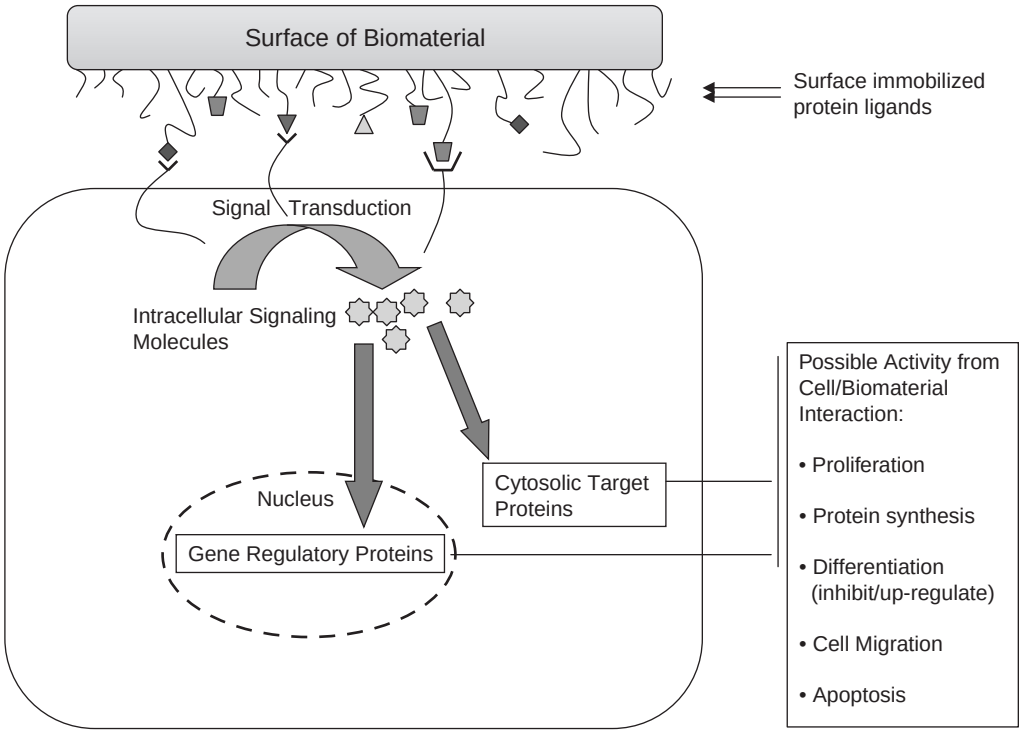


FIGURE 5.10 Like the extracellular matrix, a protein-covered biomaterial sends signals to the cell interacting with it through ligand-receptor mechanisms. Primary cell signal transduction is facilitated through multiple pathways, leading to the synthesis of various intracellular signaling molecules. Acting both on the genetic regulatory proteins in the nucleus and other cytosolic target proteins, the signaling molecules can induce various phenotypic expressions. Ideally, the biomaterials scientist can engineer the proper surface treatment to elicit the desired cellular activity.

One biomimetic surface modification approach to controlling cell–biomaterial surface reactions is to preadsorb proteins on the implant surface that mimic those most involved with cell adhesion. The three-amino-acid sequence of Arg-Gly-Asp (RGD) found in fibronectin is now well known for mediating adhesion of cells to surfaces; therefore, RGD-containing peptides are now being deposited on surfaces to promote cell attachment. Adsorption of other biological molecules such as growth factors to the surfaces of implants can control the tissue–biomaterial interaction and lead to enhanced cell activity and more differentiation than will the cell adhesion molecules alone.

There are many chemical reactions that can be used to attach a biomimetic peptide sequence to a biomaterial. For example, a protein can be immobilized on a surface through a technique known as organosilane chemistry (Figure 5.11a). The details of the chemical coupling and derivitization processes are beyond the scope of this text. Basically, there are coupling agents such as silanes used to create a covalent bond between the biomaterial surface and the protein to be attached. Well-ordered protein attachment results. A wide variety of solid surface modification techniques are available to create the reactive coupling groups, such as photochemical grafting, chemical derivation, and plasma gas discharge. Physical adsorption methods utilizing other types of bonding, such as van der Waals and electrostatic binding, can also be used to immobilize proteins (Figure 5.11b). Physical and electrostatic adsorption is the easiest technique, but it is the least specific and tends to readily release the adsorbed molecule.

Lipid groups and dye molecules can also be used to immobilize proteins on surfaces. A critical component of surface modification is the resulting ligand density. If protein adsorption is too low, the addition of more functional groups to a relatively inert polymer can be accomplished by plasma glow discharge treatment. The greater reactivity of the surfaces with higher surface energy after plasma treatment generally leads to increased tissue adhesion.

Surface modification also can be used to produce protein-resistant surfaces that are needed in blood-contacting applications such as vascular grafts. For example, polyethylene oxide has been attached to surfaces to reduce protein adsorption. Cell adhesion was significantly reduced on these treated surfaces. Anticoagulants can also be attached to biomaterial surfaces to decrease unwanted cell attachment. Various hydrophilic biomaterials have been shown to reduce platelet adhesion and thrombus formation. Hydrophilic materials have also been shown to hinder bone healing, so what is appropriate for one biomaterial application does not necessarily apply to another.

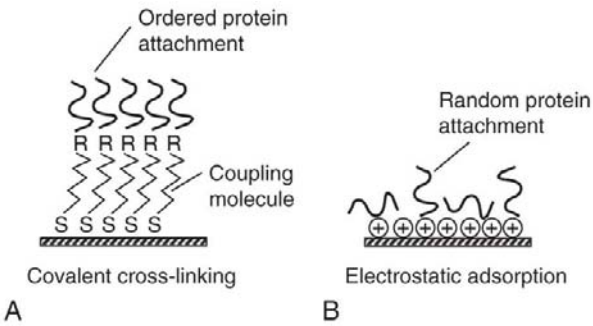


FIGURE 5.11 Two types of chemical reactions used to modify a biomaterial surface: (a) covalent coupling techniques where S = the silane/oxygen bond with the surface and R = the reactive group that attaches to the protein, and (b) physical adsorption methods utilizing electrostatic interactions. Negatively charged proteins are shown attaching to a positively charged biomaterial surface.

Hyaluronan is a biomolecule found in the cartilage extracellular matrix. It is responsible for tethering a proteoglycan (aggrecan) to the collagen matrix. Studies have shown that it can guide the differentiation of mesenchymal stem cells to cartilage chondrocytes. In those experiments, hyaluronan was chemically bound to tissue culture dishes, and undifferentiated cells were added. It was found that a specific molecular size (200,000–400,000 daltons) was optimal to initiate cartilage formation. Biomolecules such as enzymes, antibodies, antigens, lipids, cell surface receptors, nucleic acids, DNA, antibiotics, and anticancer agents can all be immobilized on or within polymeric, ceramic, or metal surfaces.

Surface deposition of calcium ions or the use of calcium-containing biomaterials strongly influences the attachment of bone cells. Hydroxyapatite-coated hip implants show decreased fibrous tissue formation and increased direct bone bonding. Better bone attachment has also been found for hydroxyapatite-coated dental implants and spinal fusion cages. Recent studies have shown that hydroxyapatite ceramics are selective in cell recruitment from the bone marrow. This may be due to an intermediate step of selective protein adsorption.

The surface modification can be applied in a pattern with a particular feature size, shape, and periodicity, thus leading to a two-dimensional feature. Patterning techniques will be discussed in the next section on surface topography and roughness (2-D).

5.5.2 Surface Topography and Roughness (2-D)

Surface topography refers to chemical discontinuities and physical discontinuities on the surface. Topography and roughness on the surface of a biomaterial have been shown to provide cues to cells that elicit a large range of cellular responses, including control of adhesion, cell morphology, apoptosis (cell death), and gene regulation. Modification of the texture of the biomaterial surface can therefore have dramatic effects on guiding tissue growth.

Photolithographic techniques can be utilized to micropattern a surface with proteins, molecules, or functional groups (Figure 5.12). This technique involves using a photoresist

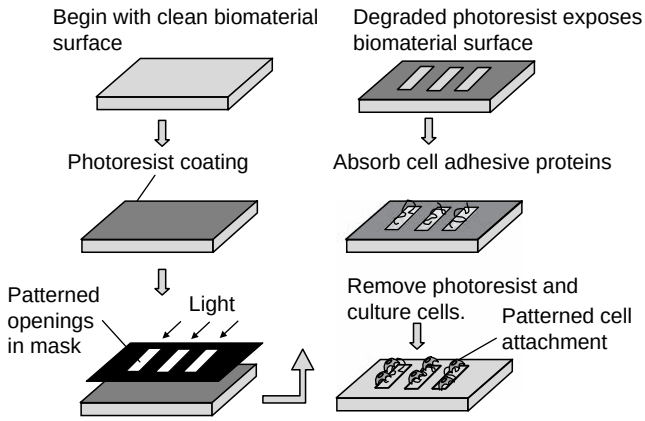


FIGURE 5.12 Photolithography techniques for micropatterning a biomaterial surface with immobilized proteins or functional molecules involve the use of standard semiconductor industry techniques applied to biomaterials.

layer in which patterns can be created by selectively exposing certain areas to light. The light degrades the exposed portions of photoresist, leaving a bare biomaterial surface. Proteins or molecules can then be selectively attached to these exposed areas. The remaining photoresist is then removed to obtain a biomaterial surface that has protein or molecular patterning. This type of photopatterning of a protein results in a substrate with specific areas for enhanced cell attachment. Using this technique, researchers were able to see that cells responded differently to being cultured on small, round-shaped areas rather than larger, square areas. When cultured in medium that supported fat cell differentiation, only the cells at a high density on the small circular areas became fat cells, while those cells allowed to spread out on the larger square areas became bone cells. Restricting or enhancing cell attachment can thus govern cell differentiation.

Instead of modifying the exposed areas with proteins, the material that is exposed can be deeply etched, leading to a pattern of grooves or pits. The directed activity of cells by a groove in a biomaterial is known as contact guidance. When cells encounter grooves or pits, they change their shape and align or elongate along topographic features. This is a technique that can lead to ordered alignment of cells on a surface, which is necessary if the goal is to mimic the normal alignment of cells in muscle, blood vessels, and nerves. The distortion of the cytoskeleton within the cell imparts biological messages to the cell nucleus, leading to changes in gene expression.

There are many methods for modifying the surface of biomaterials, and new technology continues to be developed. It has been observed that the topography or roughness must be on a biologically relevant scale (1–10 μm) to affect cell growth and attachment. Other methods for modifying surface topography that provide this type of resolution include surface roughening by laser ablation or wet etching with a corrosive solvent, or molding or casting a biomaterial in melt form into a rigid mold. Hot embossing or imprint lithography forms a relief replica of a mold by pressing it into a thermoplastic material. Microcontact printing is the most often used method to create chemical patterns for cell substrates. An elastomeric stamp is used to place a biologically active compound, such as a protein, onto a substrate through contact transfer. These techniques are not only for polymeric substrates. Metals can be roughened as well through sandblasting with ceramic particles or chemical etching by being placed in an alkaline or an acidic solution that erodes its surface, leaving pits of specific diameter and shape.

Since many metallic components are used in the orthopedic field and it is known that calcium phosphate coatings enhance bone attachment, there are many techniques for coating titanium with calcium phosphate. Ion sputtering is a process by which a thin layer of calcium phosphate is transferred over to the titanium by directing an ion beam into a block of calcium phosphate and vaporizing it to create a plasma that is recondensed on the metal implant. Plasma spraying is a similar technique and is used commercially. Biomimetic deposition through immersion of the metal part in a highly saturated solution of calcium and phosphate, leading to deposition on the metal, is a way to make a calcium phosphate coating with the most similar structure to actual bone mineral crystals.

Smooth surfaces such as pyrolitic carbon resist protein and cell attachment are ideal for heart valve applications. Bioprostheses made of bovine or porcine heart valves are even more superior for valve applications and reduce coagulation and embolism by a combination of an ideal surface topography and surface chemistry. However, they typically fail due

to calcification, which is enhanced on the natural collagen surface as compared to the smooth pyrolitic carbon. As should be now evident, the biomaterial scientist has a similar job to a chef in that many different dishes can be created from the same basic ingredients; it just depends on what components are added and how they are "cooked." Many different types of specialized equipment are used in this field.

5.5.3 Biomaterial Scaffold Fabrication Techniques (3-D)

The goal of a scaffold is to recreate important aspects of the cell microenvironment that will allow cell proliferation, migration, differentiation, and synthesis of a new extracellular matrix. When this combination of events occurs, new tissue can be regenerated. This is important clinically for attaching/integrating host tissue into a synthetic body part replacement, such as a total hip replacement implant, or for regenerating new tissue where none remains. If a large enough wound has been created that cannot be filled with nature's scaffold—a blood clot—then it is necessary to use a synthetic scaffold. Synthetically produced biomaterial scaffolds are produced commercially and are made of ceramics and/or polymers with or without a tissue-stimulating biological molecule and with or without cells. Combining growth factors, cells, and biomaterial scaffolds is the premise behind tissue engineering (see Chapter 6).

One of the most critical elements of the scaffold biomaterial is that it mimics the extracellular matrix that normally serves to maintain space, support cells, and organize cells into tissues. The section on surface chemistry in this chapter gave an example of how mimicking components of the adhesive proteins of the ECM can enhance cell attachment and differentiation. This section focuses on the three-dimentional structural and physical characteristics of the extracellular matrix scaffold that appear to be critical to imitate in synthetic scaffolds in order to stimulate cells and lead to the functional regeneration of tissues.

Pore size is a very important parameter of biomaterial scaffolds used for tissue regeneration. Through trial and error, optimal ranges of pore sizes have been determined for different tissues and for different biomaterials. Some rules of thumb are used, such as the pores must be at least 5–10 μm for a cell to fit through. Successful bone scaffolds typically have pores that traverse the full thickness of the scaffold and are 100–250 μm in size. New blood vessel formation, or neovascularization, has been shown to require pores within polymer scaffolds that are between 0.8 and 8 μm and to not be possible within polymers with pores less than 0.02 μm. Typically, the acceptable pore size in polymers is smaller than in ceramics or metals, perhaps due to pore size expansion, which can occur in the body due to degradation or swelling of the polymer.

The pore size determines many aspects of the scaffold, such as mechanical strength; degradation rate; permeability to gases, fluids, and nutrients; and extent of cell ingrowth. Interconnected porosity is essential for tissue engineering applications requiring nutrient diffusion and tissue ingrowth. Early total joint replacements had smooth surfaces, but today rough, porous coatings or grooved surfaces are used on hip implants to achieve bony ingrowth. The hip and knee implants in Figures 5.2 and 5.3 have rough, porous coatings. The porous coatings on orthopedic implants are achieved by partial fusion of small metallic spheres to the implant surface. The interparticle spaces are the pores. The pores, as well as grooves, seem to encourage bone cells to migrate into or along them. It has been observed

that the stretching out and aligning of the cells along a surface feature (greater than 5 μm) causes them to initiate bone deposition. This leads to a mechanical interlocking between the bone tissue and the implant that increases the bond strength.

There are many techniques for producing porous biomaterials for tissue engineering, and the field continues to develop rapidly. One early approach to creating pores within a biomaterial for use as a cell scaffold is to dissolve the polymer in a solvent and mix in particulate materials that are stable in the solvent but can be dissolved later (porogens). The solvent is allowed to evaporate, leaving the polymer strands tightly encasing the porogens. The porogens are then dissolved by washing them out with a different solvent that will not dissolve the polymer, such as water, to form pores. This is known as solvent casting/particulate leaching. Related to this technique is coascervation, in which the polymer is dissolved in a solvent, the porogen is added, and then a nonsolvent for the polymer is added. A polymer precipitate forms, entrapping the porogen particulates. The precipitate can be collected and then compacted prior to removal of the porogen. After washing to remove the porogen, a porous structure is revealed, as shown in the composite bone crystal/polyhydroxybutyrate hydroxyvalerate (PHBHV) polymer in Figure 5.13. Organic solvents are used with these procedures and must be thoroughly removed under vacuum or low heat to avoid killing the cells.

Phase separation/emulsification can also be used to fabricate porous polymer scaffolds. In this technique a polymer is dissolved in a solvent, and then an immiscible liquid (such as water) is added and mixed to form an emulsion. The polymer/water mixture is cast into a mold, rapidly frozen, and then freeze-dried, which is known as lyophilization. The space that was water becomes a pore, and the solvent is evaporated, leaving behind a solid structure. Scaffolds with high porosity (up to 95 percent) have been formed by this method, but the small pore sizes (13–35 μm) are a drawback. Fiber bonding methods in which preformed fibers are layered or woven and then hot-melted or glued together by solvent exposure is another technique for forming porous materials. The advantage is that the pore sizes are controllable and interconnected. The drawback is that the pore channels are rectangular

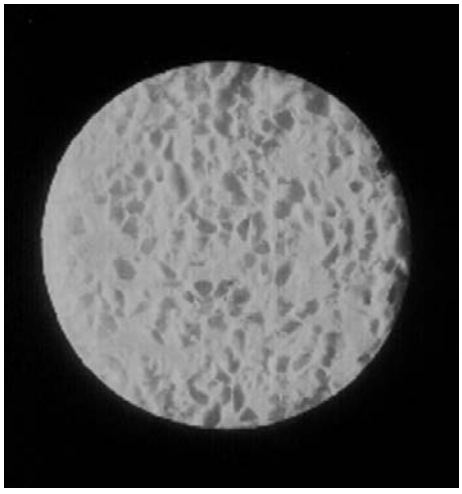


FIGURE 5.13 A composite biomaterial scaffold made of bone crystals and the resorbable polyhydroxybutyrate hydroxyvalerate (PHBHV) copolymer. The coascervation technique was used to precipitate the polymer from solution to produce this scaffold. Dissolvable porogens were used to create the large pores needed for bone repair applications. *(The scaffold was made by Marianne Manot, a former student of Dr. Kuhn's.)*

and regular, unlike natural extracellular matrix structure. Thus, alternatively, microspheres can be preformed from polymers and then sintered together to form three-dimensional scaffolds, creating a more random porosity (Figure 5.14).

Solid freeform fabrication is also used to make biomaterial scaffolds (in addition to being used for rapid prototyping of automotive or aerospace industry parts). Some techniques fall under the umbrella of "solid freeform fabrication." Stereolithography is one technique that uses a focused laser that follows a pattern dictated by computer-assisted design drawing to selectively cure only certain areas within a thin layer of liquid polymer. The depth of the liquid is raised around the part being fabricated, and the laser is again sent on a computer-assisted path to form the next layer of the component. This is repeated over and over again to produce a complex three-dimensional shape. Similarly, a technique known as 3-D printing can be used to create porous scaffolds or complex shapes layer by layer, this time using a print head to deposit "glue" over computer-specified areas of a powder bed. After all of the layers have been "printed," the final part is picked up, and the unbonded particles fall away, revealing the three-dimensional component. Pieces of replacement bone have been made using these techniques with computer assisted tomography (CAT) scans of x-ray images. The features of the scaffolds are limited to 10–1,000 μm with these techniques, which have been used to prepare polymer and ceramic scaffolds.

A scaffold fabrication technique that does result in a nanofibrous polymeric structure similar to the extracellular matrix is electrospinning. Electrospinning is the same technique used to make cotton candy. A molten polymer solution is drawn out from a nozzle by an

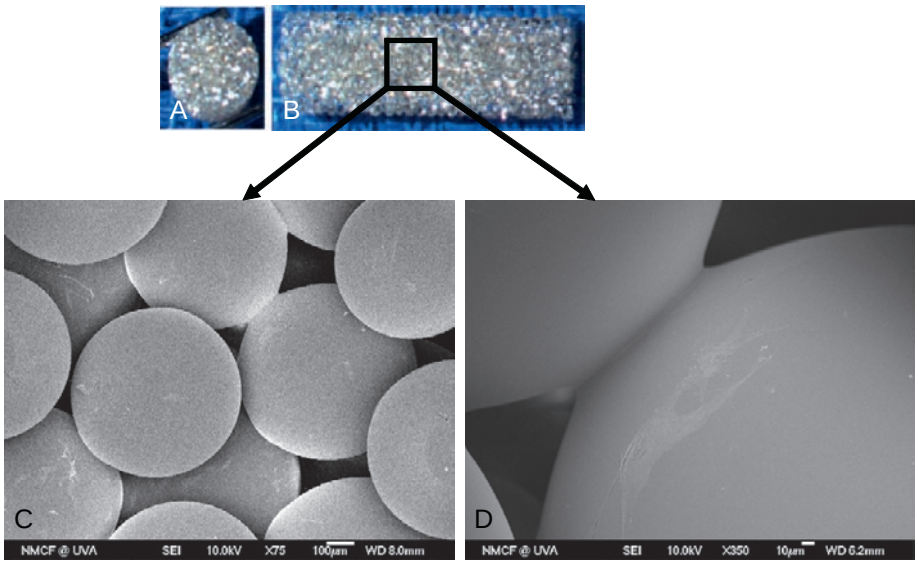


FIGURE 5.14 An example of a three-dimensional heat-sintered polylactide-co-glycolide microsphere scaffold. These 100 percent interconnected porous scaffolds offer mechanical properties in the midrange of human trabecular bone with 30–35 percent porosity and an average pore diameter of 100–150 μm. Optical micrographs (a) represent the top view and (b) the side view of the scaffold. SEM micrographs (c) and (d) show macroporous structures and bonding between adjacent microspheres. *Images courtesy of the Cato T. Laurencin Laboratories, University of Connecticut Health Center.*

electrical charge and builds up on a grounded collector. Electrospinning technology has become popular for the fabrication of tissue engineering scaffolds in recent years because of the growing interest in nanotechnology and the unique properties and relative ease of fabricating scaffolds using this process. Typically, an electrospun nanofiber matrix shows a porosity of more than 90 percent and a pore diameter range of up to 100 mm. Cells migrate through the matrix by pushing the surrounding fibers aside, since nanoscale fibers offer very little resistance to amoeboid movement of the cell. At least more than 100 different biodegradable polymers have been successfully electrospun, and over 30 of them have been used for a variety of tissue-engineered applications. Among the natural polymers, collagen, gelatin, elastin, silk fibroin, fibrinogen, hyaluronan, and chitosan have recently been fabricated into three-dimensional, nanofibrous scaffolds for orthopedic applications. Fully synthetic polymers have also been electrospun (Figure 5.15). The biomimetic nanostructure

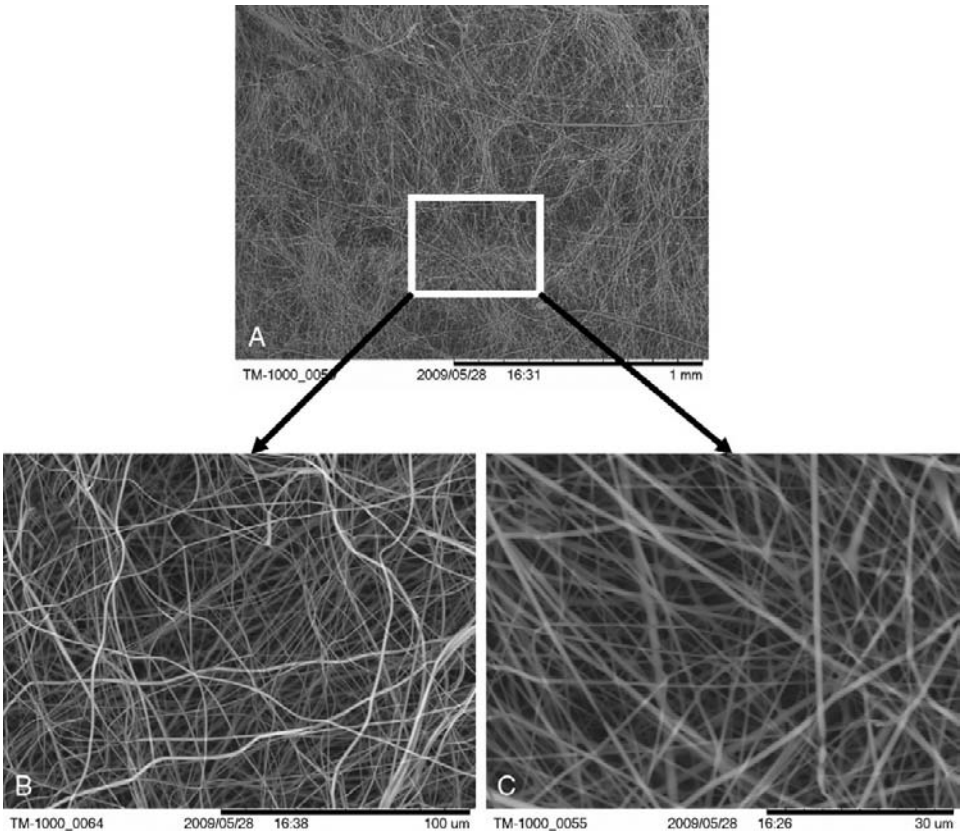


FIGURE 5.15 An example of electrospun polylactide-co-glycolide nanofiber matrices for engineering various tissues. Electrospun nanofiber matrices closely mimic the structure of a native extracellular matrix and have been found to be advantageous for a variety of soft tissue repairs and regeneration applications. SEM micrographs at various magnifications show highly porous structures with fibers in size ranges of 900–1200 μm. *Images courtesy of the Cato T. Laurencin Laboratories, University of Connecticut Health Center.*

results in enhanced cellular attachment and cellular function, and has motivated increased research on nanostructured biomaterials.

Instead of making porous scaffolds in which the cells are seeded onto the surfaces, another approach in tissue engineering for cell or growth factor delivery is hydrogel scaffolds that enclose the cells while allowing fluid and gas transport. Hydrogels can be formed from a vast array of natural and synthetic materials. Hydrogels made of natural materials such as collagen, hyaluronan, alginate, and chitosan are inherently biocompatible and bioactive. They may promote many cellular functions due to the myriad of growth factors and cytokines present in the natural materials. However, as with any natural material, the drawback is lot-to-lot variability, which makes it difficult to study biological responses to small chemistry modifications, so fully synthetic molecules such as polyethylene glycol hydrogels are better suited for fundamental studies of cell physiology.

The variables within a hydrogel that can affect the cell viability and function include the type of cross-linker and the amount of cross-linker, which further define the mechanical properties, and transport and degradation kinetics. In vivo, the extracellular matrix provides a milieu of binding ligands for cell adhesion. These ligands encourage cell receptor–ligand events that communicate the mechanics of the extracellular matrix to the cell and direct cell fate through intracellular signaling pathways. To accomplish this within synthetic hydrogels, small amounts of extracellular matrix proteins such as collagen, laminin, or fibronectin are incorporated into the network.

5.6 SAFETY TESTING AND REGULATION OF BIOMATERIALS

5.6.1 Product Characterization

Ensuring product purity and identity is one of the first steps in developing a safe product. There are extensive data documenting the safety of various biomaterials, but since processing methods may include additives and the final sterilization step may alter the biomaterial, it is crucial to always verify the end product purity and identity. The American Society for Testing and Materials (ASTM, www.astm.org) has developed many standards that manufacturers of medical device products can use as guidelines to evaluate product purity and identity, as well as safety. Under ASTM specifications for metals used in medical devices, there are restrictions on the composition, microstructure, phase and grain size, inclusion size, defect size, and macro- and microporosity to help ensure safety. There are ASTM standards for ceramic materials that specify chemical composition, phase determination, grain size, and impurities such as sintering aids, which may decrease fatigue resistance. New standards are being written by ASTM, which has created a new division to help ensure quality and reliability of tissue engineering scaffold materials. In addition to specifying the correct tests and techniques to determine the chemical identity of the tissue engineering scaffold materials, standardized tests for measuring porosity and permeability have also been developed. ASTM also has standards for dissolution testing, degradation testing, and stability testing. ASTM standards are written through a consensus process and represent the best available knowledge from a wide cross section of manufacturers,

users, and general interest groups. These standards are available electronically and are a valuable resource for both the new student and the experienced engineer.

Although many techniques are available for evaluating the composition and structure of biomaterials, no single technique is capable of providing all of the needed information. Thorough characterization requires the use of multiple analytical methods. For example, if the material is crystalline, x-ray diffraction can be used for bulk product identification (typically, powdered samples are analyzed). The x-ray beam is diffracted as it goes through the rows of atoms, and a detector measures the reflected intensities as a function of angle from the surface. The intensities at various angles provide a unique signature of the material structure and can be compared to existing data files for product identification purposes.

Fourier transform infrared spectroscopy is a method that complements the structural information gained by x-ray diffraction by providing information about the chemical groups found within the structure. The material does not need to be crystalline and can be gas, liquid, or solid. The sample is exposed to infrared radiation, and the molecular vibrations induced by the radiation are observed. Radiation at frequencies matching the fundamental modes of vibration is absorbed, causing oscillating dipoles perpendicular to the surface. These are detected as a function of wavelength and provide a chemical "fingerprint" that can be compared to existing databases to identify the material. Most characterization methods capable of identifying an unknown substance involve bombarding the material with some type of energy, quantitating the interaction with the material, and then searching a database for similar results. Other techniques utilizing this basic procedure include secondary ion mass spectroscopy (SIMS) and x-ray photoelectron spectroscopy (XPS), both of which are well suited for identifying the surface chemistry of a biomaterial.

Scanning electron microscopy (SEM) is very useful for characterizing the two-dimensional surface topography of a biomaterial. In SEM, a beam of high-energy electrons is scanned across the sample, causing the material to emit secondary electrons. The intensity of the secondary electrons primarily depends on the topography of the surface. An image can be recreated by recording the intensity of the current generated from the secondary electrons. The resolution of an SEM allows magnifications of up to 100,000×. When greater resolution of a material surface is needed, atomic force microscopy (AFM) can be used. In AFM, an atomically sharp tip attached to a cantilever is dragged across the surface of a material but actually does not touch the material. The interactions of the atoms of the material being analyzed with the tip cause either repulsion or attraction. The height adjustments or changes in interatomic forces are recorded and used to construct images of the surface topography. Under proper conditions, images showing individual atoms can be obtained.

Images in three dimensions can be obtained using computer aided x-ray tomography (micro-CT) or nuclear magnetic resonance (NMR) imaging. An image obtained by micro-CT of a biodegradable porous scaffold is shown in Figure 5.16. In both techniques, the samples are scanned in all directions, and then the image is created mathematically by a merging of all the directional information. Solid or porosity volume or volume fraction can be measured nondestructively. Direct measurement of solid or pore characteristic dimensions (width, diameter, thickness) and spacing (or period) of repeating structure can be made. In micro-CT, image contrast is achieved via attenuation of x-radiation; thus,

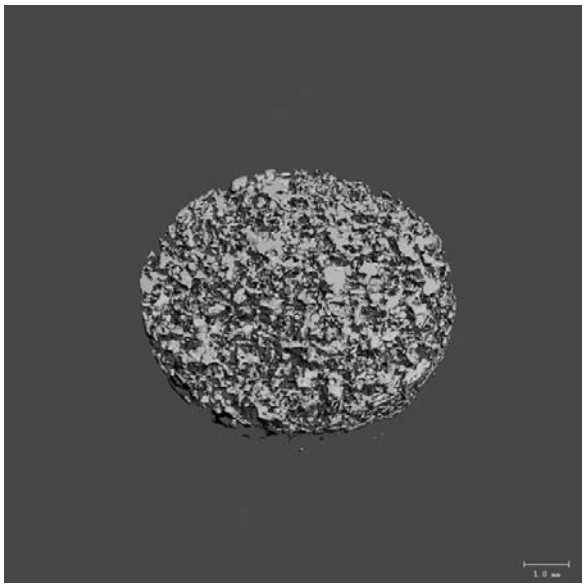


FIGURE 5.16 The scaffold from Figure 5.13 imaged by computer-aided x-ray tomography (micro-CT). *Courtesy of Douglas J. Adams, PhD, Micro-CT Facility, University of Connecticut Health Center.*

polymers (or other scaffold substrates) with relatively higher attenuation coefficients will provide higher-contrast images and improved computational segmentation of scaffolds versus the background. Commercial desktop micro-CT instruments are available with a spatial resolution of approximately 5 μm.

5.6.2 Methods for Testing and Evaluating Safety and Biocompatibility

It has been emphasized repeatedly in this chapter that tissue–biomaterial interactions are a two-way street: implantation in a body affects the biomaterial, and the biomaterial affects the body. Therefore, both of these aspects must be investigated prior to biomaterial implantation in the human body. This is known as biocompatibility and safety testing. Before initiating biocompatibility and safety testing, the structure and chemistry of the material should be fully characterized by a combination of techniques, as described earlier in this chapter. This is necessary to confirm the purity and identity of the biomaterial and to ensure that no unintended foreign substances have been introduced during the synthesis, manufacturing, and sterilization procedures.

Biocompatibility and safety tests include in vitro assays (using cells and tissues), in vivo models (in animals), and, finally, human clinical trials. Several guidelines and procedures have been developed by the standards organizations of the world (ASTM and the International Standards Organization [ISO]) and federal regulatory agencies (e.g., Food and Drug Administration). ASTM Standard F-748 and ISO 10993 provide detailed methods for completing adequate safety and biocompatibility testing and are followed by all implant and medical device manufacturers. The tests are separated into various categories based on intended use. For example, there is a matrix of tests appropriate for surface devices, external communicating devices, or implanted devices. The recommended tests are further

categorized based on contact duration (short term, prolonged contact, permanent). The basic tests include cytotoxicity, sensitization, irritation or intracutaneous reactivity, acute systemic toxicity, subacute systemic toxicity, genotoxicity, implantation, hemocompatibility, chronic toxicity, and carcinogenicity. The preferred test sample is the intact medical device that has been processed and sterilized in the same manner as the medical device that will be used in humans. However, it is not always practical to use the intact medical device due to the constraints of the biological tests. Therefore, an extract of the leachable components or the degradation components of the implant are often tested first in the in vitro assays (cytotoxicity) and also in the preliminary in vivo tests (sensitization and both systemic toxicity tests). Completion of the preclinical tests described in the standards ASTM F-748 and ISO 10993 typically takes up to two years, even if a qualified and experienced facility is conducting the testing.

It is difficult to correlate in vitro testing to in vivo testing because the in vivo system is much more complex and involves many more variables. Typically, cell culture assays are more sensitive than in vivo tests; however, demonstration of cytotoxicity in vitro may not necessarily mean that the material cannot be used in vivo. Both false negatives and false positives can be obtained by cell culture testing; therefore, animal testing is a required step in understanding safety and biocompatibility and also for an initial evaluation of the product's performance and effectiveness. There are, however, variations in response to biomaterials and drugs between species of animals. The guinea pig has been found to be the most sensitive animal for assessing delayed immune hypersensitivity (the sensitization test). The rabbit has been found to be the most sensitive animal model for detecting pyrogens in vivo. Although animal testing does provide a useful screen for restricting the implantation of most toxic components in humans, the final and ultimate biocompatibility and safety testing occurs during human clinical trials. In some cases, products that demonstrate efficacy in a mouse or dog model may not always perform as well in humans, particularly in the case of a new drug. The effective dose sometimes varies greatly between species, as well as between two humans. Therefore, dose escalation schemes are incorporated into biocompatibility and safety and efficacy testing, and large numbers of patients must be used in clinical trials.

5.6.3 The Regulatory Process

Regulatory approval by the Food and Drug Administration (FDA) is required in the United States prior to administering a new drug or biologic or implanting a new medical device in a human and also prior to marketing the new product. The FDA is currently divided into six individual centers that regulate devices and radiological health (CDRH), drugs (CDER), biologics (CBER), food and cosmetics, veterinary medicine, and toxicology. An assessment must be made as to which mode of action—drug, device, or biological—contributes the most to the therapeutic benefits of the overall product. Based on this criterion, the FDA decides which center will take the lead on the regulatory review. Each FDA center has different procedures and requirements that must be completed and met to gain FDA approval. Some products are a combination of a biologic and a device or a drug and a device, and, therefore, FDA requirements for two or more centers must be met before the product is granted approval. Tissue-engineered medical products are examples of combination products.

Combination products are typically the most difficult to regulate and take the longest to reach the market. Biomaterial scaffolds alone without cells or growth factors are typically regulated by the FDA as a device and are under CDRH jurisdiction.

The regulatory procedures are sufficiently complicated that typically a specialist in this area is hired to manage this aspect of new product development. However, it is important to have some idea of product regulation early on, since it, along with good research and development, is needed to bring a medical product to clinical success. Toward that purpose, the procedures utilized by the CDRH branch of the FDA will now be described.

The CDRH branch of the FDA utilizes classification of medical devices to assist with determining the requirements for approval and the extent of regulatory control (http://www.fda.gov/cdrh/devadvice/). Medical devices are placed into one of three classes. Class I devices are those that have limited body contact and essentially pose no significant risk. Class II devices require special controls and must usually meet some performance standards to provide some assurance of safety. A device is placed in Class III if there is insufficient information to determine that general or special controls are sufficient to provide reasonable assurance of its safety and effectiveness. A new device that is not substantially equivalent to a device on the market will automatically be placed in Class III. Examples of Class I medical devices include dental floss, a tongue depressor, a surgeon's glove, and a clinical chemistry test system such as a pregnancy test. Examples of Class II medical devices include a blood pressure cuff, an oxygen mask, dental impression material, and an electrocardiograph. Examples of Class III medical devices include a heart valve, an automated blood warming device, and a silicone inflatable or gel-filled breast prosthesis. Classification is an important step of the FDA approval process, since it will determine the extent of the testing required prior to use in humans and when the device can be sold.

At least 90 days prior to commercially distributing a new or substantially modified device, a manufacturer must submit a premarket notification to the FDA. More than 99 percent of the applications received by CDRH are cleared for marketing through the 510(k) Premarket Notification process. The goal of this process is to demonstrate to the FDA that the new device is substantially equivalent to an already approved predicate device. This is accomplished through careful characterization by several complementary methods that confirm the identity and purity of the substances involved, followed by completion of an abbreviated form of the ASTM F748 and/or ISO 10993 test protocols and adherence to quality system regulations. Quality system regulations include compiling a Device Master Record and Design History File. Together these two files contain documentation of the procurement process, the manufacturing details, all testing results—including assay verification tests—and the details of the design rationale and design verification testing.

If a device does not qualify for 510(k) approval, then a full premarket application (PMA) must be submitted containing all the required information on the safety and the effectiveness of the device as determined through preclinical and clinical testing. There is a decision tree in ISO 10993 that helps define which biocompatibility and safety testing is necessary based on length of contact with the body and/or blood. Clinical trials are highly regulated so human subjects are not exposed to significant risks without their knowledge. Carefully documented and successfully completed in vitro and in vivo animal safety testing is required at the time of application to begin a clinical trial. Human clinical trials typically are divided into four phases: safety testing, efficacy testing, blinded efficacy compared to

a clinically acceptable alternative (these three take approximately five years to complete), and finally, postmarket surveillance (gathered from product use by the general public after FDA approval). Prior to initiating a clinical trial, a manufacturer must obtain an investigational device exemption (IDE) that includes all manufacturing and quality control procedures, the plan for the clinical study, and the lists of the review boards that have reviewed the proposed plan (see Chapter 2). The FDA has 30 days to approve or disapprove the IDE. If the PMA application is considered complete, the FDA has 180 days to approve or disapprove the application. If approved, the product can be marketed for human use for the purposes declared in the application, which are to be described in the product labeling. Another FDA submission for review is required prior to legally marketing the product for a new use (off-label use). For those interested in further information on regulation by CDRH or any of the other FDA centers, FDA guidance documents that cover all aspects of regulatory approval are readily available online (www.fda.gov).

5.7 APPLICATION-SPECIFIC STRATEGIES FOR THE DESIGN AND SELECTION OF BIOMATERIALS

5.7.1 Musculoskeletal Repair

The design and selection of the biomaterials components for any implant should be based on restoring the biological function of the damaged or diseased tissue. The principal function of musculoskeletal tissues is to provide a framework to support the organs and to provide a means of locomotion. Bone, cartilage, tendons, ligaments, and muscles are all part of the group of musculoskeletal tissues; however, they have different functions and different biological properties. Each musculoskeletal tissue must be considered individually in terms of implant design and biomaterials selection.

Bone is the only tissue capable of undergoing spontaneous regeneration. It is constantly in a state of remodeling, always optimizing its structure to best meet the needs of the body. This ongoing cellular activity is why astronauts rapidly lose bone mass during zero gravity conditions and why jaw bone rapidly resorbs when teeth are pulled. Unlike bone, cartilage is acellular and has a very limited capacity for repair. Therefore, damage to cartilage is often permanent and often progressive. Cartilage provides an articulating surface enabling low-friction movements between opposing bone surfaces. Ligaments may appear on the surface to be easier to replace due to their "simple" function, but they are not simply passive joint restraints; they also provide electromechanical signals for joint-stabilizing muscle contractions.

Replacement of damaged or diseased tissues or organs is best accomplished by autograft or allograft donor tissue, but they have limited availability, and a biomimetic synthetic substitute is the next best alternative. Biomimetic calcium phosphate materials (hydroxyapatite) have been shown clearly to enhance bone cell activity and are either used alone or in combination with collagen or other polymers as bone graft substitutes. Hydroxyapatite not only influences bone cell attachment, but it also appears to control the differentiation of stem cells to bone-forming cells. This is particularly important for tissue engineering approaches that aim to not only restore function but also to restore the actual biological

tissue. In tissue engineering product development, materials selection is more complicated than it is in traditional approaches; typically, a biomaterial must have sufficient mechanical strength and also have a surface chemistry conducive to cell attachment and proliferation, must also perhaps serve as a drug delivery vehicle and release growth factors, and must also resorb or biodegrade once the new tissue has been formed.

Autografting a mixture of bone and cartilage (e.g., osteochondral) tissue from a healthy portion of a joint to the defect site is a current strategy for articular cartilage repair. Another approach involves filling the bulk of the defect with cells that can facilitate the growth of appropriate cartilage and/or bone tissue. This approach recently has been approved by the FDA and utilizes chondrocytes from the patients themselves. The chondrocytes are harvested, expanded in vitro, and then injected into a surgically created compartment over the defect site. The cartilaginous tissue created by this process has been found to degrade faster than cartilage formed during fetal development. Perhaps a vital cell signaling message has been left out during this artificially stimulated biological process.

Further research is being conducted to improve tissue-engineered cartilage reconstruction. Chondrocytes or stem cells are being seeded on soft tissue scaffolds such as collagen, fibrin, and polylactic acid and precultured in bioreactors containing exogenous growth factors to form neocartilage. This approach may offer an improvement over the cell-only approaches, mainly due to the use of a biomaterial scaffold. As an example, a mixture of mesenchymal stem cells suspended in hyaluronan is being developed for direct injection to a damaged knee meniscus. Similarly to how hydroxyapatite stimulates differentiation of stem cells to bone cells and is well suited for bony applications, hyaluronan (found in the embryonic extracellular matrix) has a chondroinductive and antiangiogenic potential and shows promise as a biomaterial scaffold for cartilaginous tissues.

Biomimetic scaffold materials are clearly more than an inert structural support for cells. When properly selected for the given application, they provide a receptive framework and lead to induction of a cell down a specific differentiation pathway through intracellular signaling (see Figure 5.10). Mechanical forces do play a role in most tissue function, and it has now been shown in cell culture experiments that scaffolds with similar properties to the substrate made by bone cells provide physiological cues to direct stem cell differentiation due to their biomimetic mechanical properties. Extended discussion of biomechanics is found in Chapter 4.

Traditionally, the load-bearing requirements of long bone repair applications have been accomplished by metallic biomaterials. Eventually, scientists working together with engineers will learn how to regenerate total joints and all of the tissues they comprise, but for now, the best option is a total prosthesis that has femoral, tibial, and/or patellar components made of traditional metals and polymers or ceramics (see Figures 5.2, 5.3, and 5.6). Based on the trends of successful current research, most musculoskeletal injuries or defects in the future will probably be treated at an earlier, smaller stage with tissue engineering approaches (see Chapter 6). Resorbable composites delivering cells and growth factors will probably become the material of choice for not only bone but also ligament and cartilage. The role of the traditional metallic load-bearing material may be reduced in the future to serving as a cast or brace that provides protection from biomechanical loading until the new tissue has fully regenerated internally.

5.7.2 Cardiovascular Devices

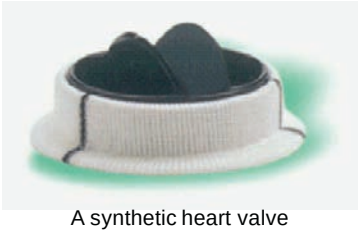
The primary requirements for biomaterials used for blood-contacting circulatory applications such as heart valves and blood vessel replacements or stents are resistance to platelet and thrombus deposition, biomechanical strength and durability, and biocompatibility and nontoxicity. These key requirements were identified by studying the primary mode of action of the tissue to be replaced. For example, arteries consist of three layers that perform various biological functions: the intima, media, and adventitia. The intima is on the blood vessel interior and has a nonthrombogenic surface; it prevents blood contact with the thrombogenic media tissue. The cells of the intima produce a myriad of biomolecules, including growth factors, vasoactive molecules, and adhesion molecules. The media or parenchymal tissue is the middle muscular layer that provides the required strength while remaining viscoelastic. The media layer is made up of multiple layers of aligned smooth muscle cells. The outer adventitia layer acts as a stiff sheath that protects the smooth muscle media layer from biomechanical overload or overdistention.

As with most tissues, autograft is the preferred biomaterial for vascular tissue replacement. For example, one of the patient's own veins can be harvested and used to replace a clogged artery. Vein grafts have a failure rate as high as 20 percent in one year. Since vein grafts from the patient are unavailable and unsuitable in approximately 30 percent of all patients, synthetic graft materials have been developed. The observation that the intact lining of blood vessels (intima) does not induce coagulation has led scientists and engineers to produce more blood-compatible biomaterials by mimicking certain properties of the endothelium. For example, very smooth materials, surfaces with negative charges, and hydrophilic biomaterials are now used with limited success in blood-contacting applications. Knitted Dacron® (polyethylene terephthalate) and Gortex® (polytetrafluoroethylene [PTFE]) vascular grafts are commonly used. The use of synthetic grafts has resulted in reasonable degrees of success (approximately 40 percent experience thrombosis at six months when synthetic grafts are used to bypass arteries that are smaller than 6 mm in diameter). Improved vascular products have incorporated anticoagulants such as heparin on the blood-contacting surfaces.

As discussed in the section on wound healing, most biomaterials are recognized by the body as foreign and lead to platelet deposition and thrombus or coagulation (blood clot formation). This limits their use in blood-contacting applications. Therefore, the tissue engineering approach of growing and implanting a living multilayered cell construct holds great promise for cardiovascular applications. It is only recently that the cell culture conditions and scaffold material selection that will promote the smooth muscle cell alignment and tight endothelial cell packing of blood vessels have been identified. A great deal of additional research must be completed prior to commercial availability of a functional tissue-engineered artery. For example, the mechanical burst strength of the highly cellular tissue-engineered arteries is not yet sufficient to withstand what the heart can generate, although these artificial arteries are nonthrombogenic.

Diseased human heart valves can currently be replaced with mechanical prostheses (synthetic biomaterials) or bioprostheses (made of biological tissue). The two mechanical heart valves shown in Figure 5.17 have four essential components: an occluder such as a disc or ball; a seating ring against which the occluder sits when the valve is closed; a capture

A A natural tissue heart valve



B A synthetic heart valve

FIGURE 5.17 Two types of artificial valves that are used to replace diseased or malformed human valves. (a) The tissue valve generally contains valve leaflets from pigs, while (b) the synthetic valve is composed entirely of manufactured materials.

mechanism, such as a cage, that constrains the occluder when the valve is open; and a sewing ring that permits attachment of the valve to the heart. The occluder bounces back and forth from the seating ring to the capture mechanism with each heartbeat. The more promptly it moves, the more efficient the opening and closing. Thus, weight, or mass, of the occluder and wear resistance of the occluder material are critical features. Early materials that were used included lightweight plastics or hollow metal balls. Silicone rubber proved very effective as a ball. The advent of low-profile disc valves for the mitral position set a new material constraint—namely, stiffness. Silicone rubber was too soft for the disc designs. Polyoxymethylene (POM) or polyacetal is stronger and stiffer than silicone rubber and therefore was used in early disc designs. However, it had a problem with wear. The discs were supposed to be free to spin in the cage so they could distribute wear evenly around the edge of the disc. However, as the disc moved up and down in the cage, wear tracks developed on the edge, preventing spinning and leading to the development of deep wear tracks and valve failure. Due to its high fatigue strength and wear resistance, pyrolytic carbon was selected as one of the prime materials for the occluder.

The capture mechanism and seating ring required strength and stiffness to maintain their shape. Furthermore, they had to be made of a material that could be sprung open for insertion of the occluders. Metal has typically been used, either as machined parts or as separate parts that are welded together. Early designs used the cobalt alloys due to their strength and corrosion resistance. More recently, titanium alloys have been used, due in part to their being lighter than cobalt alloys. Concerns of allergic reactions to cobalt alloy cages have also been expressed over the years.

Both the occluder and the containment system are in direct contact with blood. Contact with foreign materials can cause blood to clot and can lead to platelet attachment and clotting. If the design results in eddy flow or stagnation, the clotting factors can accumulate and lead to thrombus formation. Therefore, both the material selection and device design must consider problems of blood contact and blood flow. Since these problems have not been solved, patients with mechanical heart valves receive medication to reduce their tendency to form blood clots.

Biological heart valves resist thromboembolism much better than synthetic materials do, but they are less durable. The valves are harvested from 7- to 12-month-old pigs and preserved with gluteraldehyde fixation. The gluteraldehyde fixation slowly leads to unwanted calcification of the biological valves and eventual failure. Tissue engineering approaches are now being developed to synthesize living heart valves, but durability remains a large concern. The muscle of the heart has not yet withstood grafting; however, stem cell technology might be able to rebuild the damaged tissue in situ. Recent results from experiments in which mesenchymal stem cells have been injected directly into the ventricle wall have shown that a small fraction of the cells incorporate in the heart muscle and initiate regeneration of damaged heart muscle. Whole pig hearts have now been decellularized and reseeded with human cells to form a living, beating heart. Who knows what these breakthroughs will do to the life expectancy of future generations?

5.7.3 Drug Delivery

Biomaterials play an important role as delivery vehicles for pharmaceuticals and biomolecules. The pharmaceutical industry has long used powdered biomaterials such as talc and calcium carbonate to form pills and tablets containing a drug. The goal of drug delivery research is to prepare formulations that will result in sustained active drug levels in the body, leading to improved drug efficacy. Controlled-release formulations accomplish this by various techniques that involve conjugation of the drug to a biomaterial. For example, by delivering basic fibroblast growth factor bound to heparin, the blood circulation time (as measured by a half-life) is increased by a factor of three. Conjugation to polyethylene glycol (PEG) is a well-established approach for in vivo protein stabilization.

Nanoparticle drug delivery is a rapidly expanding field. Nanoparticles can be made of metals, ceramics, and polymers. One example of a nanoparticle is a liposome. Liposomes are made from phospholipids that form hydrophobic and hydrophilic compartments within an aqueous environment. A unilamellar liposome has spherical lipid bilayers that surround an aqueous core. Water-soluble drugs can be entrapped in the core, and lipid-soluble drugs can be dissolved in the bilayer. Liposomal drug formulations exhibit extended circulating half-lives after intravenous injection. The drug concentration in plasma over time is elevated three to ten times when incorporated in liposomes.

Polymers are widely used in drug delivery systems. Nondegradable hydrophobic polymers (silicone elastomers) have been used most extensively as semipermeable membranes around drug reservoirs. Alternative formulations involve mixtures of drug and resorbable polymer that release drug when the polymer degrades (Figure 5.18). Increasing the loading of the protein or drug within the matrix increases the release rate of the compound. As a rule of thumb, as the average molecular weight of the polymer in the matrix increases,

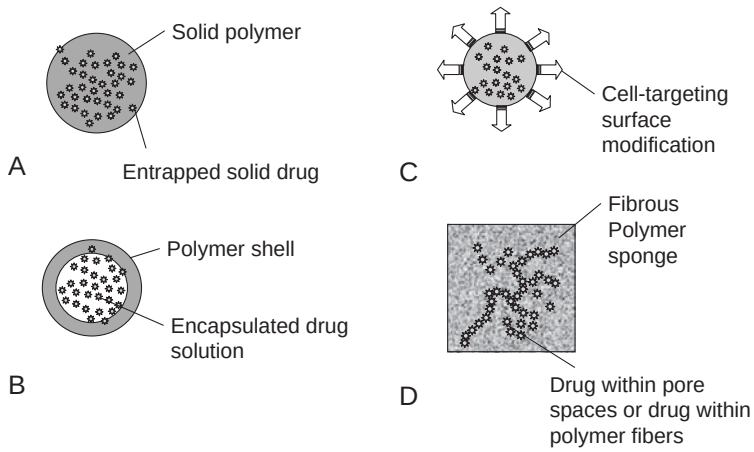


FIGURE 5.18 Polymers provide a versatile matrix for controlled release of drugs and biomolecules. Pharmaceutical agents can be contained within microparticles, microcapsules, or porous polymer blocks or conjugated to single chains of polymer. Drug targeting can be accomplished through the use of cell-specific ligands.

the rate of protein/drug release decreases. Polymer microspheres can be formed around drug solutions, thereby encapsulating the drug as another means of preparing a controlled-release formulation. Polymers that undergo bulk erosion tend to have burst release rates as compared to polymers that undergo surface erosion. The targeting of a drug to a particular cell type can be accomplished by using cell-specific ligands. Microparticles that have been modified with ligands do not diffuse uniformly throughout the body. Instead, the drug is delivered directly to only certain cell types through a receptor–ligand interaction. For example, liposomes have been prepared with folic acid ligands that are preferentially taken up by cancer cells because cancer cells express more folate receptors than do other cell types.

Direct injection of the drug into the targeted tissue is another means of obtaining high local drug concentrations. Controlled release of the drug at the site is accomplished through conjugation of the drug to a biomaterial drug delivery vehicle. This is particularly useful when the drug has toxic side effects that can be minimized by exposing only the tissue of interest to high drug levels. With local drug delivery, the toxic peak blood levels of drug are reduced, and there are sustained levels of active drug over a longer time period, leading to better efficacy (Figure 5.19). As an alternative to direct injection, chemotherapy-carrying magnetic particles (magnetoliposomes) also have been developed. After systemic injection, the drug-loaded particles are then localized to the cancer site by guidance with magnets.

Some drugs do not need to be solubilized to initiate biological activity. For example, nerve growth factor is effective when immobilized on a surface. Even higher levels of activity may be obtained after immobilization of a protein onto a surface due to conformational changes that occur upon immobilization. The active portions of the protein may be better exposed after immobilization. Alternatively, all biological activity may be lost after immobilization, so it is always necessary to conduct separate tests to confirm drug activity after immobilization or conjugation to a biomaterial surface.

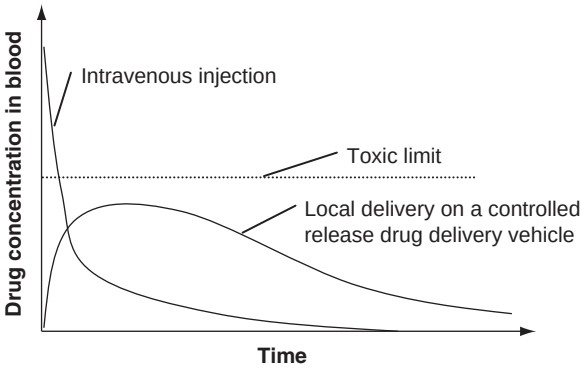


FIGURE 5.19 The primary advantage of local delivery is a reduction of the systemic blood levels of the drug to below toxic levels. High local levels at the site of injection or implantation allow for increased drug efficacy and reduced side effects. This technology is particularly applicable for chemotherapy drugs and therapeutic hormones.

Collagen is commonly used for the delivery of bone growth factors. The protein growth factors are typically not covalently immobilized on the collagen surface, leading to rapid release. Hydroxyapatite materials also are used for delivery of bone growth factors. The drug or protein often attaches quite strongly to the hydroxyapatite surface, and release is greatly delayed as compared to a collagen drug delivery vehicle. Depending on the drug and its mode of action, longer release times may or may not be desired.

Rather than attempting to design a controlled-release drug delivery vehicle with the perfect release profile, drug delivery chips that can be programmed to open up drug compartments by an external signal are being developed. The ultimate smart drug delivery vehicle is the cell. Unlike a passive drug delivery device that acts independently, cells produce cytokines, growth factors, and extracellular matrix materials based on the signals from the in vivo environment. Attempts are being made to exploit this with the implantation of encapsulated xenograft pancreatic islet cells. Pancreatic islet cells produce insulin in response to the circulating blood levels. The xenograft cells need to be encapsulated within a biomaterial (typically alginate) to evade immune surveillance activity that can be toxic to the cells. Current research in this area is focused on varying the properties of the alginate to maintain sufficient permeability to keep the cells vital yet protected from immune cell toxins, while still allowing diffusion of insulin out of the device—a tall order for one material. Materials selection for medical devices will always involve this type of balancing act between properties.

5.8. EXERCISES

1. Explain briefly how the idea and role of a biomaterial has changed since their first use.
2. List the four classes of materials used in medical devices, along with three advantages and disadvantages of each.
3. What advantage does a composite material afford over the singular components?
4. Explain the importance of the extracellular matrix in controlling cell activity and tissue maintenance.

Continued

5. Explain what is meant by biomimetic design. Considering that a biomaterial construct, when combined with cells for therapy, is in intimate contact with the cells and can be considered an extracellular matrix, explain the importance of biomimetic material design.
6. What are the stages of a typical tissue-biomaterial interaction? Include the cell types that are predominant during each stage in your answer.
7. Name three biomimetic surface modifications for biomaterials and their primary purpose. Briefly explain the mechanism by which they can be classified as biomimetic.
8. The two major testing approaches to determine the biocompatibility of a biomaterial are in vitro testing, with or without cells, and in vivo testing in a living organism. What are the difficulties in graduating from the first type to the second?
9. Most metallic biomaterial alloys have an oxide film that spontaneously forms on the alloy surface and renders the material corrosion resistant. With this understanding, explain how and why mechanical wear processes (e.g., fretting) can increase the amount of corrosion present.
10. Describe which type of biomaterial you would select for the construction of the following implantable devices. Explain which properties will be important and why. More than one material can be used in the same device.
    a. Skin substitute
    b. Guidance tube for nerve regeneration
    c. Hip replacement stem
    d. Dental braces
    e. Urinary catheter
    f. Tissue-engineered bone
11. When selecting a biomaterial to be used as an orthopedic implant, what are some of the properties or characteristics of the material that should be considered?
12. You have recently designed a new implantable biomaterial and have conducted an in vivo implantation study. Subsequent extraction of the sample reveals a thin, fibrous capsule surrounding your material. What does this experiment reveal regarding the biocompatibility of this device, specifically regarding the inflammation response?
13. What are some of the benefits of implanting a biomaterial that delivers drugs rather than giving the patients intravenous injections of the drugs?
14. What are some of the applications for which localized delivery of a biologically active molecule from an implanted medical device is advantageous?
15. Describe two methods for making a biomaterial porous.
16. Is the growth of a fibrous tissue layer around an implanted material a positive aspect in all applications? Why or why not? In what applications is a fibrous tissue layer not desired?

Suggested Readings

Annual Book of ASTM Standards, vol. 13.01, Medical and Surgical Materials and Devices, American Society of Testing and Materials (ASTM) International, West Conshohocken, Pennsylvania, 2010.

B. Basu, D.S. Katti, A. Kumar, Advanced Biomaterials, American Ceramic Society, Wiley and Sons, Hoboken, New Jersey, 2009.

K.C. Dee, D.A. Puleo, R. Bizios, An Introduction to Tissue Biomaterial Interactions, Wiley-Liss, Hoboken, New Jersey, 2002.

A.K. Dillow, A.M. Lowman, Biomimetics Materials and Design, Marcel Dekker, New York, 2002.

K.E. Gonsalves, C.R. Halberstadt, C.T. Laurencin, L.S. Nair, Biomedical Nanostructures, Wiley and Sons, Inc., Hoboken, New Jersey, 2008.

M.N. Helmus, Biomaterials in the Design and Reliability of Medical Devices, Landes Bioscience, Texas and Kluwer Academic, New York, 2003.

L.L. Hench, J. Wilson, An Introduction to Bioceramics (Advanced Series in Bioceramics, vol. 1), World Scientific, River Edge, New Jersey, 1993.

P.X. Ma, J. Elisseeff, Scaffolding in Tissue Engineering, CRC Press, Taylor & Francis Group, Boca Raton, Florida, 2006.

M.A. Meyers, P.Y. Chen, A.Y.M. Lin, Y. Seki, Biological Materials: Structure and Mechanical Properties, Prog. Mater. Sci. 53 (2008) 1–206.

J.B. Park, J.D. Bronzino, Biomaterials Principles and Applications, CRC Press, Taylor & Francis Group, Boca Raton, Florida, 2003.

B.D. Ratner, A.S. Hoffman, F.J. Schoen, J.E. Lemons, Biomaterials Science: An Introduction to Materials in Medicine, second ed., Academic Press, San Diego, California, 2004.

J.Y. Rho, L. Kuhn-Spearing, P. Zioupos, Mechanical properties and the hierarchical structure of bone, Med. Eng. Phys. 20 (1998) 92–102.

S.W. Shalaby, K.J.L. Burg, Absorbable and Biodegradable Polymers, CRC Press, Boca Raton, Florida, 2004.

J.S. Temenoff, A.G. Mikos, Biomaterials: The Intersection of Biology and Materials Science, Pearson Prentice Hall, Upper Saddle River, New Jersey, 2008.

C. van Blittersijik, Tissue Engineering, Academic Press. Elsevier, San Diego, California, 2008.

D.F. Williams, The Williams Dictionary of Biomaterials, Liverpool University Press, Liverpool, 1999.

This page intentionally left blank

CHAPTER

6

Tissue Engineering

Randall E. McClelland, Robert Dennis, Lola M. Reid, Jan P. Stegemann, Bernard Palsson, and Jeffrey M. Macdonald

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Discuss the growing area of cellular therapies.
- Discuss the three general categories of extracorporeal bioreactors.
- Understand the cellular dynamics underlying tissue function.
- Qualitatively describe the importance of stem cells in tissue function.
- Quantitatively describe cellular fate processes.
- Analytically describe mass transfer in three different configurations.
- Explain the parameters that characterize the tissue environment and how to approach mimicking them in vitro.
- Describe the issues fundamental to scale-up.
- Define functional tissue engineering and how the "-omics" sciences are driving this newly created research area.
- Discuss the issues that one encounters when implementing cellular therapies or bioartificial organs to patients.

6.1 WHAT IS TISSUE ENGINEERING?

Tissue engineering is a biomedical engineering discipline integrating biology with engineering to create tissues or cellular products outside the body (*ex vivo*) or to use the gained knowledge to better manage the repair of tissues within the body (*in vivo*). This discipline requires understanding of diverse biological fields, including cell and molecular biology, physiology and systems integration, stem cell proliferation and differentiation, extracellular matrix chemistry and compounds, and endocrinology. It also requires knowledge of many engineering fields, including biochemical and mechanical engineering, polymer sciences, bioreactor design and application, mass transfer analysis of gas and liquid metabolites, and biomaterials. Translation of tissue engineering constructs to clinical applications will involve yet other scientific disciplines so novel engineered tissues will be easily accepted and used by clinicians. The combination of these sciences has spawned the field of *regenerative medicine*, which is closely aligned with tissue engineering but has a focus on strategies that use the body's natural regeneration mechanisms to repair damaged tissues. Two of the main goals of these fields are cell therapies for the repair of damaged tissues, involving injection or engraftment of cells or cellular suspensions, sometimes in combination with scaffolding material, or establishing tissue ex vivo for use as grafts or extracorporeal organs to assist or supplement ailing in vivo organs. Clinical trials with cell therapies or extracorporeally created tissues have begun to be undertaken in the area of skin, cartilage, bone, heart, neural, and liver tissues, and the first tissue-engineered products have become available in the last decade. In addition, tissue engineering strategies are being employed to develop improved in vitro diagnostic and screening techniques, as well as creating improved tissue models to study disease. Both scientific and economic issues will define the success of these and future therapeutic modalities.

6.1.1 The Challenges Facing the Tissue Engineer

Some of the fundamental challenges that face the tissue engineer "in the implementation of cell therapies or creation of grafts and bioartificial organs" are shown in Figure 6.1. In particular, the following issues will impact the field as it progresses toward larger-scale clinical application:

1. Reconstitution of physical (mass transfer) and biological (soluble and insoluble signals) microenvironments for the development and control of tissue function.
2. Overcoming scale-up challenges in order to generate cellular microenvironments on a clinically and commercially meaningful scale.
3. Systems automation to provide appropriate process and quality control on clinically and commercially meaningful scales.
4. Implementation of tissue engineering technologies in clinical settings, including appropriate cell handling and preservation procedures that are required for cell therapies and the transplantation of viable tissues.

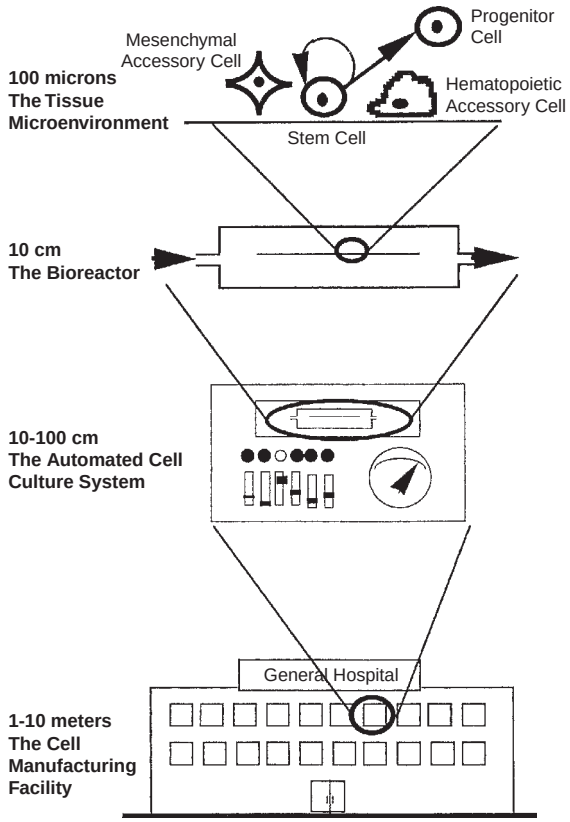


FIGURE 6.1 The four principal size scales in tissue engineering and cellular therapies.

This chapter concentrates on items 1 and 2, although some of the challenges faced with items 3 and 4 are discussed in Section 6.6. For an overall understanding, items 1 and 2 are further illustrated in Figure 6.2 from the viewpoints of "every cell in the body."

In the center of Figure 6.2 is a cell. The figure represents the environment that influences every cell in the body. This environment includes the chemical components of the microenvironment: the extracellular matrix, hormones, cyto/chemokines, nutrients, and gases. Physically, it is characterized by its geometry, the dynamics of respiration, and the removal of metabolic by-products. From these characterized observations, expanded details of each component will be discussed such that biological understandings precede the physical considerations. Finally, both biological and physical environments are combined to help integrate research and clinical activities—for the future development of tissue engineered products.

6.1.2 Cellular Therapies, Grafts, and Extracorporeal Bioartificial Organs

The development of cellular therapies initially arose from advancing knowledge within the cell and molecular biology science domains. Transferring these developments into the clinical arena is a design challenge that requires organized culture control and exploitation

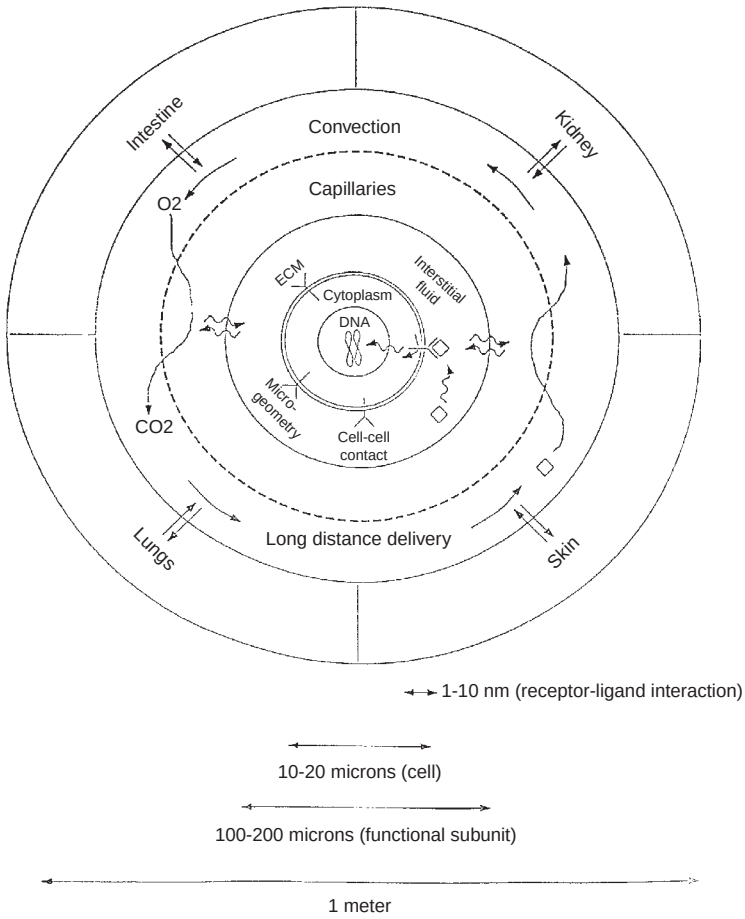


FIGURE 6.2 A cell and its communication with other body parts. *Modified from [11].*

of cell metabolites. For this reason, many scientific fields such as bioengineering, biochemical engineering, and biomaterial sciences are needed for the implementation of cell therapies. A significant challenge in tissue engineering is isolating and growing sufficient numbers of cells for device/therapy designs for clinical and commercial programs. As a historical example, it should be noted that the discovery of penicillin alone was not enough to affect the delivery of health care. Methods for the mass production of clinical grade material had to be developed. The development of such large-scale production of antibiotics arguably represents the most significant contribution of engineering to the delivery of health care. In a similar fashion, the development of industrial-scale methods for isolation, expansion, and cryopreservation of human cells will enable routine uses for cell therapies. In order for tissue engineering to have a tangible impact on modern medicine, the therapies that are developed must be both scientifically and commercially viable.

Tissues are comprised of multiple cell types that interact dynamically with each other. Therefore, tissue-specific functions are often observed only with cocultures of those multiple cell types or with cultures of a particular cell type in combination with the signals from the others. Those signals include insoluble factors in the extracellular matrix, signals from direct cell-cell contact, and soluble signals from autocrine, paracrine, and endocrine interactions. To use these signals as bioengineers, several basic concepts in cell biology need to be understood and quantitatively characterized. These include the key cellular processes of cell differentiation, hyperplastic and hypertrophic growth, migration (motion), protein synthesis, and death (necrosis or apoptosis), all of which combine to define tissue function. Basic information about stem cell and maturational lineage biology and the role of determined stem cells in organ function, genesis, and repair will be presented.

The creation of new engineered tissues requires that many bioengineering challenges be met. For example, bioengineering considerations in cell therapies include injection needle design and procedure protocols. For this application, needles must be optimized to reduce shear stress on cell membranes. Nutrient mass transfer must be analyzed to determine the range of cell aggregate sizes that can be sustained as viable tissues. Engraftment techniques and seed site selection criteria must be established so cells will prosper and assist in system homeostasis. Detrimental events, such as the formation of emboli, need to be prevented. For more complex implantable devices and bioreactors, other challenges will be faced. In these systems, the function, choice, manufacturing, and treatment of biomaterials are important for cell growth and device construction. Fluid mechanics and mass transfer play important roles in normal tissue function and therefore become critical issues in ex vivo cellular device designs. System analysis of metabolism, cell-cell communication, and other cellular processes can be used to define bioartificial organ specifications. A properly designed ex vivo culture system must appropriately balance the rates of biological and physicochemical processes to obtain desired tissue functions. By mathematically modeling this balancing effect, dimensionless parameter groups can be formulated that describe characteristic ratios of time constants. In this way, new dimensionless values will evolve to relate ratios of "physical times" with "biological times."

Finally, the implementation of cell therapies and tissue grafts in the clinic requires the recognition and resolution of several difficult issues. These include tissue harvest, cell processing and isolation, safety testing, cell activation/differentiation, assay and medium development, storage and stability, and quality assurance and quality control issues. These challenges will be described in this chapter but are not analyzed in detail.

6.1.3 Human Cells and Grafts as Therapeutic Agents

Cell therapies use human cells as therapeutic agents to alleviate a pathological condition. It is important to note that some cell therapies are already an established part of medical care. One existing type of cell therapy is blood transfusion, which has been practiced for decades with great therapeutic benefit. This therapy uses red blood cells (RBC) as the transplant product into anemic patients to help to restore adequate oxygen transport. Similarly, platelets have been transfused successfully into patients who have blood clotting problems. Bone marrow transplantation (BMT) has been practiced for almost two decades, with tens of thousands of cancer patients undergoing high-dose chemo- and radiotherapies followed by

BMT. More recently, transplantation of hemopoietic stem cells has occurred with increasing frequency to correct hematological disorders. These are all applications of cell therapies associated with blood cells and blood cell generation (hematopoiesis). (The term *hematopoiesis* comes from the Greek *hemato*, meaning "blood," and *poiesis*, meaning "generation of.") Therefore, a large population of patients already has benefitted from cell therapies, and this benefit can be extended by developing new therapies using other progenitor cell sources.

Transplants can be xenogeneic (donor and recipient are members of different species), allogeneic (donor and recipient are members of the same species but are not genetically identical), or syngeneic (donor and recipient are genetically identical—e.g., clones in the case of animals, or identical twins). Syngeneic transplants include autologous transplants (cells from a patient being isolated and given back to the same person). The issues associated with allogeneic transplants are well known because of the widespread use of organ transplantation and chiefly involve prevention of immune rejection as well as longer-term negative responses to transplanted tissues. However, with the advent of ex vivo cell culture and advances in cell manipulation procedures, autologous transplantation is becoming more common. In addition there are efforts under way in several laboratories to create "universal donor" cell sources and cell lines that would alleviate many of these issues.

The ability to reconstitute tissues ex vivo and produce cells in clinically meaningful numbers has broad implications. Table 6.1 summarizes the supply and demand of organs and tissues versus the number of procedures performed annually in the United States. Although the number of procedures is limited, the overall cost of these procedures was still estimated at a staggering $400 billion per year. The potential socioeconomic impact

TABLE 6.1 Incidence of Organ and Tissue Deficiencies, or the Number of Surgical Procedures Related to These Deficiencies in the United States[a]

Indicator	Procedure or Patients per Year
Skin	
Burns[b]	2,150,000
Pressure sores	150,000
Venous stasis ulcers	500,000
Diabetic ulcers	600,000
Neuromuscular disorders	200,000
Spinal cord and nerves	40,000
Bone	
Joint replacement	558,200
Bone graft	275,000
Internal fixation	480,000
Facial reconstruction	30,000

TABLE 6.1 Incidence of Organ and Tissue Deficiencies, or the Number of Surgical Procedures Related to These Deficiencies in the United States[a]—Cont'd

Indicator	Procedure or Patients per Year
Cartilage	
Patella resurfacing	216,000
Chondromalacia patellae	103,400
Meniscal repair	250,000
Arthritis (knee)	149,900
Arthritis (hip)	219,300
Fingers and small joints	179,000
Osteochondritis dissecans	14,500
Tendon repair	33,000
Ligament repair	90,000
Blood Vessels	
Heart	754,000
Large and small vessels	606,000
Liver	
Metabolic disorders	5,000
Liver cirrhosis	175,000
Liver cancer	25,000
Pancreas (diabetes)	728,000
Intestine	100,000
Kidney	600,000
Bladder	57,200
Ureter	30,000
Urethra	51,900
Hernia	290,000
Breast	261,000
Blood transfusions	18,000,000
Dental	10,000,000

[a]*From Langer and Vacanti (1993).*
[b]*Approximately 150,000 of these individuals are hospitalized and 10,000 die annually.*

of cellular therapies is therefore substantial. Progress in decreasing the costs of these therapies also will encourage investment into new types of treatments. In this way, new medical products will be developed to greatly improve the quality of life and productivity of affected individuals.

The concept of directly engineering tissues was pioneered by Y. C. Fung in 1985. The first symposium on this topic was organized by Richard Skalak and Fred Fox in 1988, and since then the field of tissue engineering has grown rapidly. Thousands of scholarly articles have been written on the topic, and in 1995 a peer-reviewed journal called *Tissue Engineering* was established. Since that time, the original journal has sprouted into three separate journals (*Part A: Primary Papers*, *Part B: Reviews*, and *Part C: Methods*) to more broadly cover the field, and numerous other journals in the topic area have emerged, including the *Journal of Tissue Engineering and Regenerative Medicine*. The field also has received considerable attention in the lay press because of the opportunities it presents to revolutionize medicine for an aging population.

The last two decades have seen remarkable advances in biology that have enabled tangible progress in tissue engineering. Cutting-edge cell therapies that have reached the advanced stages of development include various forms of immunotherapies, chondrocytes for cartilage repair, liver and kidney cells for extracorporeal support devices, β-islet cells for diabetes, skin cells for patients with ulcers or burns, and genetically modified myocytes for treatment of muscular dystrophy. In addition, engineered tissues such as blood vessels, bladders, urethras, and other tissues are rapidly moving toward the clinic. As would be expected based on tissue complexity, the challenges faced with each tissue are different. A few examples are provided in the following sections for illustrative purposes.

Bone Marrow Transplantation

Bone marrow is the body's most prolific organ. It produces on the order of 400 billion myeloid cells daily, all of which originate from a small number of pluripotent stem cells (Figure 6.3). The bone marrow is comprised of 500 to 1,000 billion cells and regenerates itself every two to three days, which represents normal hematopoietic function. Individuals under hematopoietic stress, such as systemic infection or sickle cell anemia, will have blood cell production rates that exceed the basal level. The prolific nature of bone marrow cells makes it especially susceptible to damage from radio- and chemotherapies. Bone marrow damage limits the extent of these therapies, and some regimens are fully myoablative. Without any hematopoietic support, patients who receive myoablative dose regimens will die due to hematopoietic failure.

Bone marrow transplantation (BMT) was developed to overcome this problem. In an autologous setting, the bone marrow is harvested from the patient prior to radio- and chemotherapies. It is cryopreserved during the time period that the patient undergoes treatment. After chemotherapeutic drug application and several half-lives of the drug have passed, the bone marrow is rapidly thawed and returned to the patient. The bone marrow cells are simply introduced into the circulation, and the bone marrow stem cells naturally "home" to the marrow cavity and reconstitute bone marrow function. In other words, the hematopoietic tissue is rebuilt in vivo by these cells. This process takes several weeks to complete, during which time the patient is immunocompromised.

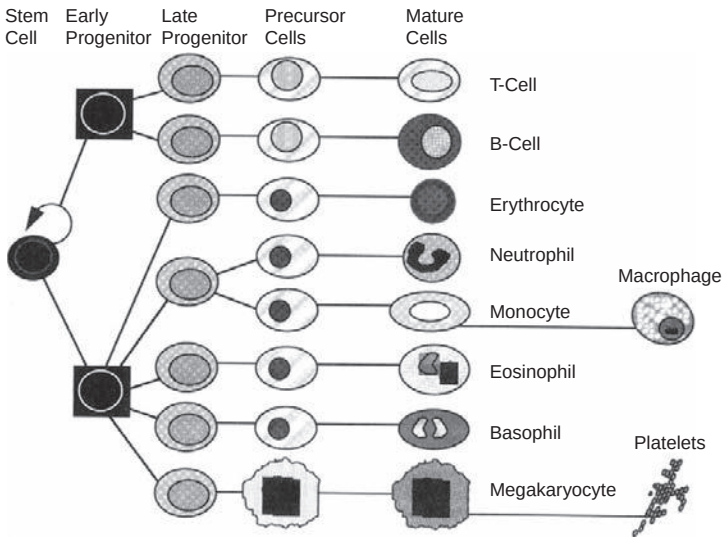


FIGURE 6.3 Hematopoietic cell production. The production fluxes through the lineages can be estimated based on the known steady-state concentration of cells in circulation, the total volume of blood, and the half-lives of the cells. Note that the 400 billion cells produced per day arise from a small number of stem cells. *From [8].*

Current forms of autologous BMT as a cellular therapy simply involve removing the cells from the patient and storing them temporarily outside the patient's body. There are several advantages to growing and increasing the number of harvested cells, and therefore newer therapies and treatments are being developed based on ex vivo culture of hematopoietic cells. In addition, new techniques to harvest bone marrow stem cells have been developed. These methods rely on using cytokines or cytotoxic agents to "mobilize" the stem cells into circulation. The hematopoietic stem and progenitor cells are then collected from the circulation using leukopheresis.

Myoablative regimens are used in allogeneic settings. In the case of leukemia, this not only removes the bone marrow but hopefully also the diseased tissue. The donor's cells migrate to the marrow and repopulate the bone cavity, just as in the autologous setting. The primary difficulty with allogeneic transplants is high mortality (10 to 15 percent), primarily due to Graft-versus-Host Disease, in which immune cells in the transplanted marrow recognize the recipient's tissues as foreign and mount an immunologic attack. Overcoming this rejection problem would significantly advance the use of allogeneic transplantation in BMT as well as other cell-based therapies.

BMT is a well-developed and accepted cellular therapy for a number of indications. These include allogeneic transplants for diseases such as leukemia and autologous transplants for diseases such as lymphoma and breast and testicular cancer. Significant growth has occurred in the use of BMT since the mid-1980s, and tens of thousands of patients are now treated using this family of therapies each year. Improvements continue to be made in tissue harvesting, processing, and transplantation, and these advances help to inform the field of tissue engineering as new cell-based therapies are developed.

Skin and Vascular Grafts

Skin is a highly proliferative and regenerative tissue. It consists of two main layers: the dermis, whose main cellular components are stromal cells or fibroblasts, and the epidermis, whose main cellular components are epidermal cells at various stages of differentiation into keratinocytes (Figure 6.4). Both cell types grow well in culture, and ex vivo cultivation is not the limiting factor with this tissue. Interestingly, transplanted dermal fibroblasts have proven to be surprisingly nonimmunogenic.

Skin transplants and engineered skin technologies have been applied to victims of burns and patients with diabetic ulcers who have severe problems with skin healing. To treat these problems, skin may be cultured ex vivo and applied to the affected areas. Technologically this cell therapy is relatively well developed, and currently there are several "engineered skin" products available, including Apligraf® (Organogenesis Inc.) and Dermagraft® (Advanced BioHealing Inc.). Figure 6.5 shows an expansion bioreactor for creating skin grafts from human foreskins developed by Advanced Tissue Sciences Inc., the now defunct company that originally developed Dermagraft®. The bioreactor is constructed with one mechanical hinge and two ports for constant "bleed-feed" flow (left panel). The graft products—which are similar to autologous skin grafts without hair follicles—are easily removed from the bioreactor, as shown in the right panel. In spite of these successes, important challenges remain in bringing such products to wide clinical use. The relatively high cost of the

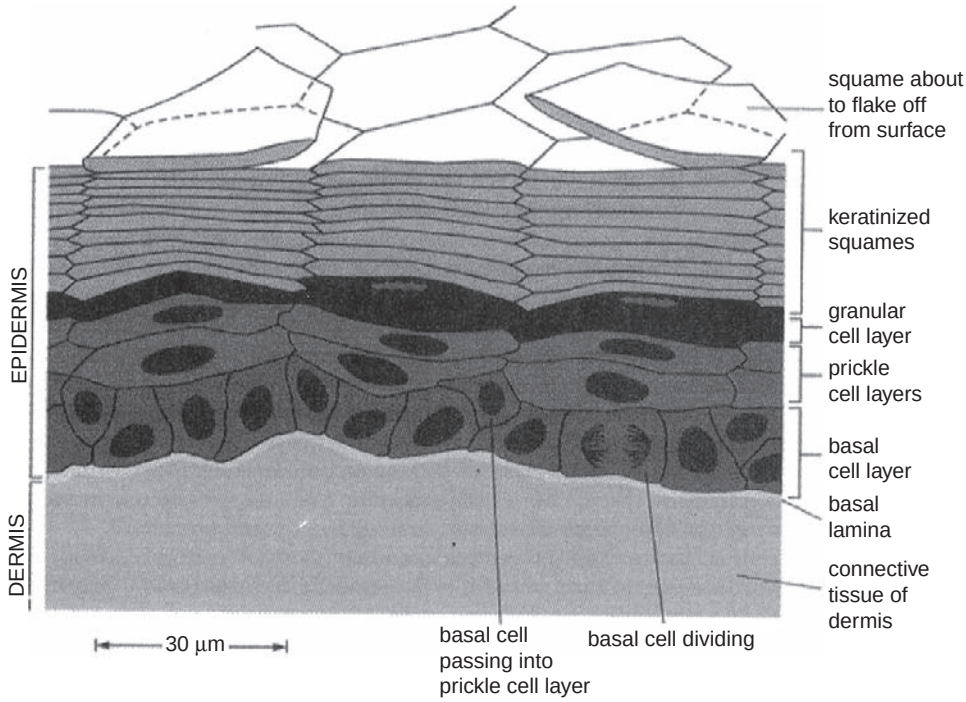


FIGURE 6.4 The cellular arrangement and differentiation in skin. The cross section of skin and the cellular arrangement in the epidermis and the differentiation stages that the cells undergo. *From [1].*

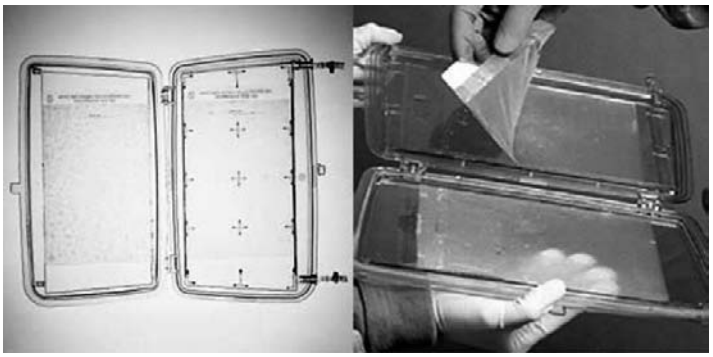


FIGURE 6.5 Advanced Tissue Sciences bioreactor for culture of their skin product, Trancyte, derived from human foreskins.

extensive testing, processing, and quality control that is required for such allogeneic products are a significant barrier to adoption for procedures for which effective therapies already exist. However, these products have introduced tissue engineering technologies to the clinic and have paved the way for future products.

Pancreatic β-Cells

Insulin-dependent diabetes mellitus is characterized by the inability of the pancreas to produce and secrete insulin, a hormone required for normal glucose metabolism. Pancreata derived from cadaveric donors can be used as sources of islets of Langerhans, which contain the insulin-secreting β-islet cells. These cells can be injected into the portal vein leading to the liver, where they lodge and secrete life-sustaining insulin. Unfortunately, at present it is not possible to expand the number of β-cells in culture without losing their essential properties, and thus this procedure is constrained by the severely limited supply of tissue. However, recently there has been exciting progress in the derivation of insulin-secreting β-cells from stem cells. This form of cellular therapy is an allogeneic transplantation procedure, and the duration of the graft can be prolonged by immunosuppressing the recipient.

The immune rejection problem is an important concern in cellular therapies and is treated in a separate section following. However, ways to overcome rejection include the physical separation of the donor's cells from the immune system by a method that allows exchange between the graft and the host across a semipermeable membrane (Figure 6.6). This approach is the basis for many of the encapsulation strategies that have been used to improve the efficacy of islet transplantation. In addition, there have been many efforts to develop "artificial pancreas" devices that allow cells to secrete insulin while being protected from the host's immune system.

Cartilage and Chondrocytes

Cartilage is an unusual tissue in that it is avascular, alymphatic, and aneural. It consists mostly of extracellular matrix in which chondrocytes are dispersed at low densities on the order of 1 million cells/cc (whereas most tissues contain several hundred million cells/cc).

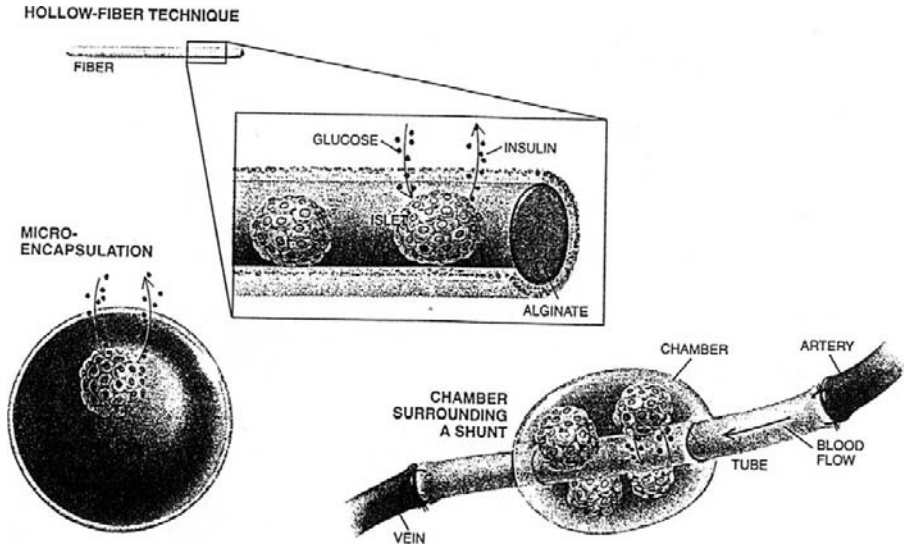


FIGURE 6.6 Encapsulation of islets in semiporous plastic is one promising way to protect them from attack by the immune system. *From Lacey, 2001.*

Chondrocytes can be cultured ex vivo to increase their numbers by up to tenfold. Using this technique, deep cartilage defects in the knee can be treated by autologous cell transplantation. In this case, a biopsy is collected from the patient's knee outside the affected area (Figure 6.7) and the chondrocytes are liberated from the matrix by an enzymatic treatment. The cells are allowed to grow in a two-stage cell ex vivo culture process to increase their number. The cells are then harvested and reintroduced into the affected area within the knee joint. This type of therapy has been shown to aid in healing of cartilage defects and is the basis of the Carticel® cell-therapy product (Genzyme Inc.). Numerous patients are afflicted with knee problems that effectively leave them immobile. It is estimated that over 200,000 patients are candidates for this type of cellular therapy annually in the United States alone.

Hepatic, Neural, and Cardiovascular Cell Therapies

Other forms of cell therapies are being developed to treat complex tissues such as the liver, brain, and heart. It is difficult to obtain and culture the main cell types of these complex tissues, and therefore progenitor and stem cell–based approaches are being actively investigated. Clinical trials for each of these tissues have been initiated using both adult mesenchymal stem cells as well as tissue-specific stem cells. The derivation of liver cells from stem cells is of interest both to treat hepatic failure, as well as to produce improved in vitro screens for drugs. Bone marrow and stem cell transplantation into the heart to treat ischemia and prevent the progression to heart failure has shown promise in the clinic, but there are still many important issues that need to be resolved for such procedures to be used widely. Cell transplantation to the brain also has been attempted clinically with mixed results. Recently, Geron Corporation was given clearance to start the first human clinical trial of

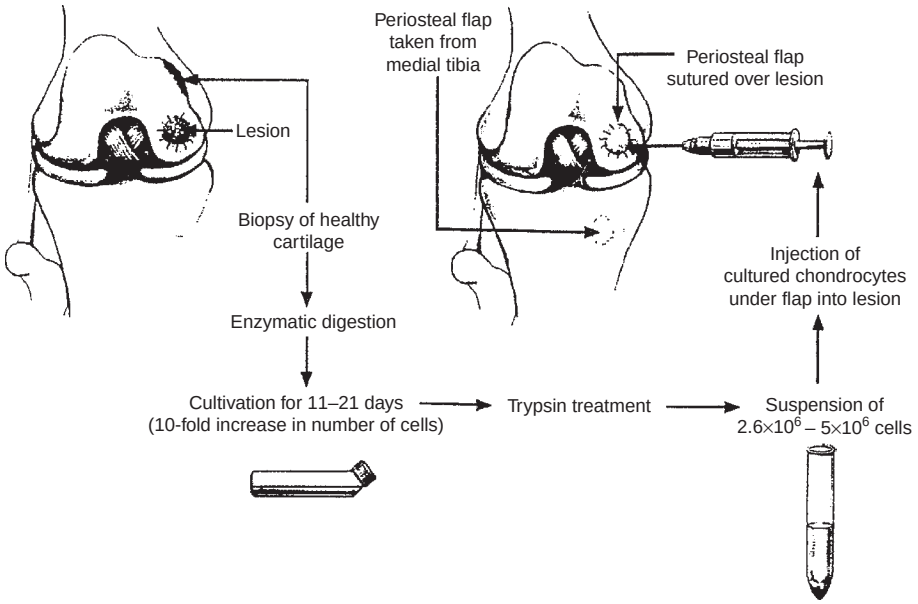


FIGURE 6.7 Chondrocyte transplantation in the right femoral condyle. *From [2].*

embryonic stem cells to treat acute spinal cord injury. These therapies, while exciting and of high potential, are still at a very early stage of development. From the tissue engineering perspective, there are still many unanswered questions regarding the most appropriate cell sources, mass transfer limitations, delivery methods, control of the microenvironment, integration with the host, and provision of an acceptable functional outcome.

6.1.4 Mechanisms Governing Tissues

Normal tissue functions can be used to define the engineering specifications of tissues created in the lab. In addition, our understanding of biology can be used to control cell function to generate tissues. Two important mechanisms that govern tissue dynamics are the relationship between epithelial cells and mesenchymal cells, and stem cell and maturational lineage biology. An understanding of these biological processes can inform the design of engineered tissues. For example, the smallest physiological unit defines mass transfer dimensions, while the number of cells required to replace the desired physiological functions defines the overall dimensions of an engineered product. These numbers help set overall design specifications and goals of the clinical device or engineered tissue.

The Epithelial-Mesenchymal Relationship

A fundamental paradigm defining tissues is the epithelial-mesenchymal relationship, since many tissues are composed of a layer of epithelial cells bound onto a layer of mesenchymal cells (e.g., fibroblasts, smooth muscle cells, etc.), as shown in Figure 6.8. Normal

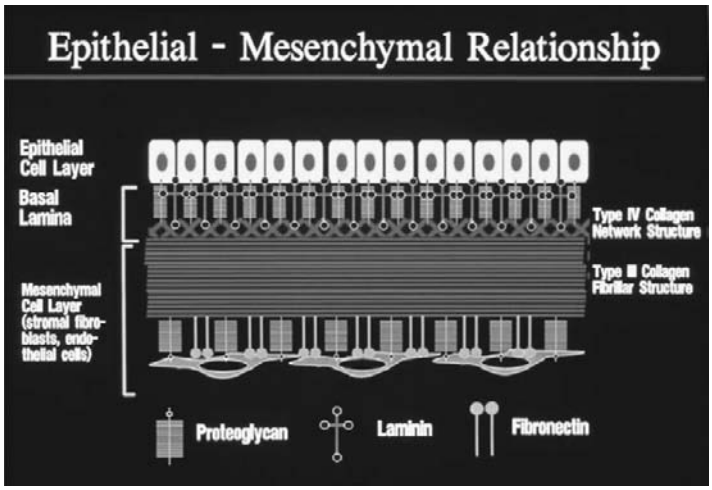


FIGURE 6.8 The epithelial–mesenchymal relationship.

epithelial cells require constant interaction with an appropriate mesenchymal partner or with matrix and soluble signals mimicking that relationship. Communication between and within the two cell layers coordinates local cell activities and is effected by soluble signals (autocrine, paracrine, and endocrine factors) working synergistically with extracellular matrix.

The Soluble Signals

Dissolved growth factors, hormones, and small molecules are often characterized in the following way:

- *Autocrine* factors are factors produced by cells that then act on those same cells.
- *Paracrine* factors are factors produced by cells that then act on neighboring cells in the same tissue.
- *Endocrine* or *systemic* factors are factors produced by cells that are then carried through the blood or lymphatic fluid to target cells in a distant tissue.

There are multiple large families of factors that operate as autocrine and paracrine signals. All of them typically can act as mitogens (i.e., elicit growth responses from cells) or can drive differentiation (i.e., induce expression of specialized tissue-specific functions), depending on the characteristics of the extracellular matrix associated with the cells. Some of the most well studied of these factors include the insulin-like growth factors (IGF), epidermal growth factors (EGF), fibroblast growth factors (FGF), colony stimulating factors (CSF), platelet derived growth factors (PDGF), transforming growth factors (TGF), and cytokines such as interleukins. It should be noted that these molecules are highly pleiotropic and can produce multiple effects on many cell types. The name "growth factor" comes from the fact that most were initially identified by assays in which a growth response was being examined.

The Extracellular Matrix

The extracellular matrix (ECM) is the material that underlies or surrounds cells in tissues. The ECM is diverse and abundant in most tissues and comprises up to half the proteins in the body. Indeed, the collagen family of ECM proteins contains over 25 members and accounts for 2 to 30 percent of the proteins in the body. For many years, the ECM was thought to play an entirely mechanical role binding together cells in specific arrays. Over the last 30 years, we have learned that the ECM is a major regulator of cell and tissue function. In addition to mechanical structure, it provides direct biochemical signals to cells in tissues and also acts as a regulator of many of the soluble growth factors in tissues.

In epithelial cells the ECM between homotypic cells can be referred to as the "lateral" ECM, while the proteins underlying the cells (and separating them from underlying mesenchymal cells) are often referred to as the basal matrix or "basement membrane." Two of the primary components of the lateral extracellular matrix are: cell adhesion molecules, or "CAMs," which are age- and tissue-specific, and proteoglycans, molecules containing a protein core to which polymers of sulfated (negatively charged) sugars called glycosaminoglycans (e.g., heparan sulfates, heparins, chrondroitin sulfates, or dermatan sulfates) are attached. The basal extracellular matrix consists of basal adhesion molecules (e.g., laminins or fibronectins) that bind the cells via matrix receptors (integrins) to one or more types of collagen scaffoldings. The collagens of one cell layer are cross-linked to those of the adjacent cell layer to provide stable coupling between the layers of cells. In addition, proteoglycans are bound to the basal adhesion molecules, to the collagens, and/or to the basal cell surface.

The lateral and basal extracellular matrix components of epithelial tissues provide direct signaling to cells in the form of chronic or persistent signaling. Indirectly, the components also facilitate signaling by stabilizing cells in appropriate configurations of ion channels, receptors, antigens, and so on; by influencing intracellular pathways; and by inducing appropriate cell shapes (flattened or three-dimensional). These interactions enable the cells to respond rapidly to soluble signals that may derive from local or distant sources. The soluble factors act by binding to high-affinity molecules called receptors, which can be on the cell surface or present in the interior of the cell. When the signal binds to its receptor, a signal initiation response is triggered.

Cell Numbers In Vivo and Orders of Magnitude

The cell densities in human tissues are on the order of 1 billion to 3 billion cells/ml. The volume of a 70 kg human is about 70,000 ml. Therefore, the human body consists of about 100 trillion (trillion $= 10^{12}$) cells. The volume of a "typical" organ is around 100 to 500 ml, and therefore a typical organ contains about 100 to 1,500 billion (10^9) cells. Organs are comprised of functional subunits. Their typical linear dimensions are approximately 100 μm. The cell number in a cube that is 100 μm on each side is estimated to be about 500 to 1,000. These cell numbers are summarized in Table 6.2.

Based on these estimations, a typical organ will have a few hundred million functional subunits, each of which is quite small. This number is dictated by the capability of each subunit and the overall physiological need for its particular function. For example, the

TABLE 6.2 Cell Numbers in Tissue Biology and Tissue Engineering: Orders of Magnitude

Cell numbers *in vivo*	
Whole body	10^{14}
Human organ	10^{9}–10^{11}
Functional subunit	10^{2}–10^{3}
Cell production *in vivo*	
Theoretical maximum from a single cell (Hayflick limit)	$2^{30–50} < 10^{15}$
Myeloid blood cells produced over a lifetime	10^{16}
Small intestine epithelial cells produced over a lifetime	5×10^{14}
Cell production *ex vivo*	
Requirements for a typical cellular therapy	10^{7}–10^{9}
Expansion potential[a] of human tissues	
Hematopoietic cells	
Mononuclear cells	10-fold
CD34 enriched	100-fold
Two or three antigen enrichment	10^{6}- to 10^{7}-fold
T cells	10^{3}- to 10^{4}-fold
Chondrocytes	10- to 20-fold
Muscle, dermal fibroblasts	$>10^{6}$-fold

[a]*Expansion potential refers to the number of cells that can be generated from a single cell in culture.*

number of nephrons in the kidney is determined by the maximal clearance need of toxic by-products and the clearance capability that each nephron possesses. These estimations provide insight into tissue structure and function. The fundamental functional subunit of most tissues contains only a few hundred cells, and in most cases this is a mixed-cell population, since most organs have accessory cells that can be as much as 30 percent of the total cell number. Further, as illustrated following, the tissue-type-specific cells may be present at many stages of differentiation.

The nature of tissue microenvironments, along with cellular dynamics, communication, and metabolic processes, must be understood in order to reconstitute tissue function accurately. In addition, generating a therapeutic dose of cells requires a large number of microenvironments. These microenvironments must be relatively similar to have all the functional subunits perform in comparable fashions. Therefore, the design of cell culture devices must provide uniformity in supporting factors, such as nutrient, oxygen, and growth factor/hormone concentrations. These inputs must be reasonably homogeneous down to 100 μm distances. Below this size scale, nonuniformity would be expected and in fact needed for proper functioning of tissue function subunits.

6.1.5 Clinical Considerations: Important Questions

What are clinically meaningful numbers of cells?

Currently, the cells estimated to be needed for clinical practice and experimental cell therapy protocols fall into the range of "a few tens of millions" to "a few billion" (Table 6.2). Since tissue-like cultures require densities above 10 million cells per milliliter, the sizes of the cell culture devices for cell therapies appear to fall into the range of 10 to a few hundred milliliters in volume.

What are the fundamental limitations to the production of normal cells?

The number of divisions a cell can undergo is dependent on its maturational lineage stage, with the key stages being (1) stem cells (diploid, pluripotent), (2) diploid somatic cell subpopulations (unipotent), and (3) polyploid cell subpopulations. Stem cells can self-replicate and undergo unlimited numbers of divisions (stem cells are also a potential source of immortalized cell lines and of tumor cells). Diploid cells are limited in the number of times they can divide and are subject to the so-called Hayflick limit. Normal, somatic, diploid human cells can undergo about 30 to 50 doublings in culture. Therefore, a single diploid cell can theoretically produce 10^{10} to 10^{15} cells in culture. Given the requirements for cell therapies, the Hayflick limit is not a major issue for cell therapies that utilize diploid cells, either stem or somatic phenotypes. However, the expansion potential is minimal or negligible for subpopulations of polyploid cells, which are found in all tissues. Conversely, expansion potential is very high in various quiescent tissues such as the liver and heart. Regenerative stimuli can cause polyploid cells to undergo DNA synthesis with negligible capability to undergo cytokinesis. This results in an increase in their level of ploidy, an increase in cell volume, and a phenomenon called hypertrophy. Thus, the regeneration of tissues is a combination of hyperplasia (the diploid subpopulations) and of hypertrophy (the polyploid subpopulations) activities.

How rapidly do normal cells grow in culture?

Normal cells vary greatly in their growth rates in culture. Hematopoietic progenitors have been estimated to have 11- to 12-hour doubling times, which represent the minimum cycle time known for adult human cells. Dermal foreskin fibroblasts grow with doubling times of 15 hours, a fairly rapid rate that may be partially attributable to the fact that they are isolated from neonatal tissue containing a high proportion of stem/progenitor cells. In contrast, adult chondrocytes grow slowly in culture, with doubling times of about 24 to 48 hours.

How are these cells currently produced?

Expansion of cells can be performed in a variety of culture containers (plates, flasks, roller bottles, bags), and in some cases also in suspension cultures (e.g., certain hemopoietic cells or cell lines). Novel scaffolding material is permitting ex vivo expansion of cell types that previously have proved difficult to culture. The ability to maximize expansion potential for specific cell types requires precise culture conditions. These conditions may comprise specific forms of extracellular matrix, defined mixtures of hormones and growth

factors, nutritional supplements, and basal media containing specific concentrations of calcium, trace elements, and gases. Reviews of these conditions for many cell types are available in recently published textbooks and journal articles.

6.2 BIOLOGICAL CONSIDERATIONS

6.2.1 Stem Cells

It is now thought that essentially all tissues contain stem cell populations that can produce cellular progeny that differentiate into mature tissue phenotypes. The maturational process includes two branches: the "commitment" branch, in which pluripotent stem cells produce daughter cells with restricted genetic potentials appropriate for a single set of cell activities (unipotent), and the "differentiation" branch, in which sets of genes are activated and/or altered in their levels of expression. The following sections discuss how these major tissue subdivisions work together to generate or repair tissues in order to provide tissue functions.

Stem Cells and the "Niche" Hypothesis

Pluripotent stem cells are cells that are capable of producing daughter cells with more than one fate; they can self-replicate, and they have the ability to produce daughter cells identical to the parent. Totipotent stem cells are cells that can generate all the cell types of the organism. Determined stem cells are cells in which the genetic potential is restricted to a subset of possible fates; they can produce some, but not all, of the cell types in the organism.

Determined stem cells of the skin can produce all the cell types in the skin but not those of the heart. Similarly, determined stem cells of the liver can produce all liver cell types but not brain. The lay press often refers to determined stem cells as "adult stem cells," which is a misnomer because determined stem cells are present in fetal and adult tissue. The determined stem cells give rise to unipotent progenitors, also called committed progenitors, with genetic potential restricted to only one fate. These unipotent progenitors rapidly proliferate into large numbers of cells that then differentiate to mature cells. The stem cells and the unipotent progenitors are the normal counterparts to tumor cells and to immortalized cell lines. Determined stem cells identified to date are small in size (typical diameters of 6–10 μm), have a high nucleus to cytoplasmic ratio (blast-like cells), and express certain early genes (e.g., alpha-fetoprotein) and antigens (e.g., CD34, CD117). They have chromatin that binds particular dyes at levels lower than that of the chromatin in mature cells, enabling them to be isolated as "side-pocket" cells using flow cytometric technologies. Stem cells express an enzyme, telomerase, that maintains the telomeres of their chromosomes at constant length, a factor in their ability to divide indefinitely in vivo and ex vivo. Multiple parameters must be used to permit isolation and purification of any determined stem cell type, since there is no one parameter (antigen, size, cell density) sufficient to define any determined stem cells. Furthermore, they appear to grow very slowly in vivo and may commit to growth and differentiation in a stochastic manner. A first-order rate constant for hemopoietic stem cells is about one day and their cycling times have been measured

by means of time-lapse videography. They commit to differentiation in culture. The first and second doubling take about 60 hours, and then the cycling rate speeds up to about 24 hours cycling time. By the fifth and sixth doubling, they are dividing at a maximal rate of 12 to 14 hours doubling time.

What evidence is there that stem cells exist?

Lethally irradiated mice that would otherwise die from complete hematopoietic failure can be rescued with as few as 20 selected stem cells. These animals reconstitute the multiple lineages of hematopoiesis as predicted by the stem cell model. In sublethally irradiated animals, genetically marked mesenchymal stem cells found in bone marrow will give rise to cells in multiple organs over a long time period. These investigations and many others have established conclusively the presence of stem cells, their multilineage potential, and their ability to persist over long periods of time in vivo.

Stem Cell Niches

The field of stem cell niches is rapidly expanding due to its importance in regulating stem cell fate. The goal is to define and understand the local microenvironment of the stem cell compartment. The field is still new but already has yielded generalizations that are proving to be useful guides for defining ex vivo expansion conditions for the cells:

- Stem cells do not have the enzymatic machinery to generate all their lipid derivatives from single lipid sources and so require complex mixtures of lipids for survival and functioning.
- Calcium concentrations are quite critical in defining whether stem cells will expand or undergo differentiation. The mechanisms underlying the phenomenology are poorly understood.
- Specific trace elements, such as copper, can cause more rapid differentiation of some determined stem cell types. It is unknown whether this applies to all stem cells, and the mechanism(s) is not known.
- Specific mixtures of hormones and growth factors are required, with the most common requirements being insulin and transferrin/Fe. Addition of other factors can result in expansion of committed progenitors and/or lineage restriction of the stem cells toward specific fates.
- The matrix chemistry of known stem cell compartments consists of age-specific and cell-type-specific cell adhesion molecules, laminins, embryonic collagens (e.g., type III and IV collagen), hyaluronans, and certain embryonic/fetal proteoglycans. With maturation of the stem cells toward specific cell fates, the matrix chemistry changes in a gradient fashion toward one typical for the mature cells. Although the matrix chemistry of the mature cells is unique for each cell type, a general pattern is the inclusion of adult-specific cell adhesion molecules, various fibrillar collagens (e.g., type I, II collagen), fibronectins, and adult-specific proteoglycans. A major variant is that for skin and neuronal cells, in which mature cells lose expression of collagens, fibronectins, and laminin. Additionally, the matrix chemistry of these cell types is dominated by CAMs and proteoglycans.

- Stem cells are dependent on signals from age- and tissue-specific stroma. The signals from the stroma are only partially defined but include signals such as Leukemia Inhibitory Factor (LIF), various fibroblast growth factors or FGFs, and various interleukins (e.g., IL 8, IL11).
- The most poorly understood of all the signals defining the stem cell niche(s) are those from feedback loops, which are initiated from mature cells, and where these signals inhibit stem cell proliferation. Implicit evidence for feedback loops is that cell expansion ex vivo requires separation between cells capable of cell division (the diploid subpopulations) and the mature nonproliferating cells.

Which tissues have stem cells?

For decades it was assumed that stem cell compartments exist only in the rapidly proliferating tissues such as skin, bone marrow, and intestine. Now, there is increasing evidence that essentially all tissues have stem cell compartments, even the central nervous system. There are now numerous reports of the isolation and characterization of tissue-specific stem cells, and they are actively being investigated as a potential cell source for tissue engineering.

The Roles of Stem Cells

The stem cell compartment of a tissue is the fundamental source of cells for turnover and regenerative processes. Stem cell commitment initiates cell replacement and genesis of the tissue, resulting in tissue repair and maintenance of tissue functions. Stem cell depletion due to disease or toxic influences (e.g., drugs) eventually leads to partial or complete loss of organ function. Mutational events affecting the stem cells can result in tumors for which both altered and normal stem cells are actively present. Thus, tumors are now considered transformed stem cells, an idea originally proposed by Van Potter, Sell, and Pierce, and now confirmed by current stem cell biologists.

6.2.2 The Maturational Lineages of Tissues

All stem cells are pluripotent, giving rise to multiple, distinct lineages of daughter cells that differentiate, stepwise, into all of the mature cells of the tissue. A general model for the production of mature cells arising from tissue-specific stem cells is shown in Figure 6.9. Determined stem cells (pluripotent) replicate slowly in vivo, with rates influenced by various systemic signals. Their immediate descendants are committed progenitors (unipotent) capable of rapid proliferation and shown in some tissues (e.g., skin) to be the acute responders to mild to moderate regenerative stimuli. The unipotent progenitors mature in a stepwise fashion through intermediate stages into fully mature cells. Characteristically, various tissue-specific functions are expressed in cells throughout the maturational lineage and in a lineage-dependent fashion. One can generalize about these gradual phenotypic changes by categorizing the functions as “early,” “intermediate,” and “late” taskings. In some cases, a specific gene is expressed uniquely only at a specific stage. In others, there are isoforms of genes that are expressed in a pattern along the maturational lineage, while in yet others there are changes in the levels of expression of the gene. Finally, the cells progress to senescence, a phenomena that typifys aging cells. The maturation of the cells is dictated in part by mechanisms inherent in the cells (e.g., changes in the chromatin)

	Stem Cell	Early Progenitor	Late Progenitor	Precursor Cell	Mature Cell
Cell Number	Potential 2^{30} to 2^{50} per cell				Need About 10^{16} total over lifetime
Cell Cycling	Very slow (t_d ~ 1/6wks.)	Slow (t_d~60–100 hrs.)	Very rapid (t_d ~ 12 hrs.)	Slow	Zero (can be activated in special cases)
Apoptosis	Inactive	Inactive	Very Active (1:5,000 survives)	Slow	Inactive (can be induced)
Motility	Zero (except during homing)	Zero	Low	Higher	Function of Physiological State
Regulation	Cell-Cell Contact	Cell-Cell Contact	Soluble Growth Factors	Soluble Growth Factors	Soluble Growth Factors

FIGURE 6.9 Model for cell production in prolific tissues. This model was derived from decades-long research in hematology. The columns represent increasingly differentiated cells, while the rows indicate the cellular fate processes (Figure 16.4) and other events that cells undergo at different states of differentiation. (t_d denotes doubling time.)

and in others by matrix and/or soluble signals in their microenvironment. The microenvironment can include signaling affecting growth, differentiation, or apoptosis. For example, some growth factors are survival factors with antiapoptotic effects.

Examples of Stem Cell–Fed Maturational Lineages

The stem cell models best characterized are the hemopoietic stem cells, the intestinal stem cells, and the skin stem cells. In addition, bone marrow–derived mesenchymal stem cells and stem cells of the liver are being increasingly investigated. These systems are described further in the following sections.

Bone Marrow and Blood Cell Formation

Hematopoiesis was the first tissue function for which a stem cell model was established (see Figure 6.3). The reconstitution of the multiple lineages of hematopoietic cells following a stem cell transplant in mice has been demonstrated to occur with a surprisingly small number of cells. Once the cells have reached certain maturational stages, they leave the bone marrow and enter the circulation, where they perform their mature cell functions. The mature cells eventually die and must be replaced. The rate of death of mature cells (by apoptosis or necrosis) sets the need for the cell production rate in the tissue. Ultimately, this death rate determines the number of stem cell commitments that are required. The specific hematopoietic lineage cell production is given in Figure 6.3.

The Villi in the Small Intestine

The lining of the small intestine is comprised of villi that absorb nutrients, as shown in Figure 6.10. The intestinal epithelial cell layer is highly dynamic. Its cellular content turns over approximately every five days and is a very prolific tissue. In between the villi are tube-shaped epithelial infoldings, known as crypts. All intestinal epithelial cell production

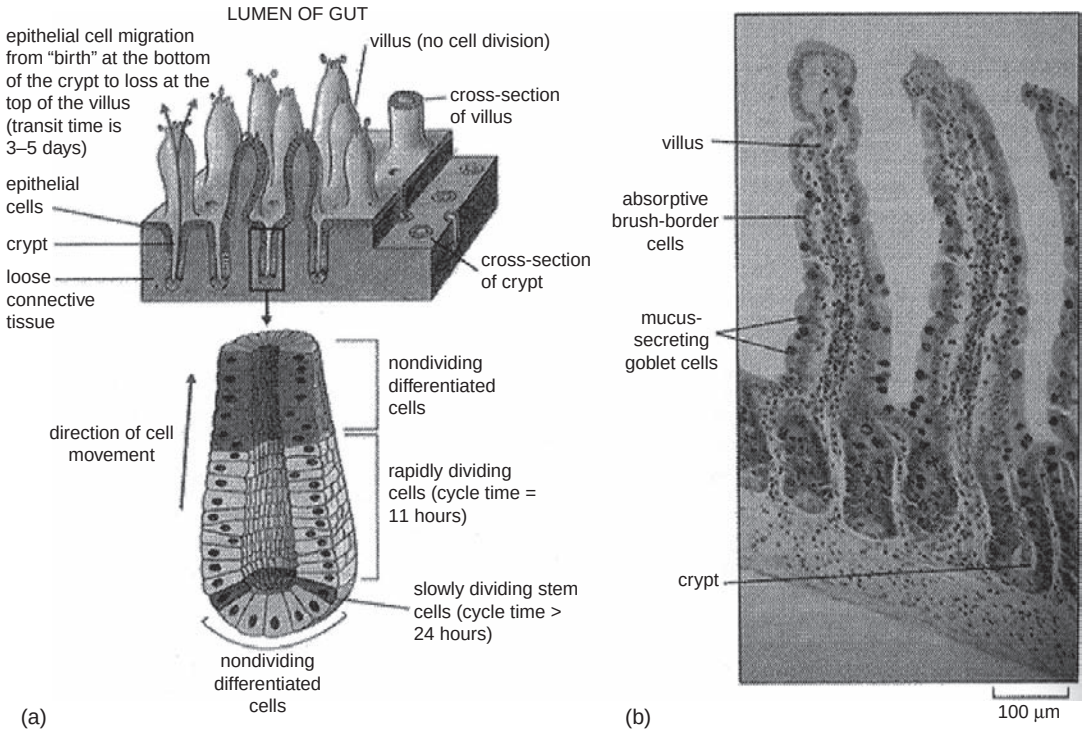


FIGURE 6.10 Villi in the small intestine. (a) The villi and the crypt indicating the mitotic state of the cells in various locations. (b) Rows of villi of epithelial intestinal cells (the diameter of a villi is about 80 mm). *From [1].*

takes place in set locations within the crypt. Once the cells are mature, they migrate to the outer edge of the crypt and then move over a period of about five days from the base of the villus to the top, where they die and slough off. During this passage, the cells carry out their organ-specific function as mature parenchymal cells. They function in the absorption and digestion of nutrients that come from the lumen of the gut. Toward the bottom of the crypt is a ring of "slowly dividing" determined stem cells. The number of stem cells per crypt is about 20. After division, the daughter cell moves up the crypt, where it becomes a rapidly cycling progenitor cell, with a cycling time on the order of 12 hours. The cells that are produced move up the crypt and differentiate. Once they leave the crypt, they are mature and enter the base of the villi.

Skin

Human skin has two principal cell layers, an epidermis and dermis, separated by a form of extracellular matrix called a basal lamina (or basement membrane) (see Figure 6.4). In this extracellular matrix, collagen VII is an important component. The two layers undulate with respect to each other, and the undulations produce deep pockets (distant from the skin surface) and other pockets that are shallower. The pattern of the relationship between layers is unique for each type of skin (e.g., that on the trunk of the body versus that in the palms of

the hand versus that on the face). The dermis is a connective tissue layer under the basal lamina and is comprised primarily of fibroblasts (also called stroma). The epidermis consists of multiple layers of epithelia comprised of differentiating keratinocytes, with the least differentiated cells located at the basal lamina. Thin skin has a squamous columnar organization (each column about 30 μm in diameter). A stem cell compartment has been identified in the deep pockets within the epidermis/dermis undulations and also in bulges near hair follicles. The stem cells produce committed progenitors, also called "transit amplifying cells," which migrate into the shallow pockets and from there into cells that line the entire basal lamina. Only epidermal cells adherent to the basal lamina are cycling, while cells that lose their attachment to the basal lamina move upward and differentiate into succeeding stages of cells. Ultimately, they turn into granular cells, and then into keratinized squames that eventually flake off. Keratins are proteins defining the differentiated cells of the skin and are evident in many forms—from forms that provide skin mechanical protection to those that are present in body hair (or in feathers in birds or scales in other vertebrates). The net proliferative rate of skin depends on the region of the body. In particular the turnover of skin is on the order of a few weeks.

The Liver

The liver is a maturational lineage system, including a stem cell compartment, that is analogous to those in the bone marrow, skin, and gut. The liver's lineage is organized physically within the acinus, the structural and functional unit of the liver. In cross section, the acinus is organized like a wheel around two distinct vascular beds: six sets of portal triads, each with a portal venule, hepatic arteriole, and a bile duct, form the periphery, and the central vein forms the hub. The parenchyma, effectively the "spokes" of the wheel, consist of plates of cells lined on both sides by the fenestrated sinusoidal endothelium. By convention, the liver is demarcated into three zones: zone 1 is periportal; zone 2 is midacinar; and zone 3 is pericentral. Blood flows from the portal venules and hepatic arterioles at the portal triads, through sinusoids that line plates of parenchyma, to the terminal hepatic venules and into the central vein. The stem cell compartment is present around the portal triads, zone 1, and identified in anatomical entities called Canals of Hering. The stem cells of the liver are at least bipotent and produce daughter cells that become either biliary cells (bile duct epithelia) or hepatocytes. Hepatocytes display marked morphologic, biochemical, and functional heterogeneity based on their zonal location. The size of hepatocytes increases from zone 1 to zone 3, and one can observe distinctive zonal variations in morphological features of the cells, such as mitochondria, endoplasmic reticulum, and glycogen granules. Hepatocytes show dramatic differences in DNA content from zone 1 to zone 3, with periportal cells being diploid and midacinar to pericentral cells shifting toward polyploid distinctions. Octaploid cells in the pericentral zone show evidence of apoptosis. Adult rodent livers (rats and mice) are 90 to 95 percent polyploid; adult human livers are 40 to 50 percent polyploid, whereas fetal and neonatal liver cells are entirely diploid. The transition to adult ploidy patterns is observed by age 3 to 4 weeks in rats and mice, and by late teenage years in humans. With age, the liver becomes increasingly polyploid in all mammalian species surveyed. This may help to explain the reduction in regenerative capacity of the liver with age.

6.2.3 Models for Stem Cell Proliferative Behavior

It should be clear by now that the replication functions of stem cells are critical to tissue function and tissue engineering. How do stem cells divide, and what happens when they divide? Three models describe the dynamic behavior of the stem cell population.

The Clonal Succession Concept

Cellular systems are maintained by a reservoir of cells that either grows very slowly or may be in a dormant state. In these cases, the reservoir of cells is available throughout the tissue's lifespan and can be routinely challenged to enter the complex process of cell proliferation and differentiation. Once triggered, such a stem cell would give rise to a large clonal population of mature cells. Any one of these clones would now have a limited life span, since feedback signals dissipate the need to maintain cell production fluxes. After time, such a clone will "burn out," and, if needed, a new stem cell clone would take over the cell production role.

Deterministic Self-Maintenance and Self-Renewal

This model relies on an assumption that stem cells can self-replicate. Following a stem cell division, there is a 50 percent probability that one of the daughter cells maintains the stem cell characteristics, while the other undergoes differentiation. The probability of self-renewal is regulated and may not be exactly 50 percent, depending on the dynamic state of the tissue.

Stochastic Models

This model considers that the progeny of a stem cell division can generate zero, one, or two stem cells as daughter cells (notice that the clonal succession model assumes zero, and the deterministic model assumes one). The assumption is that each of the three outcomes has a particular property.

6.2.4 Stem Cells and Tissue Engineering

Stem Cells Build Tissues

Stem cells are the source of the cells that make up all tissues in the body during development. In addition, certain tissues in adults have reservoirs of resident stem cells that can be mobilized to repair damage when needed. Thus, stem cells build and maintain tissues in vivo and are of great interest because of their potential use to also generate tissue ex vivo. The use of stem cells in tissue engineering has rapidly gained momentum because of their very high proliferative capacity and ability to differentiate into multiple cell types.

Ex Vivo Growth and Manipulation of Stem Cells

The classical definition of a stem cell holds that it has both the ability to divide into a new stem cell without differentiating and to differentiate into a specialized cell type. This makes stem cells attractive as a cell source for tissue engineering because the number of stem cells can be expanded greatly before they are differentiated toward a specific tissue lineage. This scenario offers a theoretically inexhaustible source of cells with which to create new tissues. However, the ex vivo conditions for truly self-replicative expansion have not been

fully determined for any stem cell family, though there have been considerable improvements toward this goal.

Isolating Stem Cells for Scientific and Clinical Purposes

Methods have been developed to isolate stem cells or enrich the stem cell content of a cell population. At present no single marker has been identified for use as a definitive stem cell marker. Therefore, most effective isolation protocols use multiparametric isolation strategies. These strategies may comprise immunoselection for cells with specific antigenic profiles that are used in combination with cell selection methods such as diameter, cell density, and levels of "granularity" (the extent of cytoplasmic particles such as mitochondria). The published protocols include flow cytometry and/or immunoselection with magnetic columns, affinity columns, and counterflow elutriation. Once purified, validation of the identity of stem cells can be achieved either ex vivo or in vivo. Ex vivo assays typically involve clonogenic expansion assays, in which a single cell is expanded in culture under precise conditions, and the ability to give rise to daughter cells of more than one fate identifies the original cell as a stem cell. In vivo assays involve transplanting the putative stem cell into an animal to observe whether multiple tissue types can be regenerated.

Types of Stem Cells

Embryonic stem cells are the most pluripotent cell type, since they can give rise to essentially any tissue in the body. These cells are typically derived from very early stages of embryonic development (usually the blastocyst stage). They have high potential to create large populations of specialized cell types, but their use is controversial because of ethical issues. In addition, their use for therapeutic cell transplantation is in question because they have the immune profile of the original embryo and therefore are necessarily an allogeneic therapy. For these reasons, scientists have very actively searched for alternatives to embryonic stem cells. The technique of somatic nuclear transfer involves replacing the nuclear material of an egg cell with that from a specialized somatic cell. The resulting cell is thereby "reprogrammed" to form a blastocyst with the genetic identity of the original somatic cell. This technique is the basis for cloning in animals and is being investigated for "therapeutic cloning," which involves creating cell banks that are genetically identical to the somatic cell donor for potential therapeutic use. It also is controversial because of the potential to be used for "reproductive cloning," which involves recreation of an entire organism based on the original reprogrammed cell.

A recent exciting discovery is the ability to reprogram cells by transfecting a few key regulatory genes (as opposed to the entire nuclear material). The resulting cells are termed *inducible pluripotent stem (iPS) cells* to reflect the fact that they have been induced to revert to a pluripotent state. In this technique, a specific set of two to six genes is introduced into the nucleus of a somatic cell. The gene set is chosen because of its ability to reprogram the target cell to a near-embryonic state, and currently there are several different gene sets that can achieve this goal. The resulting cell line then can theoretically be propagated and then differentiated into the desired specialized cell type. This technique has energized the stem cell research community, and the applications to tissue engineering are being investigated.

In the last decade there has been a large amount of research devoted to finding, characterizing, and using adult stem cells for tissue engineering and regenerative medicine. It is now recognized that there are a number of cell types in various tissues of the adult that

exhibit multipotent capabilities. The hematopoietic stem cells of the bone marrow have been widely studied and are known to create the cellular components of blood. The bone marrow also contains a rare population of so-called mesenchymal stem cells. These are isolated based on their ability to adhere to culture substrates, which distinguishes them from other cell types in the marrow, including hematopoietic stem cells. The source and fate of these cells is an active topic of discussion, and these cells are sometimes referred to as bone marrow stromal cells. A number of other adult tissue have also yielded multipotent progenitor and stem cells, including blood (endothelial progenitor cells), fat (adipose stem cells), brain (neural stem cells), and heart (cardiac stem cells). Putative stem cells also have been identified in the stroma of the umbilical cord, as well as in amniotic fluid. Not all researchers agree that each of these cell types is a true stem cell, but the potential for pluripotent and multipotent stem cells to impact tissue engineering is great.

6.2.5 Stem Cell Aging

Telomerases, DNA Stability, and Natural Cell Senescence

When linear DNA is replicated, the lagging strand is synthesized discontinuously through the formation of the so-called Okazaki fragments. The last fragment cannot be initiated, and therefore the lagging strand will be shorter than the leading strand. Linear chromosomes have noncoding repeating sequences on their ends that are called telomeres. These telomeres can be rebuilt using an enzyme called telomerase. Telomerase is a ribonucleoprotein DNA polymerase that elongates telomeres. When expressed, telomerase maintains the telomere length in growing cells. The telomere hypothesis implicates short telomere length and telomerase activation as critical players in cellular immortalization. This enzyme is active in microorganisms, in stem cells, and in transformed derivatives of stem cells (i.e., tumor cells). Commitment of stem cells toward their unipotent descendants results in the loss of telomerase activity. Thus, normal somatic cells lack this activity, and the telomeres are shortened by about 50 to 200 bp per replication. This shortening gives rise to the so-called mitotic clock. The length of the telomeres is about 9 to 11 kbp, and when it reaches about 5 to 7 kbp, the chromosomes become unstable and replication ceases (Figure 6.11). This mechanism is believed to underlie the Hayflick limit.

Telomerase activity is found in somatic hematopoietic cells but at a low activity level. There is evidence that telomeres in immature hematopoietic cells do shorten with ontogeny and with increased cell doublings in vitro. The rate of telomere shortening in stem cells is finite, but it may be slower than in other somatic cells. Numerical evaluation of the consequences of stem cell aging strongly suggests that there has to be some form of self-renewal of stem cells in adults (see Exercise 1).

6.2.6 Tissue Dynamics

Tissues are comprised of many different cell types of various developmental origins (Figure 6.12). The dynamic behavior of cells and their interactions determine overall tissue formation, state, and function. The activities of individual cells are often substantial.

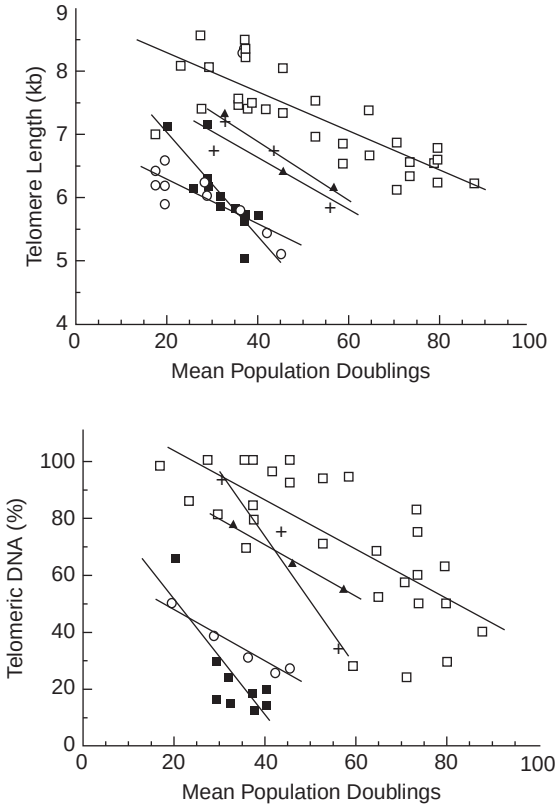


FIGURE 6.11 Primary experimental data showing the shortening of telomere length with increasing cellular doubling in cell culture. *From [5].*

However, the time scales that relate cellular activities with tissue function are relatively long, and therefore their importance tends to be overlooked. As examples, some cellular-level activities that underlie dynamic states of tissue function fall into the following categories:

1. *Hyperplasia:* Cell replication or proliferation resulting in an increase in cell number due to complete cell division (both DNA synthesis and cytokinesis).
2. *Hypertrophy:* DNA synthesis not accompanied by cytokinesis and instead results in an increase in the cell volume.
3. *Cell maturation and differentiation:* Changes in gene expression and the acquisition of particular functions, with the changes occurring in a maturationally lineage-dependent pattern.
4. *Cell apoptosis:* Aging cells undergoing "programmed cell death," a process distinguishable from necrosis.
5. *Cell adhesion:* The physical binding of a cell to its immediate environment, which may be a neighboring cell, an extracellular matrix, or an artificial surface.
6. *Cell motility:* The motion of a cell into a particular niche or location.

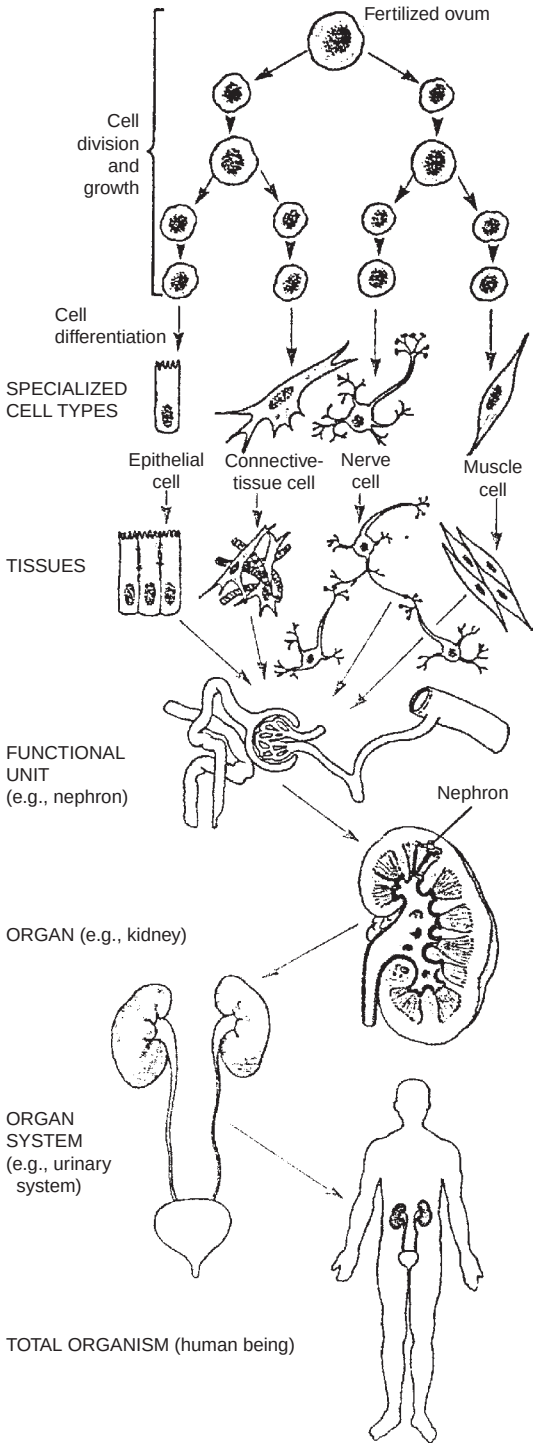


FIGURE 6.12 Levels of cellular organization in tissues and the diverse developmental origins of cells found in tissues. *From [17].*

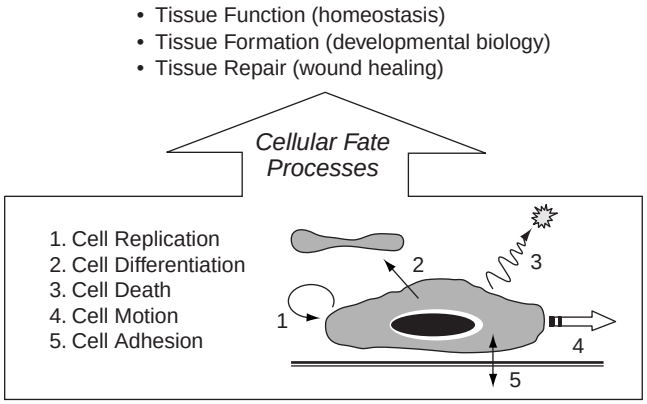


FIGURE 6.13 Tissue dynamics. The three dynamic states of tissues and the underlying cellular fate processes.

These processes are illustrated in Figure 6.13. What is known about each one of these processes will be briefly described in the following sections, with particular emphasis on quantitative and dynamic descriptions. The processes contribute to three dynamic states at a tissue level:

1. *Histogenesis:* The maturational lineages of cells derived from a tissue's stem cell compartment. A tissue's overall functions are the net sum of contributions from all the cells within the maturational lineages of a tissue.
2. *Tissue formation:* The formation of tissue has been characterized by studies comprising the field of developmental biology. Tissues vary in their proportion of stem cells, diploid cells, and polyploid cells, depending on age. The tissue, when isolated from young donors (e.g., infants) will have tissues with maturational lineages skewed toward the young cells in the lineage (stem cells, diploid somatic cells), whereas those from geriatric donors will have tissues skewed toward the later stages of the lineage (polyploid cells). This phenomenon is the explanation for why tissues procured from young donors have greater expansion potential ex vivo and are probably going to have greater potential for most forms of cell therapy programs.
3. *Tissue repair:* Repair of damaged tissues involves production of cells from the stem cell compartment, proliferation of the cells, and their differentiation into the fully mature cells. The repair process can also involve migration of cells, if need be, to a site of damage.

The preceding dynamic processes involve interplay among many different cell types. The cells communicate and coordinate their efforts through the principal cellular processes shown in Figure 6.13. The biology and dynamics of these processes are discussed in detail in Section 6.3.4.

Tissue Histogenesis

All tissues are dynamic. For instance, tissue dynamics can be illustrated by comparing the cell numbers that some organs produce over a lifetime to the total number of cells in the human body. As stated before, the human bone marrow produces about 400 billion myeloid cells daily in a homeostatic state. Over a 70-year lifetime, the cell production from

bone marrow accumulates to a staggering 10^{16} cells. This cell number is several hundred times greater than the total number of cells that are in the body at any given time. Similarly, the intestinal epithelium, the body's second most prolific tissue, produces about 5×10^{14} cells over a lifetime—ten times the total number of cells in the human body.

Tissues have their own characteristic turnover rates (Table 6.3). Bone marrow is the most proliferative tissue in the body, followed by the lining of the small intestine, and then by the epidermis. The turnover rate of these two tissues is on the order of a few days; that for the epidermis is in terms of weeks. The turnover in quiescent tissues is on the order of months to even years. For example, the turnover of the liver of rodents is estimated to be about one year; the turnover rate in the livers of humans is also slow. Even the tissue of the central nervous system is now known to turn over, though at a very slow rate, through the action of stem cells in compartments that line the ventricles of the brain.

Tissue Genesis

The preceding overview relates mainly to steady-state tissues that develop and regenerate through stem cell compartments and maturing lineages of cells. This steady state is achieved gradually during embryonic development and the process of organogenesis, which is quite complicated, as exemplified by hematopoiesis or the formation of blood cells. During vertebrate ontogeny, hematopoiesis sequentially occupies the yolk sac, fetal liver, spleen, and bone marrow. Variations in this pattern exist among vertebrate species. The earliest identification of hematopoietic cells is their assignment to the progeny of the C4 blastomere in the 32-cell embryo. The blastula grows to about 1,000 cells (ten doublings) and assumes a spherical shape. Then, the blastula undergoes gastrulation, not unlike

TABLE 6.3 Cell Renewable Rates in Tissues

Tissue	Species	Turnover Time (days)
Erythropoiesis	Rat	2.5
Myelopoiesis	Rat	1.4
Hematopoiesis	Human	2.5
Small intestinal epithelium	Human	4–6
	Rat	1–2
Epidermis	Human	7–100
Corneal epithelium	Human	7
Lymphatic cells	Rat (thymus)	7
	Rat (spleen)	15
Epithelial cells	Rat (vagina)	3.9
	Human (cervix)	5.7
Spermatogonia	Human	74
Renal interstitial cells	Mouse	165
Hepatic cells	Rat	400–500

pushing a finger into an inflated balloon. The point of invagination is the endoderm, which eventually forms the gut. After gastrulation, the ectoderm is brought into position relative to the endoderm and a third germ layer is formed between the two: the mesoderm. This middle layer is formed via cell-cell interactions and soluble growth factor action. Several determined stem cells originate from the mesoderm, including hematopoietic tissue, mesenchymal tissue, muscle, kidney, and notochord. Blood cells originate from the ventral mesoderm. Some hematopoietic cells migrate into the yolk sac to form blood islands consisting mainly of erythroid cells ("primitive" hematopoiesis). Intraembryonic hematopoiesis originates from the aortic region in the embryo and leads to "definitive" hematopoiesis. It appears that the embryonic origin of hematopoietic cells is from bipotent cells that give rise to both the vasculature (the endothelium) and hematopoietic cells. Hematopoietic stem cells are then found in the liver in the fetus. Around birth, the hematopoietic stem cells migrate from the liver into the bone marrow, where they reside during postnatal life. Interestingly, the umbilical cord blood contains hematopoietic stem cells capable of engrafting pediatric, juvenile, and small adult patients. This developmental process illustrates the asymmetric nature of stem cell division during development and increasing restriction in developmental potential. Furthermore, the migration of stem cells during development is important. Understanding the regulatory and dynamic characteristics of the stem cell fate processes is very important to tissue engineering.

Tissue Repair

When tissue is injured, a healing response is induced. The wound healing process is comprised of a coordinated series of cellular events. These events vary with ontological age. Fetal wound healing proceeds rapidly and leads to the restoration of scarless tissue. In contrast, postnatal healing is slower and often leads to scarring, which generally permits satisfactory tissue restoration, while not always fully restoring normal tissue structure. Some pathological states resemble wound healing. A variety of fibrotic diseases involve similar processes to tissue repair and subsequent scarring. The increasing appreciation of stem cell compartments and their descendant maturational lineages and the changes that occur in them with age are likely to provide an improved understanding of wound healing phenomena.

THE SEQUENCE OF EVENTS IN WOUND HEALING

Immediately following injury, control of bleeding starts with the rapid adhesion of circulating platelets to the site of damage. Within seconds, the platelets are activated, secrete contents from their storage granules, spread, and recruit more platelets to the thrombus that has started to develop. Within minutes of injury, the extent of hemorrhaging is contained through the constriction of surrounding blood vessels.

The next phase of the wound healing process involves the release of agents from the platelets at the injured site that cause vasodilatation and increased permeability of neighboring blood vessels. The clotting cascade is initiated and results in the cleavage of fibrinogen by thrombin to form a fibrin plug. The fibrin plug, along with fibronectin, holds the tissue together and forms a provisional matrix. This matrix plays a role in the early recruitment of inflammatory cells and later in the migration of fibroblasts and other accessory cells.

Inflammatory cells now migrate into the injured site. Neutrophils migrate from circulating blood and arrive early on the scene. As the neutrophils degranulate and die, the abundance

of macrophages at the site increases. All tissues have resident macrophages, and their number at the injury site is enhanced by macrophages migrating from circulation. They act in concert with the neutrophils to phagocytose cellular debris, combat any invading microorganisms, and provide the source of chemoattractants and mitogens. These factors induce the migration of endothelial cells and fibroblasts to the wound site and stimulate their subsequent proliferation. If the infiltration of macrophages into the wound site is prevented, the healing process is severely impaired.

The result of these initial processes is the formation of a so-called granulation tissue. It is comprised of a dense population of fibroblasts, macrophages, and developing vasculature that is embedded in a matrix comprised mainly of fibronectin, collagen, and hyaluronic acid. The invading fibroblasts begin to produce collagen, mostly types I and III. The collagen increases the tensile strength of the wound. Myofibroblasts actively contract at this time, shrinking the size of the wound by pulling the wound margins together.

Over time, the matrix then undergoes remodeling, which involves the coordinated synthesis and degradation of connective tissue proteins. Remodeling leads to a change in the composition of the matrix as healing progresses. For instance, collagen type III is abundant early on but gives way to collagen type I with time. The balance of these processes determines the degree of scar formation. Although the wound appears healed at this time, chemical and structural changes continue to occur within the wound site. The final step of the wound healing process is the resolution of the scar, though in most cases this process is incomplete and some form of scar tissue remains after healing. The formation and degradation of matrix components take place over many months. The healing process is essentially complete when the composition of the matrix and the spatial location of the cells have returned to close to the original state. Understanding the wound healing process is important to the tissue engineer, since the placement of disaggregated tissues in ex vivo culture induces responses reminiscent of the wound healing process.

6.2.7 Cell Differentiation

The coordinated activity of the cellular fate processes determines the dynamic state of tissue function (see Figure 6.13). There is growing information available about these processes in genetic, biochemical, and kinetic terms. The dynamics considerations that arise from the interplay of the major cellular fate processes are introduced at the end of the chapter, and the associated bioengineering challenges are described.

Describing Cellular Differentiation from a Biological Perspective

Differentiation is the process by which a cell undergoes phenotypic changes to become a particular specialized cell type. This specialized cell type is characterized by its physiological function and its corresponding role as part of a tissue and/or organ. This process begins with a lineage and differentiation commitment and is followed by a coordinated series of gene expression events.

The term *differentiation* is derived from differential gene expression. Differentiation involves a change in the set of genes that are expressed in the cell, and this change is usually an irreversible change toward a particular functional state. This process involves a carefully orchestrated switching off and on of gene families. The final set of genes expressed is those that pertain to the function of the mature cell.

EXAMPLE PROBLEM 6.1

How can the production of red blood cells be modeled?

Solution

Erythropoiesis replaces decaying mature red blood cells. About 200 billion of these cells need to be produced daily in a human adult. Many of the changes that a cell undergoes during the process of erythropoiesis are well known (Figure 6.14). The cell size, the rates of RNA and

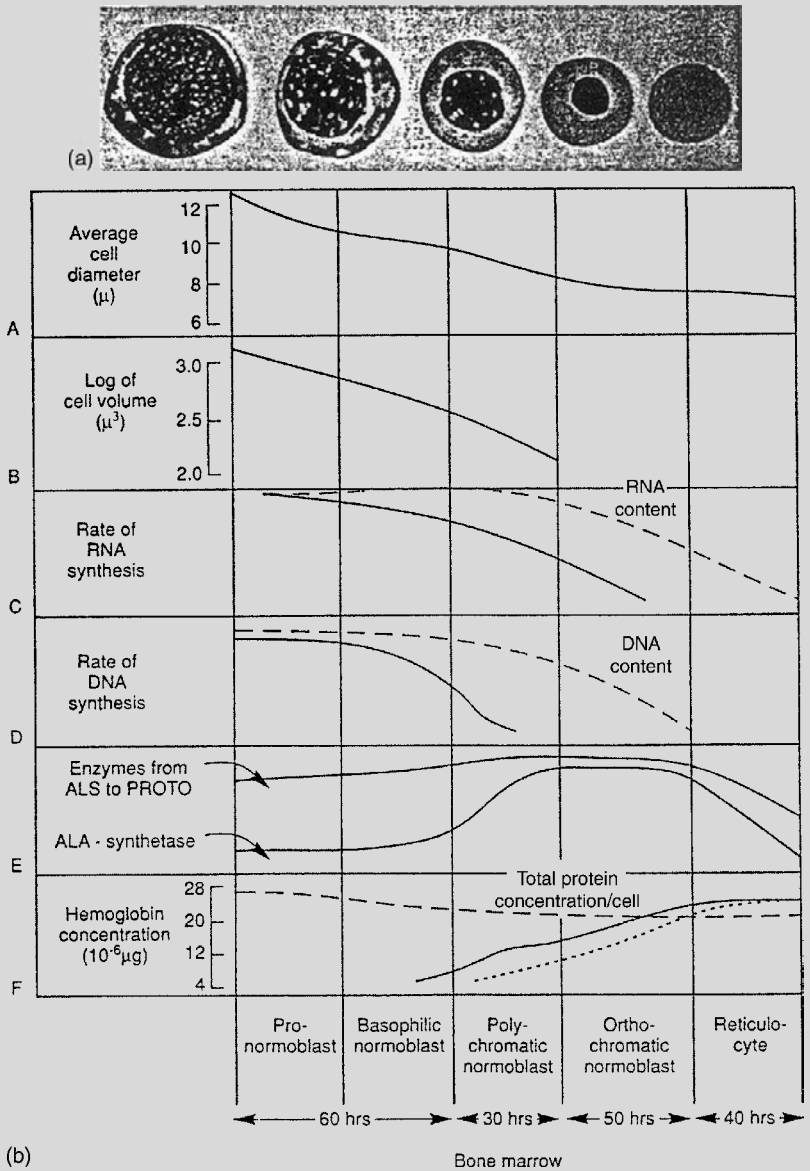


FIGURE 6.14 The erythroid maturation sequence. *From [4].*

Continued

DNA synthesis, and protein content all change in a progressive and coordinated fashion. The differentiation from a pronormoblast (earlier precursor stage) to a fully mature enucleated erythrocyte takes about 180 hours, or about one week. The replication activity is the highest at the preprogenitor and progenitor stages, but once the precursor stage is reached, replication activity ceases sharply. This information can be used to derive and solve equations that describe the process of erythropoeisis.

Experimental Observations of Differentiation

The process of differentiation can be observed directly using fluorescent surface markers and/or light microscopy for morphological observation. Ultrastructural changes are also used to define the stage of differentiation.

A flow cytometer is often used to monitor the process of cellular differentiation. The basis for this approach is the fact that characteristic surface proteins are found on cells at different stages of differentiation. These surface markers can be used as binding sites for fluorescently conjugated monoclonal antibodies. The flow cytometer can be used to trace the expression of several surface markers, and by performing such studies over time, it is possible to track the differentiated state of the cells and cell population.

For example, erythropoiesis can be traced based on expression of the transferrin receptor (CD71) and glycophorin-A. The latter is an erythroid-specific surface protein that is highly negatively charged and serves to prevent red cell aggregation in dense red cell suspensions. The transferring receptor plays a critical role during the stages in which iron is sequestered in hemoglobin. The measurement of this process is shown in Figure 6.15.

DESCRIBING THE KINETICS OF CELL DIFFERENTIATION

The process of differentiation is a slow one, often taking days or weeks to complete. The kinetics of this complex process can be described mathematically using two different approaches:

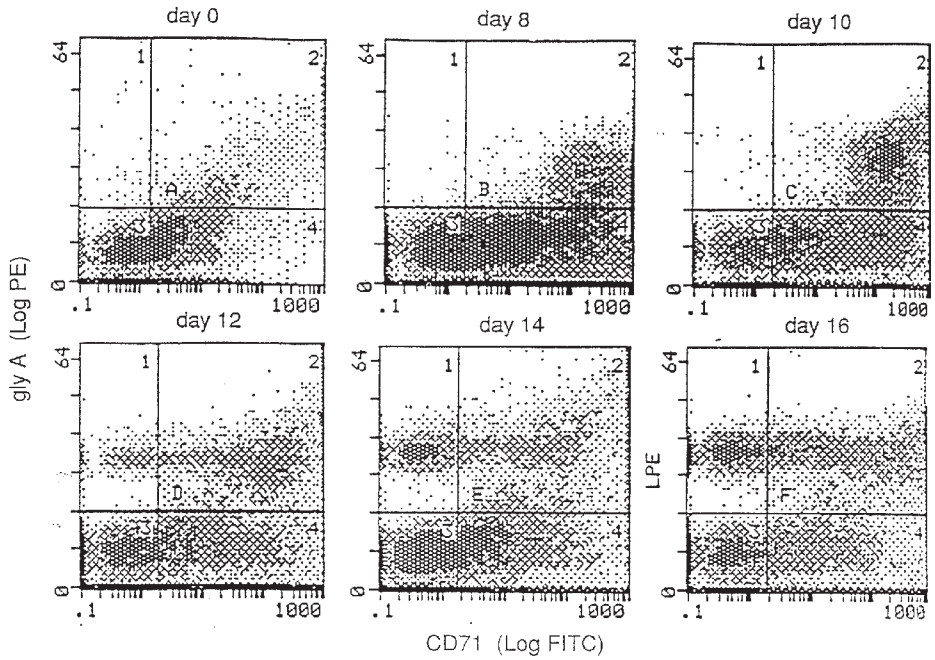
A. Compartmental models. The traditional approach to describing cell growth and differentiation is to use compartmental models. The differentiation process involves a series of changes in cell phenotype and morphology, typically becoming more pronounced at the latter stages of the process.

$$X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \ldots\ldots X_i \rightarrow \ldots\ldots X_n \rightarrow \textit{turnover} \tag{6.1}$$

where n can be as high as 16 to 18.

In the use of compartmental models in the past, the transition from one stage to the next was assumed to represent cell division. Thus, these models coupled the dual process of cell differentiation with cell replication. Mathematically, this model is described by a set of ordinary differential equations as

$$\frac{dX_i}{dt} = 2k_{i-1}X_{i-1} - k_iX_i \tag{6.2}$$



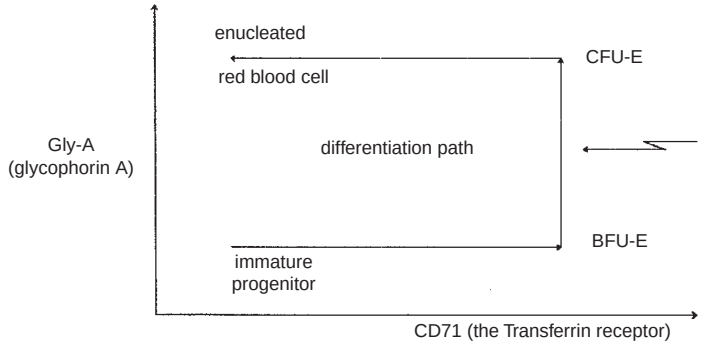


FIGURE 6.15 Two-parameter definition of erythropoietic differentiation. Glycophorin A is found on erythroid cells post the blast-forming unit-erythroid (BFU-E) stage, whereas transferrin (CD71) is expressed at the progenitor stage (BFU-E and colony-forming unit-erythroid (CFU-E)). By measuring the two simultaneously using a flow cytometer, this differentiation process can be traced as a U-shaped path on a bivariate dot plot. *From [16].*

The transition rate from one stage to the next is proportional to the number of cells present at that stage. This transition rate can clearly be a function of growth factor concentration and a number of other variables.

B. Differentiation as a continuous process. An alternative view is to consider the differentiation process to be a continuous process. Once the commitment to differentiation has been

made, the differentiation process is assumed to proceed at a fixed rate. This viewpoint leads to a mathematical description in the form of first-order partial differential equations:

$$\frac{dX}{dt} + \delta \frac{dX}{da} = (\mu(a) - \alpha(a))X \tag{6.3}$$

where δ is the rate of differentiation and a is a parameter that measures the differentiation state of the cell. μ and α are the growth and death rates, respectively, and vary between zero and unity. Both μ and α can be a function of a.

Cell Motion

THE BIOLOGICAL ROLES OF CELL MIGRATION

Cell migration plays an important role in all physiological functions of tissues and also some pathological processes. Cell migration is important during organogenesis and embryonic development. It plays a role in the tissue repair response in both wound healing and angiogenesis. The immune system relies on cell migration, and pathological situations like cancer metastasis are characterized by cell motility. Cell migration represents an integrated molecular process.

Animal cells exhibit dynamic surface extensions when they migrate or change shape. Such extensions, called lamellipodia and filopodia, are capable of dynamic formation and retraction. Local actin polymerization at the plasma membrane is a key process in the generation of these structures. These extensions are also regulated by a complex underlying process involving multiprotein interactions. In neurites, filopodia are believed to play a role in the progression of cell elongation by aiding the assembly of microtubules that are a significant component of these cells. The filopodia in neurites extend from the lamellipodial region and act as radial sensors. Filopodia on neurites have been found to be crucial to growth cone navigation. The filopodia on neurite growth cones have also been found to carry receptors for certain cell adhesion molecules. Mature leukocytes also can extend cytoplasmic extensions. Recently, structures termed uropods have been found on T lymphocytes. These cytoplasmic projections form during lymphocyte-endothelial interaction. There is a redistribution of adhesion molecules, including ICAM-1,-3, CD43, and CD44, to this structure. T cells have been found to use the uropods to contact and communicate directly with other T cells. Uropod development was promoted by physiologic factors such as chemokines. Cytoplasmic extensions, therefore, can perform a spectrum of functions in different cells that are related to migration and communication.

DESCRIBING CELL MOTION KINETICALLY

Whole populations. The motion of whole, nonreplicating cell populations can be described by

$$\frac{dX}{dx} = J \tag{6.4}$$

where J is the flux vector of the system boundary (cells/distance/time in a two-dimensional system), X is the cell number, and x is the flux dimension. The cellular fluxes are then related to cellular concentration and chemokine concentrations using

$$\begin{aligned} \mathrm{J} = &\ \text{random motility} + \text{chemokinesis} + \text{chemotaxis} \\ = &\ \sigma \frac{dX}{dx} - \left(\frac{X}{2}\right)\left(\frac{d\sigma}{da}\right)\left(\frac{da}{dx}\right) + \chi \frac{X da}{dx} \end{aligned} \tag{6.5}$$

where σ is the random motility coefficient, χ is the chemotactic coefficient, and a is the concentration of a chemoattractant. The first term is similar to a diffusion term in mass transfer and is a measure of the dispersion of the cell population. The chemokinesis term is a measure of the changes in cell speed with concentration and is normally negligible.

If $\chi = 0$ and chemokinesis is negligible, then the motion of the cells is represented by a random walk and is formally analogous to the process of molecular diffusion. The biased movement of the cells directly attributable to concentration gradients is given by the chemotaxis term, which is akin to convective mass transfer.

Individual cells. As described, motile cells appear to undergo a random walk process that mathematically can be described in a fashion similar to the diffusion process. Unlike the diffusion process, the motion of each moving entity can be directly determined (Figure 6.16). Thus, the motility characteristics of migrating cells can be measured on an individual cell basis.

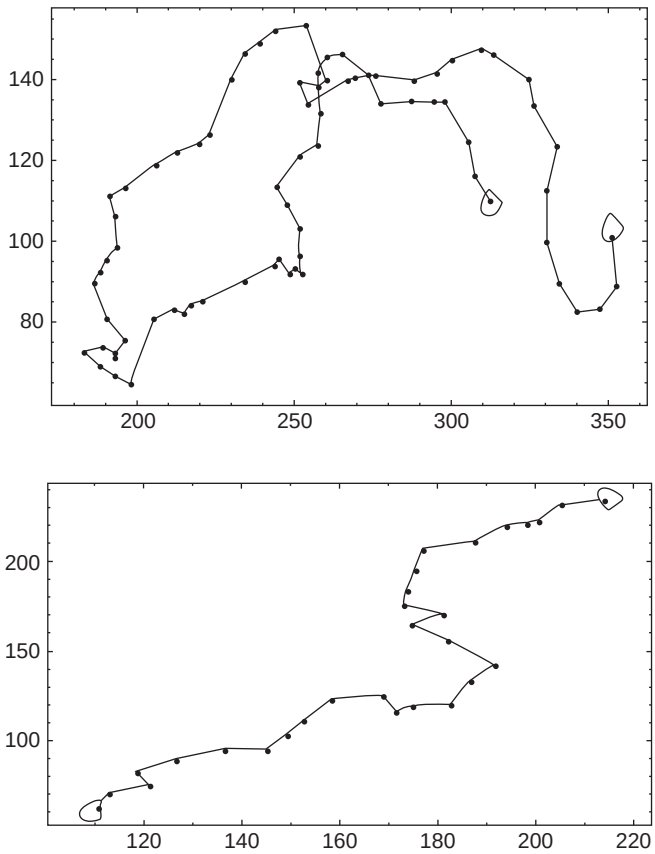


FIGURE 6.16 Experimental paths of individual neutrophil leukocytes undergoing random motility in uniform environments. *From [9].*

TABLE 6.4 Random Motion—Measured Cell Speeds and Persistence Times

Cell Type	Speed	Persistence Time
Rabbit neutrophils	20 μm/min	4 min
Rat alveolar macrophages	2 μm/min	30 min
Mouse fibroblasts	30 μm/h	1 h
Human microvessel endothelial cells	25–30 μm/h	4–5 h

Such a description is given in terms of the cell speed(s), the persistence time (p, the length of time that the cell moves without changing its direction), and the orientation bias (θ) that is due to action of chemoattractants. The random motility coefficient is related to these parameters as

$$\sigma = s^2 p \tag{6.6}$$

Typical values for these parameters are given in Table 6.4.

Cell Replication

THE CELL CYCLE

The process of cell division is becoming increasingly well understood in terms of molecular mechanisms (Figure 6.17). The so-called "cell cycle" of division is driven by a family of regulatory proteins known as cyclin-dependent kinases (cdk proteins). The cyclins exist in phosphorylated and dephosporylated states, basically a binary system. The network goes through a well orchestrated series of switches, during which the cyclins are turned off and on. When in an "on" state, they serve to drive the biochemical processes that are needed during that part of the cell cycle. It should be noted that there are several important decisions associated with moving "in and out of cycle," which is the process of a cell moving from a quiescent state (the so-called G_o) to a cycling (proliferative) state and vice versa. The cell cycle

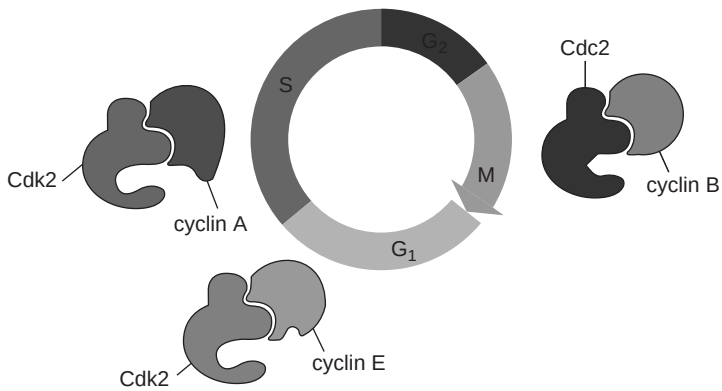


FIGURE 6.17 Schematic representation of the eukarotic cell cycle (G1–S–G2–M) and the presence of cyclin-dependent kinases.

is such that the time duration of the S, G_2, and M phases is relatively fixed. Once a cell determines it needs to and can divide, it initiates DNA synthesis and subsequent cell division. This process has the overall characteristics of a zero-order kinetic process. Once initiated, it proceeds at a certain rate. The minimal cycling time of human cells is about 12 hours. Progenitor cells can cycle at this rate, whereas more differentiated cell types tend to have longer cycle times.

DESCRIBING THE CELL CYCLE DYNAMICS

The dynamics of growth can be described in different ways.

1. *Exponential growth*. If growth is unconstrained, the rate of formation of new cells is simply proportional to the number of cells present

$$\frac{dX}{dt} = \mu X, \Rightarrow X(t) = X_o \exp(\mu t) \tag{6.7}$$

and exponential growth results. The growth rate μ is equal to $\ln(2)/t_d$, where t_d is the doubling time.

2. *Age-time structured descriptions.* If the phases of the cell cycle are to be described, then the status of the cell in the cycle needs to be incorporated in the dynamic description. This leads to first-order partial differential equations in time and cell cycle status:

$$\frac{dX}{dt} + v\frac{dX}{da} = \alpha(a)x \tag{6.8}$$

where v is the rate at which the cell moves through the cell cycle, a is a variable that describes the cell cycle status ($a = 0$ newborn cell, $a = 1$ cell completing mitosis), and α is the death rate of the cell that can be cell cycle dependent. This population balance equation can be solved under the appropriate initial and boundary conditions.

3. *Molecular mechanisms.* The cascade of cyclin-dependent kinases that constitute the molecular mechanism for cell cycling has largely been unraveled. Based on this knowledge, it is possible to describe the cell cycle in terms of the underlying molecular determinants. Such models have many components and typically need to be solved using computational methods.

Interacting Cellular Fate Processes Determine Overall Tissue Dynamics

The differentiation process involves a series of changes in cell phenotype and morphology that typically become more pronounced at the latter stages of the process. The key event is milieu-dependent differentiation (or differential gene expression), which encompasses the organogenic process that yields mature cells of a certain type and function. Similarly, embryonic induction can be described as a series of such events. This process is schematically presented in Eq. (6.1).

This progression of changes is typically coupled to fundamental "driving forces"—for example, cell cycling (mitosis) and cell death (apoptosis). Thus, the basic cellular processes can be accounted for in a population balance on each stage of differentiation:

$$\Delta\text{'s in cell\#} = \text{exit by input} - \text{differentiation} - \text{exit by apoptosis} + \text{entry by cell division}$$

or in mathematical terms:

$$\begin{aligned}\frac{dX}{dt} &= I - \delta X - \alpha X + \mu X \\ &= I - (\delta + \alpha - \mu)X\end{aligned} \tag{6.9}$$

where δ, α, and μ are the rates of differentiation, apoptosis, and replication, respectively. This equation can be rewritten as

$$\frac{dX}{dt} + kX = I, \tag{6.10}$$

where $k = \delta + \alpha - \mu$, $t = 1/k$. The parameter k is the reciprocal of the time constant t that characterizes the dynamics of changes in the number of cells of type X. It is therefore evident that the ratios of μ/δ, α/μ are the key dimensionless groups that determine the overall cell production in tissue.

6.2.8 Cellular Communications

How Do Cells Communicate?

Cells in tissues communicate with one another in three principal ways (Figure 6.18).

1. They secrete soluble signals, known as cyto and chemokines.
2. They touch each other and communicate via direct cell-cell contact.
3. They make proteins that alter the chemical microenvironment (ECM).

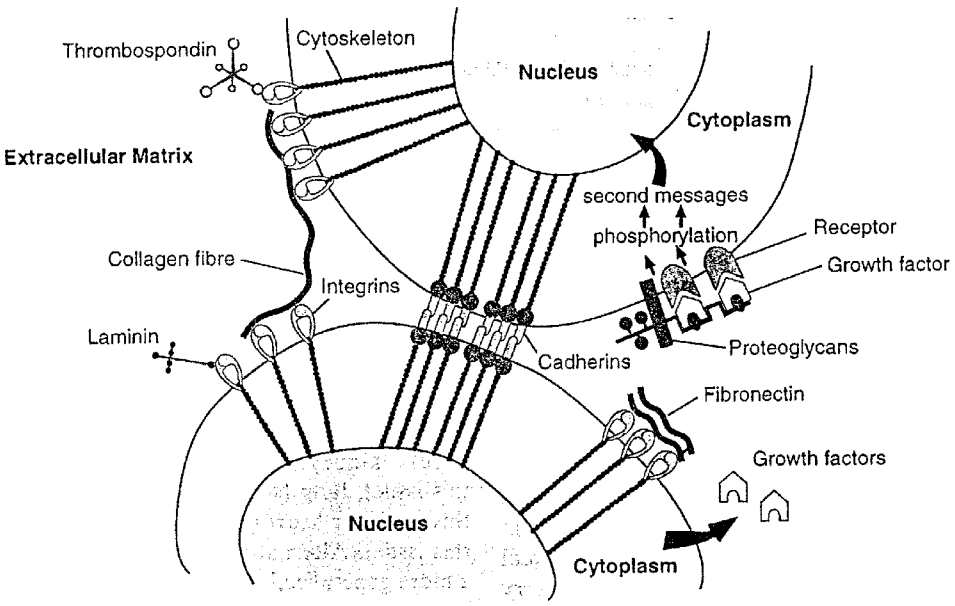


FIGURE 6.18 Cell and extracellular matrix (ECM) protein interactions. *From [13].*

These means of cellular communication differ in terms of their characteristic time and length scales, as well as in their specificity. Consequently, each is most suitable to convey a particular type of a message.

Soluble Growth Factors

Growth factors are small proteins that are on the order of 15,000 to 20,000 Dalton in size (one Dalton is the weight of the hydrogen atom and is a typical unit used to describe the size of molecules). They are relatively chemically stable and have long half-lives unless they are specifically degraded. Initially, growth factors were discovered as active factors that originated in biological fluids. For instance, erythropoetin was first isolated from urine and the colony stimulating factors from conditioned medium from particular cell lines. The protein could be subsequently purified and characterized. With the advent of DNA technology, growth factors can now be cloned directly as ligands for known receptors. This procedure was used to isolate thrombopoietin, as the *c-mpl* ligand and the stem cell factor as the *c-kit* ligand. Growth factors are produced by a signaling cell and secreted to reach a target cell.

EXAMPLE PROBLEM 6.2

What is the maximal secretion rate of protein?

Solution

The maximal secretion rate of a protein from a single gene in a cell can be estimated based on the maximal rates of transcription and translation. Such estimates have been carried out for the production of immunoglobulins (MW = 150,000 D), whose genes are under the control of strong promoters. The estimate shows that the maximal secretion rate is on the order of 2,000 to 8,000 protein molecules per cell per second, which corresponds to about 1 pg per cell per hour. This estimate compares favorably with measured maximal secretion rates. Since growth factors tend to be about one-tenth the size of immunoglobulin, a correspondingly higher secretion rate in terms of molecules per cell per time would be expected, although the total mass would stay the same. The secretion rates of protein from cells are expected to be some fraction of this maximum rate, since the cell is making a large number of proteins at any given time.

Growth factors bind to their receptors, which are found in cellular membranes, with high affinities. Their binding constants are as low as 10 to 100 pM. The binding of a growth factor to a receptor is described as

$$G + R \Leftrightarrow G{:}R, \quad \text{and} \quad \frac{[G][R]}{[G{:}R]} = K_a \tag{6.11}$$

where [G], [R], and [G:R] are the concentrations of the growth factor receptor and the bound complex, respectively, and K_a is the binding constant. Since the total number of receptors (R_{tot}) in the system is constant, we have the mass conservation quantity:

$$R_{tot} = [R] + [G{:}R] \tag{6.12}$$

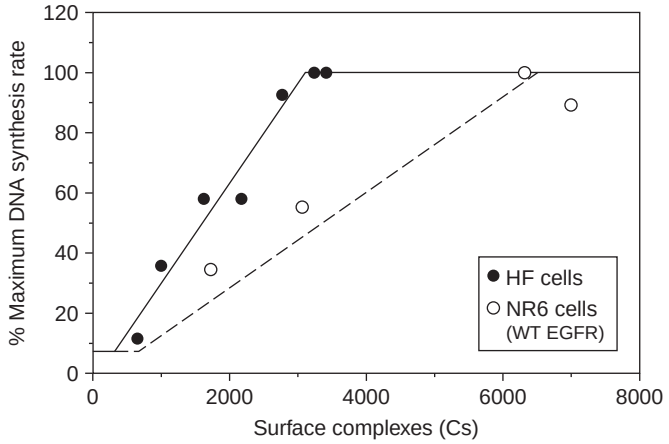


FIGURE 6.19 Relationship between steady-state surface complexes and mitogenic response to EGF for NR6 cells (○) and human fibroblasts (●). *Data from [7].*

and therefore Eq. (6.12) can be written as

$$R_{tot} = [G{:}R]\frac{K_a}{[G]} + [G{:}R] = [G{:}R]\left(1 + \frac{K_a}{[G]}\right) \tag{6.13}$$

or

$$([\mathrm{G{:}R}]/\mathrm{R_{tot}}) = [\mathrm{G}]/([\mathrm{G}] + \mathrm{K_a}) \tag{6.14}$$

Receptor occupancy rates ($[\mathrm{G{:}R}]/\mathrm{R_{tot}}$) need to be on the order of 0.25 to 0.5 to reach a significant stimulation (Figure 6.19), and therefore growth factor concentrations as low as 10 pM are sufficient to generate cellular response.

In many cases, the receptor:ligand complex is internalized, and a typical time constant for internalization is 15 to 30 minutes. The absolute values for growth factor uptake rates have been measured. For instance, interleukin-3 (IL-3) and the stem cell factor (SCF) are consumed by immature hematopoietic cells at rates of about 10 and 100 ng per million cells per day. Further, it has been shown that 10,000 to 70,000 growth factor molecules need to be consumed to stimulate cell division in complex cell cultures.

EXAMPLE PROBLEM 6.3

How far can soluble signals propagate and how long does it take?

Solution

The maximum signaling distance can be estimated from a simple diffusion model that describes secretion from a sphere. Under steady-state conditions, it can be shown that the concentration of a secreted molecule as a function of the distance from the cell is described as

$$\frac{c}{K_a} = \alpha\frac{R}{r} \quad \textit{where} \quad \alpha = \frac{(R^2/D)}{(K_a R/F)} \tag{6.15}$$

where R is the radius of the cell, F is the secretion rate, and D is the diffusion coefficient of the growth factor. The distance that a signal reaches when $c = K_a$ is estimated by

$$r_{critical}/R = \alpha \tag{6.16}$$

Thus, the maximal secretion rate is given by the ratio of two time constants α, the time constant for diffusion away from the cell (R^2/D), and the secretion time constant (K_aR/F). Since it takes an infinite amount of time to reach a steady state, a more reasonable estimate of the signal propagation distance is the time at which the signal reaches one-half of its ultimate value. This leads to a time constant estimate for the signaling process of about 20 minutes and a maximal distance of about 200 μm. This distance is shortened proportionally to the secretion rate (F) and inversely proportionally to the affinity (K_a), both of which are limited.

The binding of a growth factor to its receptor triggers a complex signal transduction process (Figure 6.20). Typically the receptor complex changes in such a way that its intracellular components take on catalytic activities. These activities are in many cases due to

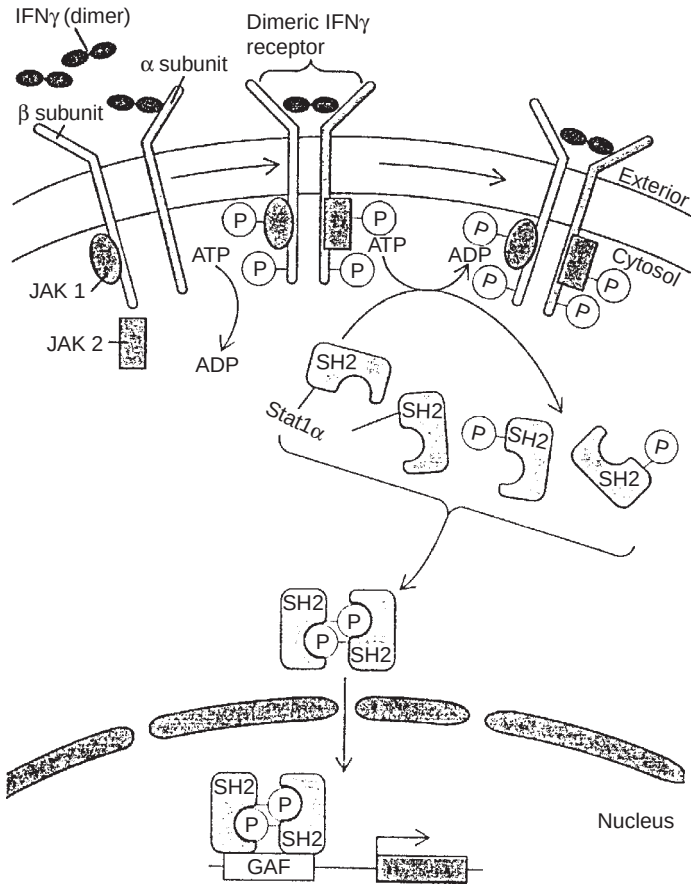


FIGURE 6.20 A schematic representation of the interferon gamma signal transduction pathway. *From [10].*

the so-called JAKs (Janus kinases). These kinases then operate on the STATs (signal transducers and activators of transcription) that transmit the signal into the nucleus. The kinetics of this process are complex, and detailed analyses are becoming available for the epidermal growth factor and other signal transduction processes.

Direct Cell-Cell Contact: Insoluble Factors 1

Cells are equipped with proteins on their surface called cell adhesion molecules (CAMs). These include the cadherins (adhesion belts, desmosome) and connexins (gap junctions). These molecules are involved in direct cell-to-cell contact. Some CAMs are known as the cell junction molecules, since junctions are formed between the adjacent cells, allowing for direct cytoplasmic communication. Such junctions are typically on the order of 1.5 nm in diameter and allow molecules below about 1,000 D to pass between cells.

A growing body of literature is showing how fluid mechanical shear forces influence cell and tissue function. Tissue function has a significant mechanical dimension. At the cellular level, the mechanical role of the cytoskeleton is becoming better understood. Signals may be delivered to the nucleus by cellular stretching in a way that is similar to the method in which growth factor binding to a receptor delivers signals. The cell-surface integrin receptors thus can perform as "mechanical transducers" of important signals. Further, cells have specific locations within a tissue, and long-distance information must be transmitted between cells via weak mechanical interactions. The mechanical characteristics of the cellular microenvironment are thus of importance, as are the mechanical properties of the cells themselves. The study of mechanically induced changes in cell function and mechanotransduction of cell signals is a rapidly growing area with great relevance to tissue engineering.

Cell-Matrix Interactions: Insoluble Factors 2

The extracellular matrix (ECM) is composed of a complex weave of fibers, struts, and gels that fills the spaces between cells in a tissue and interconnects the cells and their cytoskeletal elements. The ECM is multifunctional and provides tissue with mechanical support and cells with a substrate on which to migrate, as well as a place to locate signals for communications. The ECM thus has structural as well as instructional functions. It is dynamic and is constantly being modified. For instance, ECM components are degraded by metalloproteases and can be regenerated via cellular production. Many tissues remodel their matrix to some degree. For example, in cardiac muscle about 3 percent of the matrix is turned over daily.

The ECM is comprised of a large number of components that have varying structural and regulatory functions. On the cell surface, there are a number of adhesion and ECM receptor molecules that facilitate cell-ECM communications. These signals contain instructions for migration, replication, differentiation, and apoptosis. The nature of these signals is governed by the composition of the ECM, which in turn can be altered by the cells found in the tissue. Thus, many cellular fate processes can be directed by the ECM, and it provides a means for cells to communicate. The signals in the ECM are more stable and can be made more specific and stronger than those delivered by diffusible growth factors. The components of the ECM and their functions are summarized in Table 6.5.

TABLE 6.5 Components of the Extracellular Matrix[a]

Component	Function	Location
Collagens	Tissue architecture, tensile strength	Ubiquitously distributed
	Cell–matrix interactions	
	Matrix–matrix interactions	
Elastin	Tissue architecture and elasticity	Tissues requiring elasticity (e.g., lung, blood vessels, heart, skin)
Proteoglycans	Cell–matrix interactions	Ubiquitously distributed
	Matrix–matrix interactions	
	Cell proliferation	
	Binding and storage of growth factors	
Hyaluronan	Cell–matrix interactions	Ubiquitously distributed
	Matrix–matrix interactions	
	Cell proliferation	
	Cell migration	
Laminin	Basement membrane component	Basement membranes
	Cell migration	
Epiligrin	Basement membrane component (epithelium)	Basement membranes
Entactin (nidogen)	Basement membrane component	Basement membranes
Fibronectin	Tissue architecture	Ubiquitously distributed
	Cell–matrix interactions	
	Matrix–matrix interactions	
	Cell proliferation	
	Cell migration	
	Opsonin	
Vitronectin	Cell–matrix interactions	Blood
	Matrix–matrix interactions	Sites of wound formation
	Hemostasis	
Fibrinogen	Cell proliferation	Blood
	Cell migration	Sites of wound formation
	Hemostasis	
Fibrillin	Microfibrillar component of elastic fibers	Tissues requiring elasticity (e.g., lung, blood vessels, heart, skin)

Continued

TABLE 6.5 Components of the Extracellular Matrix[a]—Cont'd

Component	Function	Location
Tenascin	Modulates cell–matrix interaction	Transiently expressed associated with remodeling matrix
	Antiadhesive	
	Antiproliferative	
SPARC[b] (osteonectin)	Modulates cell–matrix interaction	Transiently expressed associated with remodeling matrix
	Antiadhesive	
	Antiproliferative	
Thrombospodin	Modulates cell–matrix interaction	Platelet α granules
Adhesion molecules	Cell surface proteins mediating cell adhesion to matrix or adjacent cells	Ubiquitously distributed
	Mediators of transmembrane signals	
von Willebrand factor	Mediates platelet adhesion	Plasma protein
	Carrier for procoagulant factor VIII	Subendothelium

[a]*Mutsaers et al., 1997.*
[b]*SPARC, secreted protein acidic and rich in cysteine.*

EXAMPLE PROBLEM 6.4

How many receptor sites are needed for various cellular functions?

Solution

The RGD (arginine-glycine-aspartic acid) tripeptide binding sequence has been immobilized on a cell growth surface at varying densities. Cell attachment, spreading, and growth were examined as a function of the surface density of RGD binding sites for fibroblasts. The results showed that an average receptor spacing of 440 nm was sufficient for cell attachment and spreading, and 160 nm for focal point adhesion formation. This type of experiment and analysis can provide insight into how cells adhere and migrate on different substrates.

A main thrust of tissue engineering efforts has been the design and fabrication of materials that mimic the ECM. These matrices can be used as scaffolds to support tissue growth. A variety of natural and synthetic materials have been used for this purpose. In some cases these materials are designed to be bioresorbable to allow the cells to replace the supplied ECM as they establish themselves and reconstruct tissue function. It is challenging to design such matrices because many of the functions of the ECM are still incompletely understood and involve two-way communication between cells and the matrix. The full spectrum of ECM functionality is perhaps something that can only be provided by the cells themselves (Table 6.6).

TABLE 6.6 Adhesion Molecules with the Potential to Regulate Cell-ECM Interactions

Adhesion Molecule	Ligand
Integrins	
$\alpha_1\beta_1$	Collagen (I, IV, VI), laminin
$\alpha_2\beta_1$	Collagen (I–IV, VI), laminin
$\alpha_3\beta_1$	Collagen (I), laminin, fibronectin, entactin, epiligrin
$\alpha_4\beta_1$	$\text{Fibronectin}_{\text{ALT}}$, VCAM-1, thrombospondin
$\alpha_5\beta_1$	Fibronectin, thrombospondin
$\alpha_6\beta_1$	Laminin
$\alpha_V\beta_1$	Fibronectin
$\alpha_L\beta_2$	ICAM-1, ICAM-2, ICAM-3
$\alpha_M\beta_2$	ICAM-1, iC3b, fibrinogen, factor X, denatured protein
$\alpha_x\beta_2$	Fibrinogen, iC3b, denatured protein
$\alpha_V\beta_3$	Vitronectin, fibrinogen, fibronectin, thrombospondin
$\alpha_V\beta_5$	Vitronectin
$\alpha_V\beta_6$	Fibronectin
$\alpha_0\beta_4$	Laminin
$\alpha_4\beta_7$	$\text{Fibronectin}_{\text{ALT}}$, VCAM-1, MAdCAM-1
$\alpha_{1\text{th}}\beta_3$	Fibrinogen, fibronectin, vitronectin, vWF
LRI^h	Fibrinogen, fibronectin, vitronectin, vWF, collagen (IV), entactin

6.3 PHYSICAL CONSIDERATIONS

6.3.1 Organization of Tissues into Functional Subunits

The body has 11 major organ systems (Table 6.7), with the muscular and skeletal systems often considered together as the musculoskeletal system. These organ systems carry out major physiological functions, such as respiration, digestion, circulation, and mechanical motion. Each one of these organ systems in turn is comprised of a set of tissues and organs. The major organs that participate in digestion are shown in Figure 6.21. Each organ system and organ has homeostatic functions that can be defined based on their physiological requirements. These can be thought of as providing the specifications that define an organ's function and provide the parameters needed to design the tissue. An example of such a "spec sheet" is given in Table 6.8.

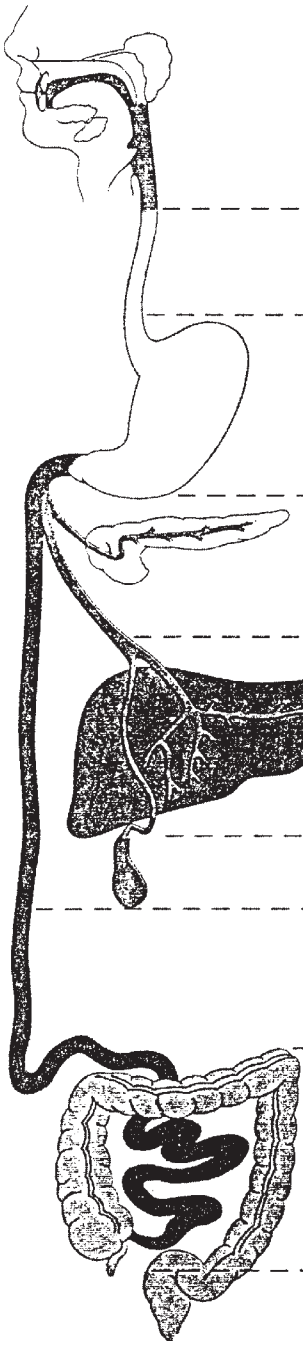
There are several important conclusions that can be arrived at using a simple order of magnitude analysis of the information found in such spec sheets. Insightful and judicious order of magnitude analysis is an important mode of analysis of tissue function, and it

TABLE 6.7 Major Organ Systems of the Body

Circulatory	Heart, blood vessels, blood (some classifications also include lymphatic vessels and lymph in this system)	Transport of blood throughout the body's tissues
Respiratory	Nose, pharynx, larynx, trachea, bronchi, lungs	Exchange of carbon dioxide and oxygen; regulation of hydrogen–ion concentration
Digestive	Mouth, pharynx, esophagus, stomach, intestines, salivary glands, pancreas, liver, gallbladder	Digestion and absorption of organic nutrients, salts, and water
Urinary	Kidneys, ureters, bladder, urethra	Regulation of plasma composition through controlled excretion of salts, water, and organic wastes
Musculoskeletal	Cartilage, bone, ligaments, tendons, joints, skeletal muscle	Support, protection, and movement of the body; production of blood cells
Immune	Spleen, thymus, and other lymphoid tissues	Defense against foreign invaders; return of extracellular fluid to blood; formation of white blood cells
Nervous	Brain, spinal cord, peripheral nerves and ganglia, special sense organs	Regulation and coordination of many activities in the body; detection of changes in the internal and external environments; states of consciousness; learning; cognition
Endocrine	All glands secreting hormones: Pancreas, testes, ovaries, hypothalamus, kidneys, pituitary, thyroid, parathyroid, adrenal, intestinal, thymus, heart, pineal	Regulation and coordination of many activities in the body
Reproductive	Male: Testes, penis, and associated ducts and glands	Production of sperm; transfer of sperm to female
	Female: Ovaries, uterine tubes, uterus, vagina, mammary glands	Production of eggs; provision of a nutritive environment for the developing embryo and fetus; nutrition of the infant
Integumentary	Skin	Protection against injury and dehydration; defense against foreign invaders; regulation of temperature

can be combined with experimental studies to characterize and set organ function parameters. Even with our current advanced state of biological knowledge, detailed theoretical analysis and calculations cannot substitute for focused and well-justified experimental work.

Organs function at a basal rate but have the ability to respond to stress. For example, the total circulation rate and organ distribution under strenuous exercise differs significantly from that at rest. Similarly, under hematopoietic stress, such as infection or in patients with sickle cell anemia, the basal blood cell production rate can significantly exceed the basal rate given in Figure 6.3.



ORGAN	EXOCRINE SECRETIONS	FUNCTIONS
Mouth and pharynx		Chewing (mechanical digestion); initiation of swallowing reflex
Salivary glands	Salt and water	Moisten food
	Mucus	Lubrication
	Amylase	Polysaccharide-digesting enzyme
Esophagus		Move food to stomach by peristaltic waves
	Mucus	Lubrication
Stomach		Store, mix, dissolve and begin digestion of food; regulate emptying of dissolved food into small intestine
	HCl	Solubilization of food particles: kill microbes
	Pepsin	Protein-digesting enzyme
	Mucus	Lubricate and protect epithelial surface
Pancreas		Secretion of enzymes and bicarbonate: also has nondigestive endocrine functions
	Enzymes	Digest carbohydrates, fats, proteins, and nucieic acids
	Bicarbonate	Neutralize HCl entering small intestine from stomach
Liver		Secretion of bile; many other nondigestive functions
	Bile salts	Solubilize water-insoluble fats
	Bicarbonate	Neutralize HCl entering small intestine from stomach
	Organic waste products and trace metals	Elimination in feces
Gallbladder		Store and concentrate bile between meals
Small intestine		Digestion and absorption of most substances: mixing and propulsion of contents
	Enzymes	Food digestion
	Salt and water	Maintain fluidity of luminal contents
	Mucus	Lubrication
Large intestine (colon)		Storage and concentration of undigested matter: mixing and propulsion of contents
	Mucus	Lubrication
Rectum		Defecation

FIGURE 6.21 Functions and organization of the gastrointestinal organs. *From [17].*

TABLE 6.8 Standard American Male[a]

Age	30 years
Height	5 ft 8 in or 1.86 m
Weight	150 lb or 68 kg
External surface	19.5 ft^2 or 1.8 m^2
Normal body temperature	37.0°C
Normal mean skin temperature	34°C
Heat capacity	0.86 cal/(g) (°C)
Capacities	
Body fat	10.2 kg or 15%
Subcutaneous fat layer	5 mm
Body fluids	ca.51 liters or 75%
Blood volume	5.0 liters (includes formed elements, primarily red cells, as well as plasma)
	Hematrocrit = 0.43
Lungs	
Total lung capacity	6.0 liters
Vital capacity	4.2 liters
Tidal volume	500 ml
Dead space	150 ml
Mass transfer area	90 m^2
Mass and energy balances at rest	
Energy conversion rate	72 kcal/h or 1730 kcal/day [40 kcal/(m^2)(h)]
O_2 consumption	250 ml/min (respiratory quotient = 0.8)
CO_2 production	200 ml/min
Heart rate	65/min
Cardiac output	5.01/min (rest)
	3.0 + 8 M in general (M = O_2 consumption in liters per minute)
Systemic blood pressure	120/80 mmHg

[a]*From Lightfoot (1974).*

Organs are comprised of functional subunits. Clear examples of such subunits include the alveoli in the lung and the nephron in the kidney (Figure 6.22). These functional units are comprised of a mixture of different kinds of cells that together constitute tissue function. Separating the functional subunits into their individual cell components often leads to the loss of tissue-specific function, but specific cell properties can be studied using such purified preparations.

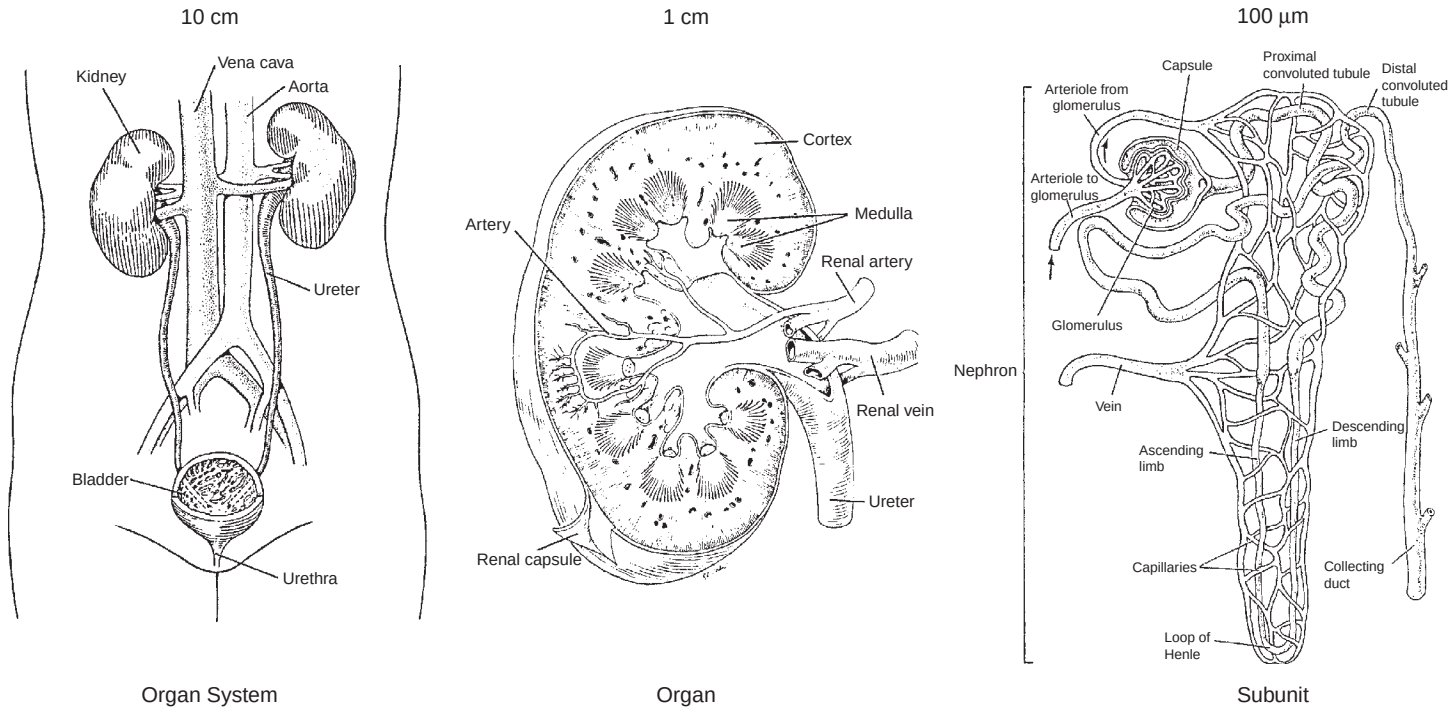


FIGURE 6.22 An organ system, an organ, and a functional subunit.

This observation leads to the conclusion that tissue function is a property of cell-cell interactions. The size of the functional subunits is on the order of 100 μm, whereas the size scale of a cell is 10 μm. Each organ is then comprised of tens to hundreds of millions of functional subunits. The sizing of organs represents an evolutionary challenge that is also faced by tissue engineering in scaling up the function of reconstituted tissues ex vivo.

The tissue and organ microenvironment is thus very complex. To achieve proper reconstitution of organ function, these dynamic, chemical, and geometric variables must be accurately replicated. This is a challenging task, and the following sections are largely devoted to developing quantitative methods to describe the microenvironment. These methods can then be used to develop an understanding of key problems, formulation of solution strategy, and analysis for its experimental implementation.

The microcirculation connects all the microenvironments in every tissue to the larger "whole body" environment. With few exceptions, essentially all metabolically active cells in the body are located within a few hundred μm from a capillary. The capillaries provide a perfusion environment that connects every cell (the cell at the center of Figure 6.22) to a source of oxygen and sink for carbon dioxide (the lungs), a source of nutrients (the small intestine), the clearance of waste products (the kidney), and so forth. The engineering of these functions ex vivo is a main focus of bioreactor design. Such culture devices have to appropriately simulate or substitute for respiratory, gastrointestinal, and renal functions. Further, these cell culture devices have to allow for the formation of microenvironments and thus must have perfusion characteristics that allow for uniformity down to the 100-micron-length scale. These are stringent design requirements.

6.3.2 Estimating Tissue Function from "Spec Sheets"

Most analysis in tissue engineering is performed with approximate calculations and estimations based on physiological and cell biological data—a tissue spec sheet, so to speak (see Table 6.8). These calculations are useful to interpret organ physiology, and they provide a starting point for an experimental program. Some examples follow.

The respiratory functions of blood: Remarkably insightful calculations leading to interpretation of the physiological respiratory function of blood have been carried out. The basic functionalities and biological design challenges can be directly derived from tissue spec sheets. Blood needs to deliver about 10 mM of O_2 per minute to the body. The gross circulation rate is about 5 liters per minute. Therefore, blood has to deliver to tissues about 2 mM oxygen per liter during each pass through the circulation. The pO_2 of blood leaving the lungs is about 90 to 100 mmHg, while pO_2 in venous blood at rest is about 35 to 40 mmHg. During strenuous exercise, the venous pO_2 drops to about 27 mmHg. These facts state the basic requirements that circulating blood must meet to deliver adequate oxygen to tissues.

The solubility of oxygen in aqueous media is low and can be represented by

$$[O_2] = \alpha_{O2} pO_2 \tag{6.17}$$

where the Henry's law coefficient is about 0.0013 mM per mmHg. The oxygen that can be delivered with a partial pressure change of 95 – 40 = 55 mmHg is thus about 0.07 mM, far below the required 2 mM (by a factor of about 30-fold). Therefore, the solubility or oxygen content of blood must be substantially increased, and the concentration dependency

of the partial pressure must be such that the 2 mM are given up when the partial pressure changes from 95 to 40 mmHg. Furthermore, during strenuous exercise, the oxygen demand doubles and 4 mM must be liberated for a partial pressure change from 95 to 27 mmHg.

The evolutionary solution is to put an oxygen-binding protein into circulating blood to increase the oxygen content of blood. To stay within the vascular bed, such a protein would have to be 50 to 100 kD in size. With a single binding site, the required protein concentration for 10 mM oxygen is 500 to 1,000 g/l, which is too concentrated from an osmolarity standpoint, and the viscosity of such a solution may be 10-fold that of circulating blood, which is clearly impractical. Further, circulating proteases would lead to a short plasma half-life of such a protein.

Having four sites per oxygen-carrying molecule would reduce the protein concentration to 2.3 mM, and confining it to a red cell would solve both the viscosity and proteolysis problems. These indeed are the chief characteristics of hemoglobin. A more elaborate kinetic study of the binding characteristics of hemoglobin shows that positive cooperativity will give the desired oxygen transfer capabilities both at rest and under strenuous exercise (Figure 6.23).

Perfusion rates in human bone marrow cultures: The question of how often the growth medium that supports cells should be replenished is important in designing cell culture conditions. Normally, the in vivo situation provides a good starting point for experimental optimization. A dynamic similarity analysis can give insight into this question.

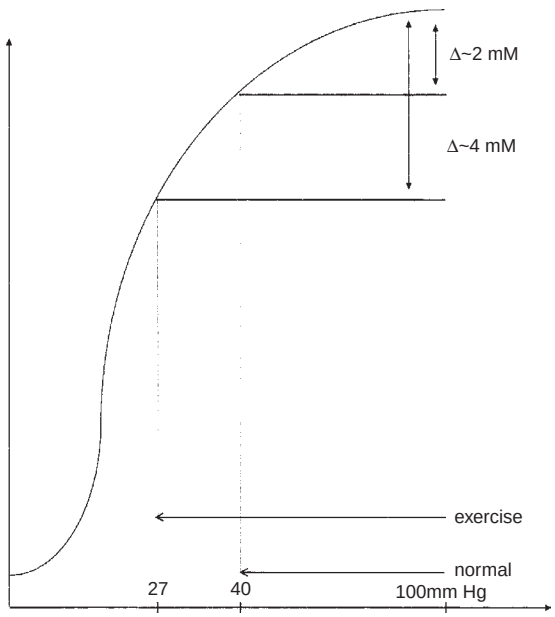


FIGURE 6.23 A schematic of hemoglobin–oxygen binding curves and oxygen delivery. The change in oxygen concentration during passage through tissues (i.e., oxygen delivery) is shown as a function of the concentration of oxygen in blood. *Modified from [3].*

The blood perfusion though bone marrow is about 0.08 ml per cc per minute. Cellularity in marrow is about 500 million cells per cc. Therefore, the cell-specific perfusion rate is about 2.3 ml/10 million cells/day. Cultures of mononuclear cell populations from murine bone marrow were developed in the mid- to late 1970s. These mouse cultures had long-term viability, but attempts to use the same culture protocols for human bone marrow cultures in the early 1980s were largely unsuccessful. The culture protocol called for medium exchange about once per week.

To perform a dynamic similarity analysis of perfusion rates, or medium exchange rates, between in vivo and in vitro conditions, the per cell medium exchange rate in culture is calibrated to that calculated for the preceding in vivo situation. Cell cultures are typically started with cell densities on the order of a million cells per ml. Therefore, 10 million cells would be placed in 10 mls of culture medium, which contains about 20 percent serum (vol/vol). A full daily medium exchange would hence correspond to replacing the serum at 2 ml/10 million cells/day, which is similar to the number calculated previously.

Experiments using this perfusion rate and the cell densities were performed in the late 1980s and led to the development of prolific cell cultures of human bone marrow. These cultures were subsequently scaled up to produce a clinically meaningful number of cells and are currently undergoing clinical trials. Thus, a simple similarity analysis of the in vivo and in vitro tissue dynamics led to the development of culture protocols that are of clinical significance. Such conclusions can be derived from tissue spec sheets.

These examples serve to illustrate the type of approximate calculations that assist the tissue engineering in performing an analysis of a tissue or organ system. Accurate measurements or estimations and well-organized facts (the spec sheets) provide the basic data. Characteristic time constants, length constants, fluxes, rates, concentrations, and so on can then be estimated. The relative magnitudes of such characteristics serve as a basis for order of magnitude judgments, which in turn inform the design of a tissue as well as the process for fabricating that tissue.

6.3.3 Mass Transfer in 3-D Configurations

An understanding of mass transfer in biological systems is essential in tissue engineering in order to design and deliver cell therapy products or extracorporeal organs. Although biological and physiological spec sheets give a global target for mass transfer, the details of the capillary bed of the smallest physiological units of the tissue are required. Mass transfer depends on the diffusion and convection of nutrients and waste to and from tissue, and the consumption of nutrients and production of waste by the tissue. Table 6.9 shows typical convection of blood in the vasculature, and Table 6.12 and Figure 6.29 give the consumption and diffusion rates of nutrients in the human body. Convection is driven by pressure differences and dominates in the vasculature, while diffusion is driven by concentration gradients and dominates in the tissues. Diffusion can be described by Fick's Law, and one can modify this equation to describe mass transfer in three basic configurations: rectangles, cylinders, and slabs.

The key to a successful bioreactor design for extracorporeal support or expansion of stem cells is maintaining adequate mass transfer while at the same time providing a local

TABLE 6.9 Peak Convection Rates of Blood in the Vasculature

Compartment	cm/sec
Aorta	140 ± 40
Common carotid	100 ± 20
Vertebral[a]	36 ± 9
Superficial femoral	90 ± 13
Liver sinusoid[b]	10^{-2}–10^{-3}

Data from DeWitt and Wechsler (1988).
[a]*Data from Jager et al. (1985).*
[b]*Data from McCuskey (1984).*

environment conducive to the differentiated state. The dominant mechanism of transport of low molecular weight nutrients (e.g., oxygen and glucose) within tissue or cell aggregates is diffusion. The length scale for diffusive transport (i.e., the distance over which oxygen penetrates into the tissue from nutrient stream before it is completely consumed by the cells) depends on the volumetric concentration of metabolite in the nutrient stream, C_0; the rate at which the cells consume the nutrient, Q_i; the diffusion coefficient for the metabolite in the tissue, D_t; and the system geometry. The metabolic consumption rate Q_i is generally a function of the nutrient concentration in the cell mass, C. The most common rate expression is the Michaelis-Menton type, which reduces to zero-order kinetics for high concentrations of nutrient (i.e., $C \gg K_m$) and to first-order kinetics for low nutrient concentrations (i.e., $C \ll K_m$). It is reasonable to consider the zero-order limit because the condition $C \gg K_m$ is met for many important nutrients under normal physiological conditions.

An estimate of the length scale for nutrient diffusion in a 3-D cell mass can be obtained by mathematical modeling. The general equation governing the balance between steady-state diffusion and metabolic consumption is

$$D_i \nabla^2 C = Q_i \tag{6.18}$$

where C is the concentration of the nutrient within the cell mass. Three simple geometries amenable to analytical solutions are shown in Figure 6.24: a slab, cylinder, or sphere of cells bathed in medium containing the nutrient at concentration C_0. Expanding the gradient operator for each of these geometries, Eq. (6.18) can be written as

$$\text{slab: } D_t \frac{d^2C}{dx^2} = Q_i \tag{6.19a}$$

$$\text{cylinder: } D_t \frac{1}{r}\frac{d}{dr}\left(\frac{rdC}{dr}\right) = Q_i \tag{6.19b}$$

$$\text{sphere: } D_t \frac{1}{r}\frac{d}{dr}\left(\frac{r^2dC}{dr}\right) = Q_i \tag{6.19c}$$

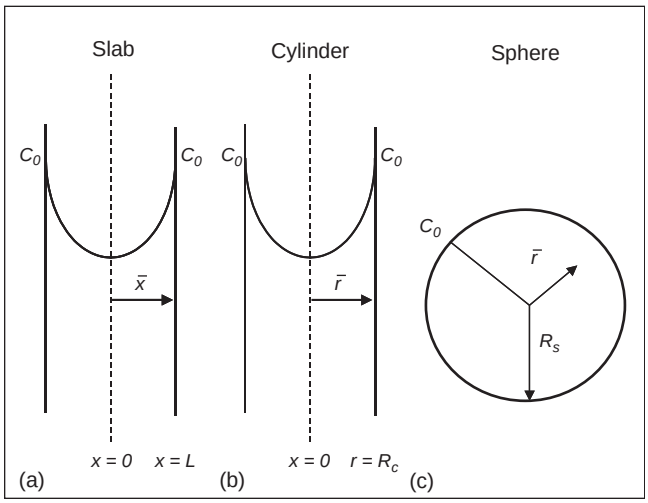


FIGURE 6.24 The three possible configurations considered by the mass transfer equations. *Reproduced from [12].*

The manner in which the distance variables are defined in each case is shown in Figure 6.24. A standard technique in the analysis of differential equations describing transport phenomena is to scale the variables so they are dimensionless and range in value from 0 to 1. This allows the relative magnitude of the various terms of the equations to be evaluated easily and allows the solutions to be plotted in a set of graphs as a function of variables that are universally applicable. For all geometries, C is scaled by its maximum value, C_o—that is, $\bar{C} = C/C_o$. Distance is scaled with the diffusion path length, so for a slab, $\bar{x} = x/L$; for a cylinder, $\bar{r} = r/R_c$; and for a sphere, $\bar{r} = r/R_s$. With these definitions, the boundary conditions for all three geometries are no flux of nutrient at the center $(d\bar{c}/d\bar{x},\, d\bar{c}/d\bar{r} = 0\ at\ \bar{x}, \bar{r}, = 0)$ and that $\bar{C} = 1$ at the surface $(\bar{x}, \bar{r}, = 1)$.

The use of scaling allows the solutions for all three geometries to collapse to a common form:

$$\bar{C} = 1 - \frac{\phi^2}{2(1 - \bar{x}^2)}$$

All of the system parameters are lumped together in the dimensionless parameter ϕ^2, which is often called the Thiele modulus. The Thiele modulus represents the relative rates of reaction and diffusion and is defined slightly differently for each geometry:

$$\phi^2_{slab} = \frac{Q_i/L^2}{C_o D_t}, \qquad \phi^2_{cylinder} = \frac{Q_i/Rc^2}{C_o D_t}, \qquad \phi^2_{sphere} = \frac{Q_i/(L^2 = (R/3)^2)}{C_o D_t}$$

6.3.4 The Tissue Microenvironment: Cell Therapy and Bioreactor Design Principles

A main task in tissue engineering is to recreate tissue function by constructing a tissue. Achieving this goal is challenging, and it is based on some fundamental axioms:

- The developmental program and the wound healing response require the systematic and regulated unfolding of the information on the DNA through coordinated execution of genetic subroutines and programs. Participating cells require detailed information about the activities of their neighbors. Proper cellular communications are of key concern and in many cases the concept of the "stem cell niche" is important.
- Upon completion of organogenesis or wound healing, the function of fully formed organs is strongly dependent on the coordinated function of multiple cell types. Tissues function based on multicellular aggregates, called functional subunits of tissues.
- The microenvironment has a major effect on the function of an individual cell. The characteristic length scale of the microenvironment is 100 μm.
- The microenvironment is characterized by (1) neighboring cells: cell-cell contact, soluble growth factors, and so on; (2) the chemical environment: the extracellular matrix, the dynamics of the nutritional environment; and (3) the local geometry.

Cellular Function In Vivo: The Tissue Microenvironment and Communication

An important requirement for successful tissue function is a physiologically acceptable environment in which the cells will express the desired tissue function. One common way to approach this goal is to attempt to recreate or mimic the physiological in vivo environment. This task involves mimicking a variety of features of the microenvironment including cell-cell and cell-matrix communications, the biochemical and mechanical milieu, as well as communication between tissues.

Examples of tissue microenvironments are shown in Figure 6.25, and some are discussed in Section 6.2.2. The communication of every cell with its immediate environment and other tissues can be examined using a topological representation of the organization of the body. The DNA in the center of the diagram contains the information that the tissue engineer wishes to express and manage. This cell is in a microenvironment that has important spatiotemporal characteristics.

Signals to the nucleus are delivered at the cell membrane and are transmitted through the cytoplasm by a variety of signal transduction mechanisms. Some signals are delivered by soluble growth factors after binding to the appropriate receptors. These growth factors may originate from the circulating blood or from neighboring cells. Nutrients, metabolic waste products, and respiratory gases traffic in and out of the microenvironment in a highly dynamic manner. The microenvironment is also characterized by its cellular composition, the ECM, and local geometry, and these components also provide the cell with important signals. The size scale of the microenvironment is on the order of 100 μm. Within this domain, cells function in a coordinated manner. If this arrangement is disrupted, appropriate cell function will not be obtained.

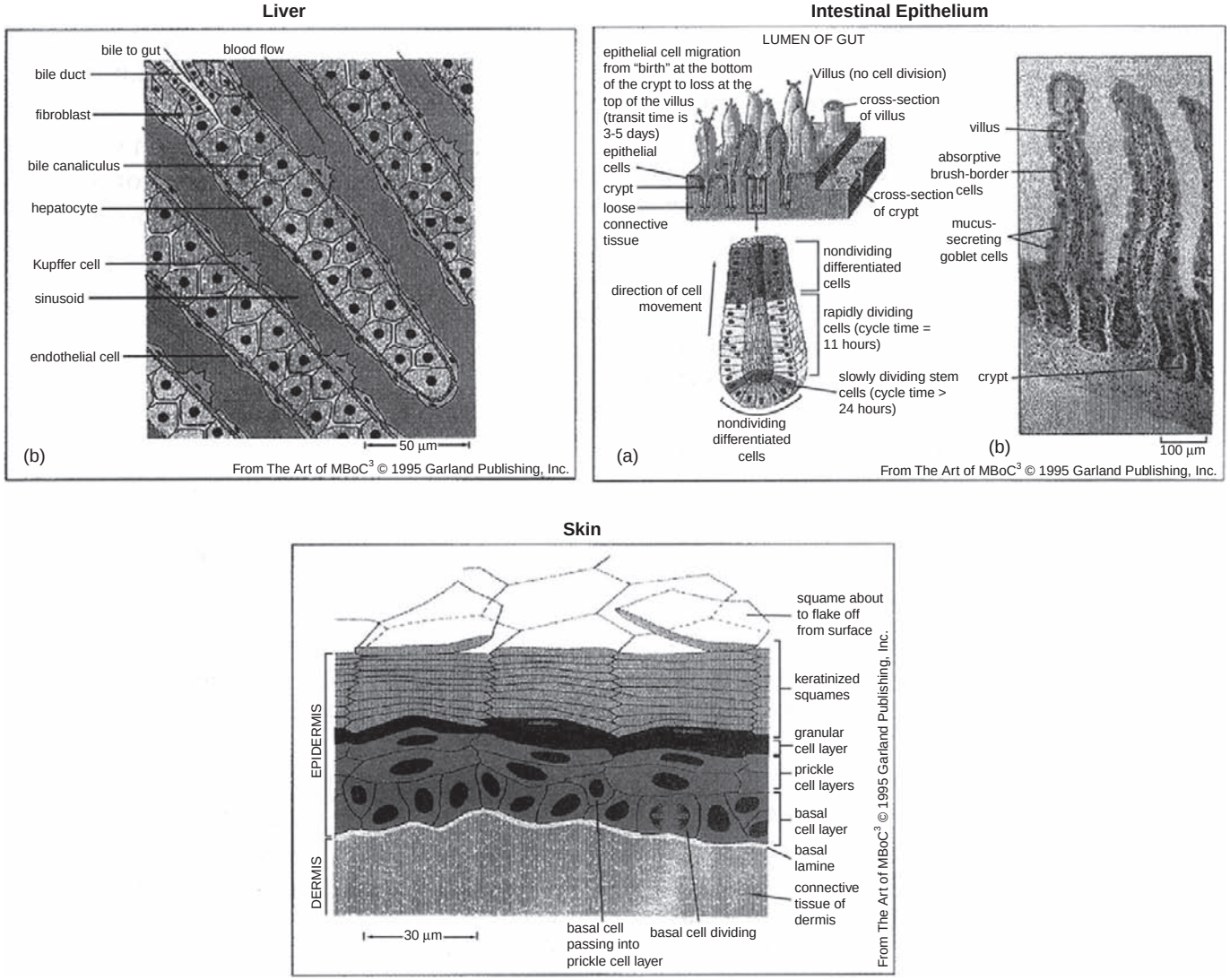


FIGURE 6.25 In vivo tissue microenvironments. *From Alberts et al., 1995.*

EXAMPLE PROBLEM 6.5

The hepatic stellate cell is a mesenchymal cell that is located in the Space of Disse, the area between the sheets of hepatocytes and the hepatic sinusoidal endothelium (Figure 6.26) in the liver. In the normal state, hepatic stellate cells are found in close proximity to a basement membrane-like matrix. This matrix consists of collagen type IV, laminin, and heparan sulphate proteoglycans. If the liver is injured, hepatic stellate cells are activated, and they begin to make matrix proteins, mostly collagen type I and III. As a result of hepatic stellate cell activation, the phenotype of both the hepatocytes (lose their brush border) and endothelium (lose their fenestration) changes. If this condition persists, liver fibrosis results. Thus, a lack of coordination in cellular function in the tissue microenvironment can lead to loss of tissue function. This example can be used to examine how liver dysfunction can be prevented or corrected, including using a tissue engineering approach.

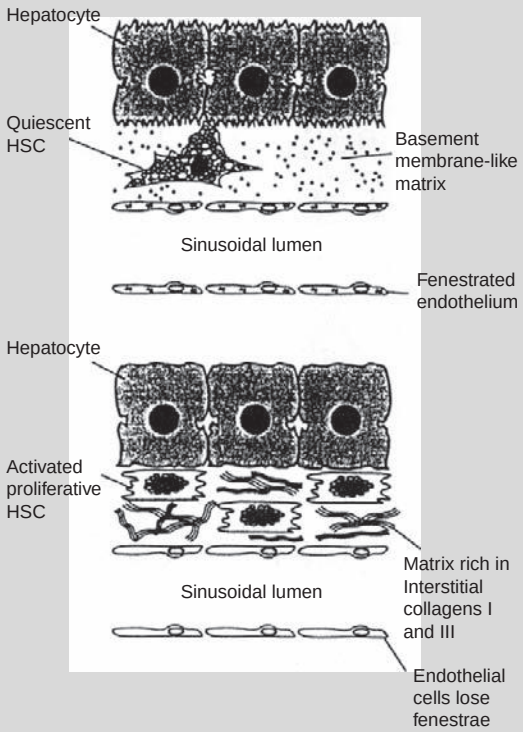


FIGURE 6.26 Schematic of hepatic stellate cell activation and the process of liver fibrosis. *From Iredale, 1997.*

Tissues are perfused by the microcirculation and can be viewed as essentially homogeneous down to a length scale of about 100 μm. The microcirculation then interfaces with the long-distance convective transport system that connects all the tissues in the body and, thus, all cells, to a nutritional supply, exchange of respiratory gases, removal of toxic products, and so on.

From a tissue engineering standpoint, a key challenge is to design the environment at both the micro and macro scales. This problem is characterized by two important considerations:

- The microenvironment (its chemical, geometric, cell architectural, and diffusional characteristics)
- Interactions with other tissues (source of nutrients, exchange of respiratory gases, removal of waste products, and delivery of soluble protein such as growth factors)

Cellularity

The packing density of cells is on the order of a billion cells per cubic centimeter (cc). Tissues are typically operating at one-third to one-half of packing density, such that typical cell densities in tissues are on the order of 100 to 500 million cells per cc. Since the characteristic length scale is about 100 μm, the order of magnitude of the number of cells found in a tissue microenvironment can be estimated: a 100 μm cube at 500 million cells per cc contains about 500 cells. Simple multicellular organisms, such as *C. elegans*, a much-studied small worm, that have about 1,000 cells, provide an interesting comparison.

The cellularity of the tissue microenvironment varies among tissues. An example of a tissue with low cellularity is cartilage. The function of chondrocytes in cartilage is to maintain the extracellular matrix. Cartilage is essentially avascular, alymphatic, and aneural. Thus, many of the cell types found in other tissues are not present, which limits the range of functions that the tissue can perform. The cellularity of cartilage is about a million cells per cc, or about one cell per cubic 100 μm. Most other tissues, in particular those that are highly metabolically active and/or remodel at a high rate, are much more cellular, with hundreds of cells per 100 μm cube.

The tissue microenvironment is characterized by having a number of cell types (Table 6.10). In addition to the parenchymal cells (the tissue type cells—e.g., hepatocytes in liver), a variety of accessory cells are found in all tissues, as described briefly following.

TABLE 6.10 Cells That Contribute to the Tissue Microenvironment

Stromal cells: derivates of a common precursor cell
Mesenchyme
Fibroblasts
Myofibroblasts
Osteogenic/chondrogenic cells
Adipocytes
Stromal-associated cells: histogenically distinct from stromal cells, permanent residents of a tissue
Endothelial cells
Macrophages

TABLE 6.10 Cells That Contribute to the Tissue Microenvironment—Cont'd

Transient cells: cells that migrate into a tissue for host defense either prior to or following an inflammatory stimulus
B lymphocytes/plasma cells
Cytotoxic T cells and natural killer (NK) cells
Granulocytes
Parenchymal cells: cells that occupy most of the tissue volume, express functions that are definitive for the tissue, and interact with all other cell types to facilitate the expression of differentiated function

- Mesenchymal cells (such as fibroblasts and smooth muscle cells) are present in all tissues. These cells are of connective tissue type.
- Monocytes are present in all tissues and can take on a variety of different morphologies (see Figure 6.27). Monocytes can differentiate into macrophages that, once activated, produce a variety of cyto- and chemokines that influence the behavior of neighboring cells.

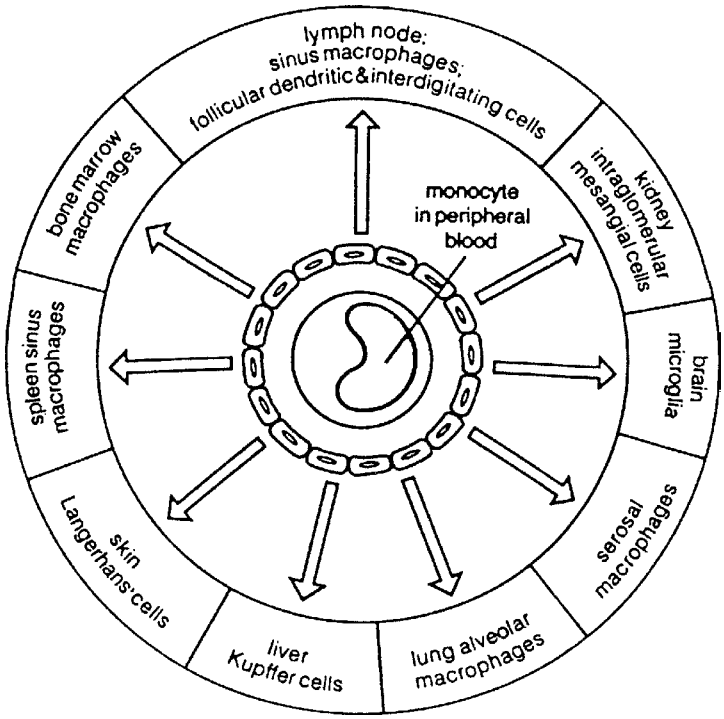


FIGURE 6.27 Distribution of macrophages and their presence in different tissues. *From [6].*

TABLE 6.11 Cells Contributing to the Hepatic Microenvironment

Cell Type	Size (μm)	Relative Percentage of Total Cells
Stroma		
Kupffer cells	12–16	8
Vascular endothelia	11–12	9
Biliary endothelia	10–12	5
Fat-storing cells	14–18	3
Fibroblasts	11–14	7
Pit cells	11–15	1–2 (variable)
Parenchymal cells		
Mononuclear (type I)	17–22	35
Binuclear (type II)	20–27	27
Acidophilic (type III)	25–32	5

- Endothelial cells are associated with the vasculature found in almost all tissues. These cells play a major role in the trafficking of cells in and out of tissue and may play a major role in determining tissue metabolism.
- Lymphocytes and neutrophils have a transient presence in tissues, typically as a part of a host defense response or other cleanup functions.

These accessory cells are typically about 30 percent of the cellularity of tissue, while the parenchymal cells make up the balance. An example is provided in Table 6.11 that lists the cellularity of the liver.

EXAMPLE PROBLEM 6.6

What happens if these accessory cells are removed?

Solution

The role of accessory cells (in some cases called stromal cells) in bone marrow cultures has been systematically studied. Since the immature cells can be isolated based on known antigens, the relative abundance of the key parenchymal cells and the accessory cells can be varied. Such an experiment amounts to a titration of the accessory cell activity. The results from such an experiment are shown in Figure 6.28. All cell production indices (total cells, progenitor cells, and preprogenitor cells) decline sharply as the accessory cells are removed. Supplying preformed irradiated stroma restores the production of total and progenitor cells but not the preprogenitors. This result is consistent with the expectation that specific interactions between accessory cells and parenchymal cells are important for immature cells.

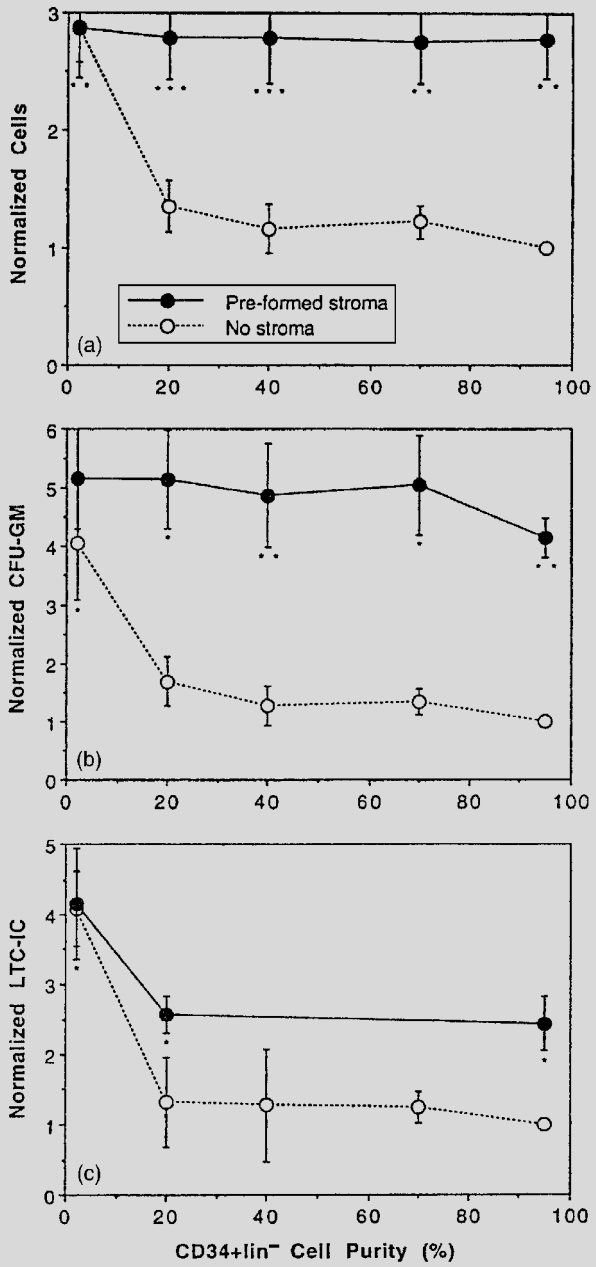


FIGURE 6.28 Effects of CD34+ lin$^-$ cell (a population of primitive hematopoietic cells) purity on culture output. With increasing purity, the performance on a per-cell basis drops due to loss of accessory cell function. CFU-GM ¼ colony-forming units granulocyte/macrophage; LTC-IC ¼ long-term cultureinitiating cells. *From Koller and Palsson, 1993.*

Dynamics

Temporal aspects of the cellular microenvironment also are critical to obtaining desired cell phenotype and function. The microenvironment is highly dynamic and displays a multitude of time constants. Some of the key dynamic processes are described following.

OXYGENATION

Generally, mammalian cells do not consume oxygen rapidly compared to microorganisms, but their uptake rate is large compared to the amount of oxygen available at any given time in blood or in culture media (Table 6.12). At 37°C, air-saturated aqueous media contain only about 0.21 mM oxygen per liter. Mammalian cells consume oxygen at a rate in the range of 0.05–0.5 μmol/10^6cells/hr. With tissue cellularities of 500 million cells per cc, these oxygen uptake rates call for volumetric oxygen delivery rates of 25 to 250 μmol oxygen/cc/hour. This rate needs to be balanced with the volumetric perfusion rates of tissues and the oxygen concentration in blood.

Metabolically active tissues and cell cultures, even at relatively low cell densities, quickly deplete the available oxygen. For instance, at cell densities of 10^6cells/ml, oxygen is depleted in most tissues in about 0.4 to 4 hours. Oxygen thus must be continually supplied. To date, a number of primary cell types (e.g., hepatocytes, keratinocytes, chondrocytes, and hematopoietic cells) have been grown ex vivo for the purpose of cell therapy. The effects of

TABLE 6.12 Measured Oxygen-Demand Rates of Human Cells in Culture

Human	μmol $O_2/10^6$ cells/h
HeLa	0.1–0.0047
HLM (liver)	0.37
LIR (liver)	0.30
AM-57 (amnion)	0.045–0.13
Skin fibroblast	0.064
Detroit 6 (bone marrow)	0.43
Conjunctiva	0.28
Leukemia MCN	0.22
Lymphoblastoid (namalioa)	0.053
Lung	0.24
Intestine	0.40
Diploid embryo WI-38	0.15
MAF-E	0.38
FS-4	0.05

oxygen on hepatocytes have been systematically investigated. The reported specific oxygen uptake rate (OUR) for hepatocytes is around 1.0 μmole/10^6cells/hr, which is relatively high for mammalian cells. Conversely, a much lower oxygen consumption rate of about 0.02 μmole/10^6cells/hr has been reported for rat bone marrow cells.

In addition to the supply requirements, the concentration of oxygen close to the cells must be within a specific range. Oxygenation affects a variety of physiological processes, ranging from cell attachment and spreading to growth and differentiation. An insufficient concentration retards growth, while an excess concentration may be inhibitory or even toxic. For instance, several studies have shown that the formation of hematopoietic cell colonies in colony assays is significantly enhanced by using oxygen concentrations that are 5 percent of saturation relative to air, and an optimal oxygen concentration for bioreactor bone marrow cell culture has been shown to exist. The oxygen uptake rate of cells is thus an important parameter in the design of cell culture and tissue engineering studies.

METABOLISM AND CELL SIGNALING

Typically, there is not a transport limitation for major nutrients, although cells can respond to their local concentrations. The reason is that their concentrations can be much higher than that of oxygen, especially for the nutrients consumed at high rates. Typical uptake rates of glucose are on the order of 0.2 μmoles/million cells/hr, while the consumption rates of amino acids are in the range of 0.1–0.5 μmoles per billion cells per hour. The transport and uptake rates of growth factors face more serious transport limitations. The expected diffusional response times are given in Figure 6.29.

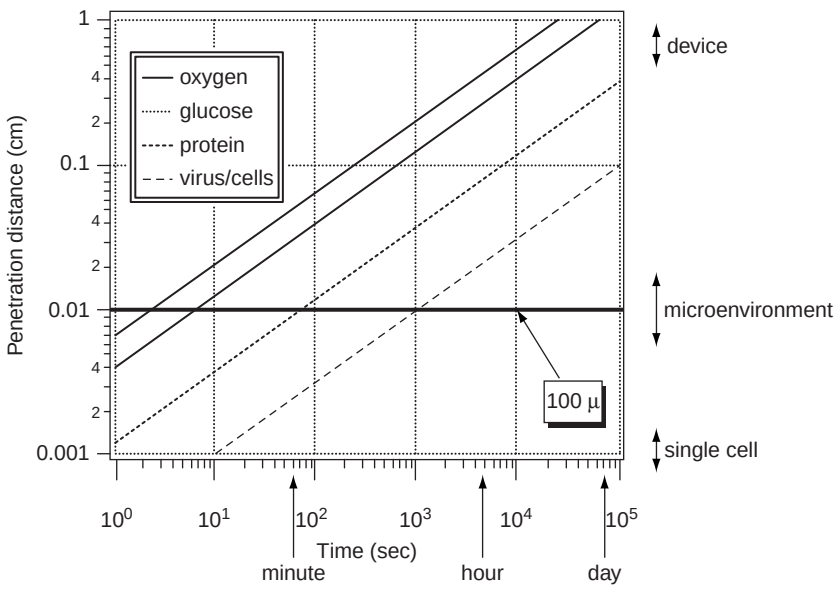


FIGURE 6.29 The diffusional penetration lengths as a function of time for several classes of biomolecules.

PERFUSION

The circulatory system provides blood flow to organs that is then distributed into the microenvironments. Overall, the perfusion rates in a human are about 5 liters/min/70 Kg, or about 0.07 ml/cc/min. With 500 million cells per cc, this is equivalent to 0.14 μl/million cells/min. These numbers represent a whole body average. There are differences in the perfusion rates of different organs that typically correlate with their metabolic activity.

CELL MOTION

As described earlier, cells are motile and move at different rates. Neutrophils can move several cell diameters in a minute, while adherent cell types, such as keratinocytes, fibroblasts, and endothelial cells, move on the order of a cell diameter per hour. These motilities represent rapid processes compared to cell replication and differentiation. Neutrophils have to be able to respond rapidly to invading microorganisms, while the adherent cell types mentioned move in response to dynamic tissue needs.

Size and Geometry

GEOMETRY

The geometric shapes of microenvironments vary (see Figure 6.25), and so do their dimensionalities. Many microenvironments are effectively curved 2-D surfaces. The cellular arrangement in bone marrow has been shown to have a fractal geometry with effective dimensionality of 2.7, while the brain is a 3-D structure.

WHAT DETERMINES THE SIZE OF THE MICROENVIRONMENT?

The answer to this question is not clear. At the lower limit is the size of a single cell, about 10 μm. A cell aggregate must be some multiple of this distance. The factors determining the upper limit are not clear. However, estimates of effective growth factor signal propagation distances and experimental evidence of oxygen penetration distances suggest that the dynamics of cell communication and cell metabolic rates are important in determining the size scale of the microenvironment. These distances are determined by the process of diffusion. In both cases, the estimated length scale is about 100 to 200 μm. The stability issues associated with coordinating cellular functions grow with an increased number of cells and may represent a limitation in practice.

6.3.5 Biomaterials

Biomaterials for tissue engineering present several challenges. There are basically three length scales of interest. The shortest is at the biochemical level, where concerns include the specific chemical features of the ECM and interactions with cellular receptors. Intact ECM components can be used to coat support material to ensure appropriate interactions among the cells and their immediate environment. More sophisticated treatments involve the synthesis of specific binding sequences from ECM protein and presenting them in various ways to the cells. Particular cellular arrangements can be obtained by micropatterning such materials. A combination of material manufacturing, biochemistry, and genetics is required to address these issues.

The next size scale of interest is the 100 micron size scale, the size of a typical organ microenvironment. Many organs have highly specific local geometries that may have to be engineered in an ex vivo system. Hence, a particular microgeometry with particular mechanical properties may have to be produced. Clearly, challenging material manufacturing issues arise. Further, the support matrix may have to be biodegradable after transplantation and the degradation products nontoxic. Lactic and glycolic acid-based polymers are promising materials in this regard. If little restructuring of implants occurs following grafting, then the geometry of the support matrix over larger size scales may be important.

The largest size scale is that of the bioreactor itself. Bioreactors in tissue engineering are likely to be small with dimensions on the order of about 10 cm. The materials issues that arise here are primarily those of biocompatibility. Although manufacturing technology exists for tissue culture plastic, it is likely that additional issues will arise. The tissue culture plastic that is commercially available is designed to promote adhesion, binding, and spreading of continuous cell lines. Although such features may be desirable for continuous cell lines, they may not be so for various primary cells.

6.4 SCALING UP

6.4.1 Fundamental Concept

As just discussed, tissue dynamics are comprised of intricate interplay between the cellular fate processes of cell replication, differentiation, and apoptosis. They are properly balanced under in vivo conditions. The dynamics of the in vivo conditions are a balance of these biological dynamics and the constraining physicochemical processes. The basic concept of design in tissue reconstruction is to engineer a proper balance between the biological and physicochemical rates so normal tissue function can occur.

6.4.2 Key Design Challenges

Within this framework, many of the engineering issues associated with successful reconstitution of tissues can be examined. This section provides an engineering perspective of tissue engineering and helps to define the productive and critical role that engineering needs to play in the ex vivo reconstruction of human tissues.

Important design challenges in tissue engineering include the following:

- Oxygenation—that is, providing adequate flux of oxygen at physiological concentrations
- Provision and removal of cyto- and chemokines
- Physiological perfusion rates and uniformity in distribution
- Biomaterials, including functional, structural, toxicity, and manufacturing characteristics

Other issues associated with the clinical implementation of cellular therapies include the design of disposable devices, optimization of medium composition, initial cell distribution, meeting FDA requirements, and operation in a clinical setting. These cannot be discussed in detail here, but some of these issues are addressed generally in the following section.

6.4.3 Time Scales of Mass Transfer

The importance of mass transfer in tissue and cellular function is often overlooked. The limitations imposed by molecular diffusion become clear if the average displacement distance with time is plotted for diffusion coefficients that are typical for biological entities of interest in tissue function (Figure 6.29). The diffusional penetration lengths over physiological time scales are surprisingly short and constrain the in vivo design and architecture of organs. The same constraints are faced in the construction of an ex vivo device, and high mass transfer rates into cell beds at physiological cell densities may be difficult to achieve.

The biochemical characteristics of the microenvironment are critical to obtaining proper tissue function. Much information exists about the biochemical requirements for the growth of continuous cell lines. For continuous cell lines, these issues revolve around the provision of nutrients and the removal of waste products. In cultures of primary cells, the nutrients may have other roles and directly influence the physiological performance of the culture. For instance, recently it has been shown that proline and oxygen levels play an important role in hepatocyte cultures.

In most cases, oxygen delivery is likely to be an important consideration. Too much oxygen will be inhibitory or toxic, while too little may alter metabolism. Some tissues, like the liver, kidney, and brain, have high oxygen requirements, while others require less. Controlling oxygen at the desired concentration levels, at the desired uniformity, and at the fluxes needed at high cell densities is likely to prove to be a challenge. Further, the oxygen and nutritional requirements may vary among cell types in a reconstituted tissue that is comprised of multiple cell types. These requirements further complicate nutrient delivery. Thus, defining, designing, and controlling the biochemical characteristics of the microenvironment may prove difficult, especially given the constraints imposed by diffusion and any requirements for a particular microgeometry.

EXAMPLE PROBLEM 6.7

How are specific oxygen uptake rates measured?

Solution

Most data on specific oxygen uptake rates are obtained with cells in suspension using standard respirometers. However, to obtain accurate and representative data for primary cells, they need to be adherent. This challenge has led to the design of a novel in situ respirometer, shown in Figure 6.30. This respirometer has been used to measure the oxygen uptake rates of hepatocytes in culture (OUR of 1.0 μmoles/million cells/hr). A similar device has been used to measure the OUR in bone marrow cultures, giving results of 0.03 to 0.04 μmoles/million cells/hr which are similar to the in vivo uptake rates.

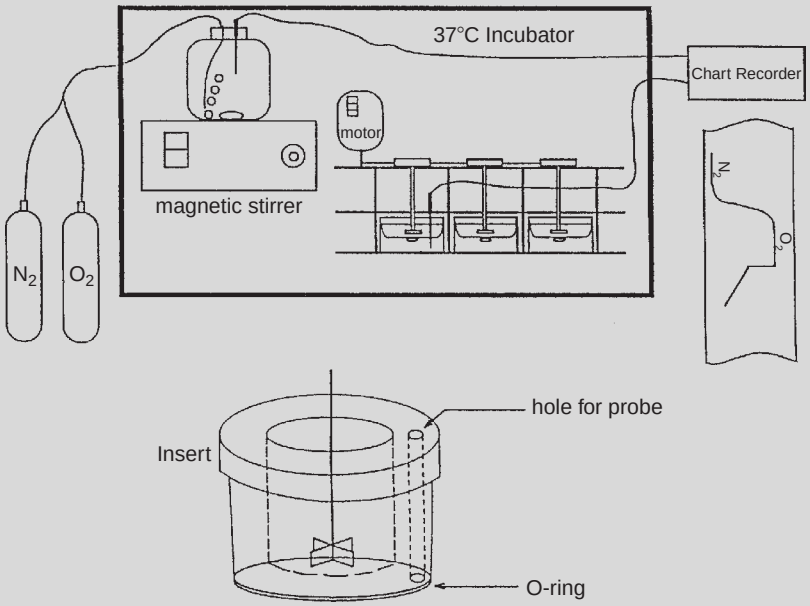


FIGURE 6.30 Schematic diagram of the apparatus for measuring oxygen uptake rate. *From [14].*

EXAMPLE PROBLEM 6.8

Oxygen can be delivered in situ over an oxygenation membrane, as illustrated in Figure 6.31. At slow flow rates compared to the lateral diffusion rate of oxygen, the oxygen in the incoming

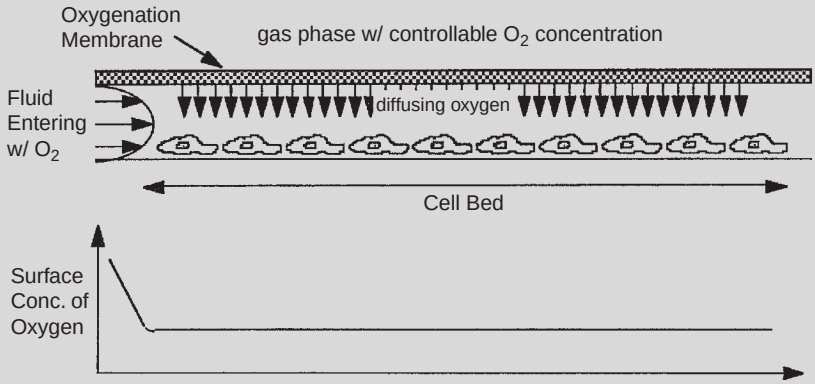


FIGURE 6.31 Oxygen delivery across a membrane that is placed on top of a fluid that is flowing across a cell bed. If the fluid transit time is much slower than the diffusional time for oxygen, then oxygen is delivered primarily via diffusion. This leads to a small entrance effect in which the oxygen in the incoming stream is consumed, while the oxygen concentration over the rest of the cell bed is relatively constant.

Continued

stream is quickly depleted, and the bulk of the cell bed is oxygenated via diffusion from the gas phase. This leads to oxygen delivery that can be controlled independently of all other operating variables and that is uniform. The gas phase composition can be used to control the oxygen delivery. This is an example of how technology can be used to overcome some of the limitations imposed by mass transfer considerations.

6.4.4 Fluid Flow and Uniformity

A device that carries tens of millions of microenvironments (i.e., a reconstituted tissue or organ) must provide uniform delivery and removal of gases, nutrients, and growth factors. Achieving such uniformity is difficult. This difficulty arises partly from the fact that fluid has a no-slip condition as it flows past a solid surface. Thus, there will always be slower-flowing regions close to any side walls in a bioreactor. These slow-flowing regions under conditions of axial Graetz numbers of unity lead to mass transfer boundary layers that extend beyond the hydrodynamic boundary layer.

This problem is illustrated further in Figure 6.32. Fluid is flowing down a thin slit, representing an on-end view of the chamber shown in Figure 6.33. Thin slits with high aspect

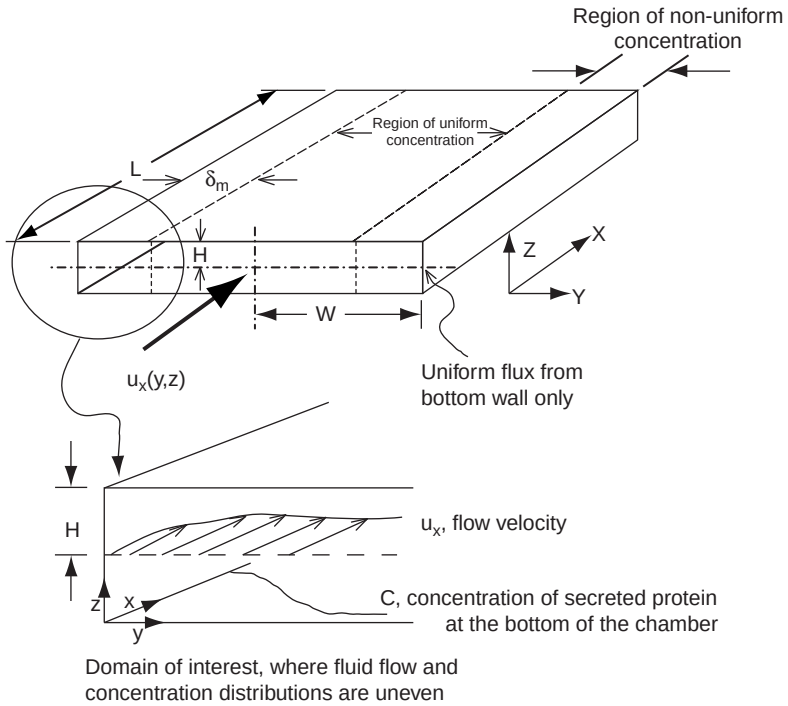


FIGURE 6.32 Coordinate system for a rectangular chamber with production of biological factors secreted by cells lodged on the bottom wall. The fluid is slowed down close to the side wall, creating a different concentration than that found in the middle of the slit. This leads to a very different microenvironment for cell growth and development of tissue function at the wall than elsewhere in the chamber. *From [15].*

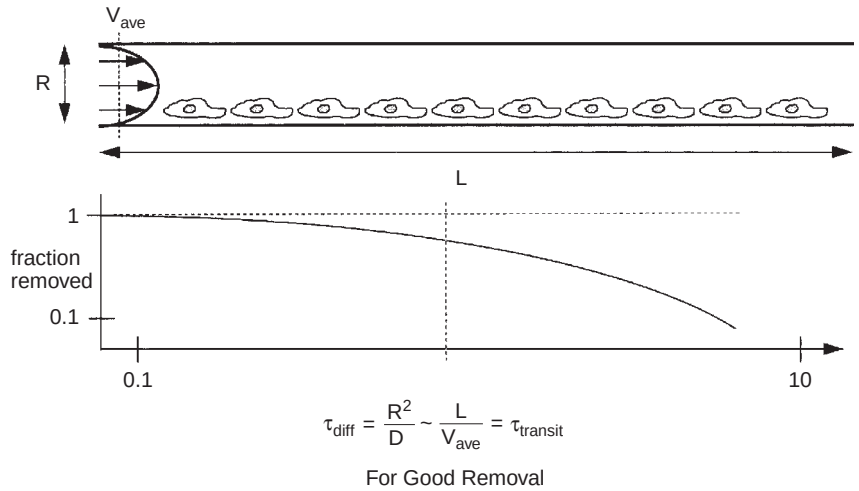


FIGURE 6.33 A schematic of the concentration of growth factors in a liquid that is flowing across a bed of cells that consume it. (Bottom) The fractional removal of the growth factor as a function of the relative time constants of lateral diffusion and transit across the cell bed. The ratio of the two is the Graetz number (Gz). If the diffusional response time is slower than that of transit (Gz >1), then there is insufficient time for the diffusing growth factor molecule to make good contact with the cell bed. Most of the growth factor leaves the system in the exit stream. Conversely, if the diffusion time is much shorter than that of transit, there will be ample time for the growth factor to make it to the cell bed. If it is rapidly consumed there, then negligible amounts will leave the system.

ratios will satisfy the Hele-Shaw flow approximations, in which the flow is essentially the same over the entire width of the slit except close to the edges. The width of the slow-flowing regions is on the order of the depth of the slit (R). Thus, if the aspect ratio is 10, fluid will flow the same way over about 90 percent of the width of the slit. In the remaining 10 percent, the slow flow close to the wall can create microenvironments with a different property than in the rest of the cell bed. Such nonuniformity can lead to differences in cell growth rates and to migration of cells toward the wall. Such nonuniformity in growth close to walls has been reported. This problem can be overcome by using radial flow configurations. The example shown is for one type of a bioreactor for cell culturing. Other configurations will have similar difficulties in achieving acceptable uniformity in conditions, and careful mass and momentum transfer analyses need to be performed to guide detailed design.

6.5 IMPLEMENTATION OF TISSUE ENGINEERED PRODUCTS

6.5.1 Delivering Cellular Therapies in a Clinical Setting

In the last section, the design challenges that the tissue engineer is faced with when it comes to scaling up microenvironments to produce cell numbers that are of clinical significance were surveyed. These problems are only a part of the challenges that must be met in implementing cellular therapies. In this section, some of the other problems that need to be solved will be discussed.

Donor-to-Donor Variability

The genetic variability in the human population is substantial. Therefore, outcome from cell isolations and tissue engineering approaches can vary significantly between donors. Even if the cell growth process, the production of materials, and the formulation of medium are essentially identical, a large variation in outcomes among donors can still result. This variability is due to intrinsic biological factors. Some regularization in performance can be achieved by using a full complement of accessory cells. Although interdonor variability is considerable, the behavior of the same tissue source is internally consistent. An example is shown in Figure 6.34, where the relative uptake rates of growth factors are shown for a number of donor samples. The uptake rates of growth factors can be highly correlated within many donor samples. The quantitative nature of the correlation changes from donor to donor, making it difficult to develop a correlation that would represent a large donor population.

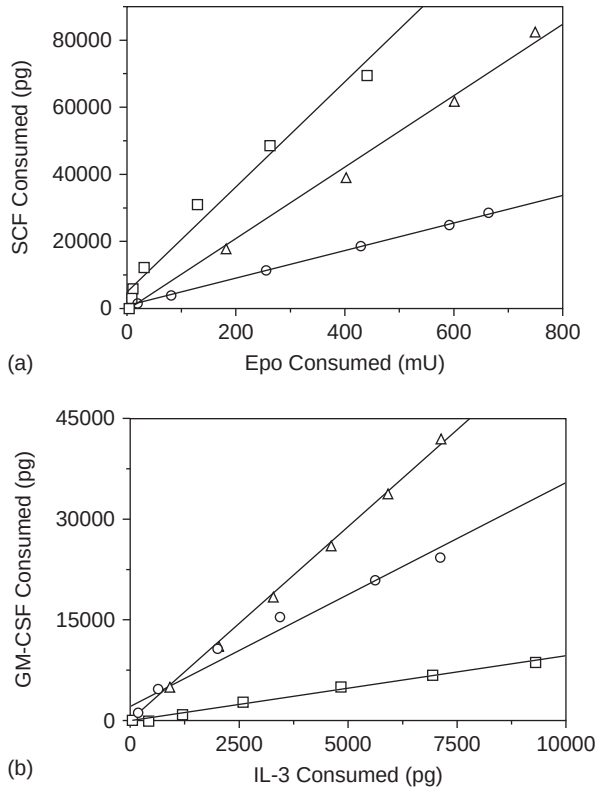


FIGURE 6.34 The uptake of growth factors—that is, stem cell factor (SCF) and erythropoietin (Epo)—is highly correlated in cultures of human bone marrow. The correlation is strong for a single tissue sample. The slopes of the curves vary significantly among donors. *From [8].*

Strongly Interacting Variables

Primary cell cultures are sensitive to many of the variables that define the microenvironment. Many of these variables interact strongly. Thus, any change in a single cell culture variable will change the optimal value for all the others. Statistical experimental design procedures that optimize the search over many experimental variables can be employed to lessen this effect. An example of such a two-dimensional search is shown in Figure 6.35, where the optimal progenitor production performance of human cell cultures is shown as a function of the inoculum density and the medium flow rate. The top panel shows the optimal number of progenitors produced per unit cell growth area. Optimizing this objective would lead to the smallest cell culture device possible for a specified total number of cells that is needed. The bottom panel shows the optimal expansion of progenitor cells—that is, the output number relative to the input. This objective would be used in situations where the starting material is limited and the maximum number of additional progenitors is desired. Note that the two objectives are found under different conditions. Thus, it is critical to clearly delineate the objective of the optimization condition from the outset.

Immune Rejection

Allogeneic transplants face immune rejection by the host. A variety of situations are encountered in such transplantations. Dermal fibroblasts, used for skin ulcers, seem to be effectively nonimmunogenic, though the reason is not clear. This fact makes it possible to make a large number of grafts from a single source and transplant into many patients. In contrast, pancreatic beta cells and islets face almost certain rejection (see Figure 6.6). In allogeneic bone marrow transplantation, the graft may reject the immunocompromised host. This so-called "Graft-vs-Host disease" is the main cause for the mortality resulting from allogeneic bone marrow transplants.

The cellular and molecular basis for the immune response is becoming better understood. The rejection problem is a dynamic process that relies on the interaction between subsets of $CD4^+$ cells ($CD4^+$ is a surface antigen on certain T-cells) Th1 and Th2 that differ in their cytokine secretion characteristics and effector functions. The components of the underlying regulatory network are becoming known. Quantitative and systemic analysis of this system is likely to lead to rational strategies for manipulating immune responses for prophylaxis and therapy.

Tissue Procurement

The source of the starting material for a growth process is of critical importance. For example, the source of dermal fibroblasts used for skin replacements is often human foreskin obtained from circumcisions. This source is prolific and can be used to generate a large number of grafts. Since the source is always the same, the biological variability in the growth process is greatly diminished. The costs associated with tissue procurement as reflected in the final product are minimal. Conversely, an adult autologous source may be expensive to obtain and will display highly variable performance.

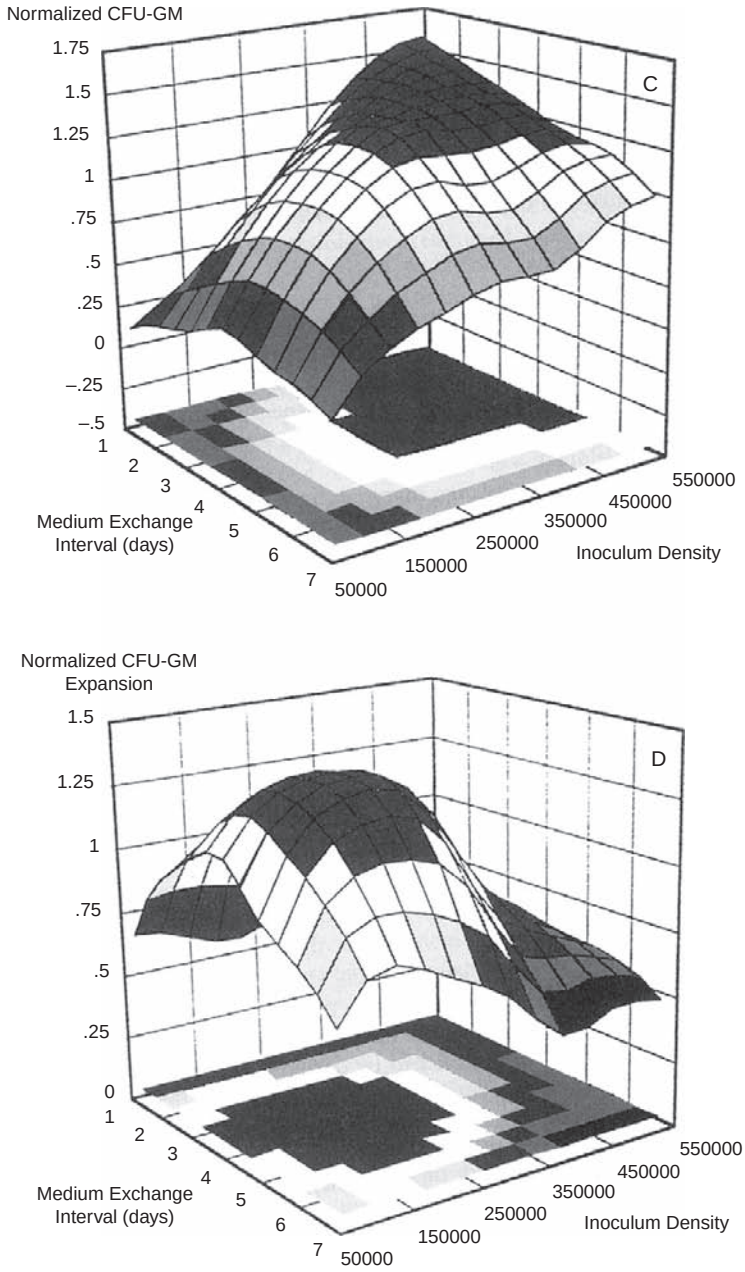


FIGURE 6.35 The performance of a bone marrow culture system over a wide range of inoculum densities and medium exchange intervals. Note that the variables interact. The performance (on the *z*-axis) in the top panel is the total number of progenitor cells (CFU-GM) per unit area, whereas in the bottom panel the expansion ratio is depicted (number out divided by number in). Note that the two measures are optimized under vastly different conditions. *From [8].*

Ultimately, the most desirable starting material for cellular therapies is a tissue-specific stem cell whose rate of self-renewal and commitment to differentiation can be controlled. Further, if such a source could be made nonimmunogenic, most of the problems associated with tissue procurement would be solved. The variability in the tissue manufacturing process would be reduced and would make any quality assurance (QA) and quality control (QC) procedures easier.

Cryopreservation

The scientific basis for cryopreservation involves several disciplines, including basic biophysics, chemistry, and engineering. Current cryopreservation procedures are clinically accepted for a number of tissues, including bone marrow, blood cells, cornea, germ cells, and vascular tissue. Recent experience has shown that, in general, the same procedures cannot be applied to human cells that have been grown ex vivo. New procedures need to be developed and implemented. Any cryopreservation used in existing cellular therapies that rely on ex vivo manipulation call for freezing the primary tissue prior to the desired manipulation.

6.6 FUTURE DIRECTIONS: FUNCTIONAL TISSUE ENGINEERING AND THE "-OMICS" SCIENCES

6.6.1 Cellular Aspects

Some cell populations that are to be transplanted may contain subpopulations of unwanted cells. The primary example of this is the contamination of autologously harvested hematopoietic cell populations with the patient's tumor cells. Ideally, any such contamination should be removed prior to transplantation. Similarly, many biopsies are contaminated with accessory cells that may grow faster than the desired parenchymal cells. Fibroblasts are a difficult contaminant to eliminate in many biopsies, and they often show superior growth characteristics in culture. For this reason, it is difficult to develop primary cell lines from many tumor types.

6.6.2 Functional Tissue Engineering

Tissue engineering is distinguished from cell biology by the focus on the emergent function that arises from the organization of large numbers of cells into higher-order structures, variously called *tissues* or *organs*, depending on the level of anatomical complexity and structural integration. The reengineering of complex human anatomical structures such as limbs or organ systems is by definition a systems-engineering problem. Though it can be argued that all tissue functions arise from fundamental cellular mechanisms, the system-level organization of tissues and organs confers function that is not possible to achieve with individual cells or masses of unorganized cells in a scaffold. A pile of bricks does not provide the functionality of a house, nor does a crate full of car parts function like an automobile. Analogously, engineered tissues must be viewed at the systems level, and the success or failure of the engineering effort ultimately rests upon a quantitative assessment of the organ-level function of the engineered tissue or organ. The use of molecular biological

techniques to verify the presence of one or more critical subcellular constituents is simply not an adequate demonstration of end organ function. Thus, it is important to develop the necessary facilities to actually quantify the organ-level function of tissue engineered constructs. Of course, different tissues and organ systems have different functions. When designing a tissue or organ, it is therefore essential to develop a design specification for the engineered tissue, with well-defined, quantitative functional assessments, also called "figures of merit (FoM)," as well as a defined method by which to assess these values.

Presumably, the tissues or organs will be cultured for some time prior to their use to permit growth and development. Many tissues have measurable function that changes during development, so it is most desirable to identify one or more quantities that may be measured nondestructively during the course of the development of the tissues in culture. A specific example is instructive: the contractility of mammalian skeletal muscle changes throughout the early stages of development into adulthood. Muscle phenotype is defined largely by the myosin heavy chain content of individual muscle fibers, but these can be quantitatively inferred by nondestructive measurements of the isometric and dynamic contractility of the muscle tissue. The same is true for tendon. The tangent modulus, tensile strength, and fracture toughness increase during development, whereas the size of the "toe region" of the stress strain curve tends to decrease, presumably due to the increasingly well-ordered collagen structure during pre- and early postnatal development. It is not possible to nondestructively test the tensile strength and fracture toughness of a cultured tendon specimen, but the tangent modulus and the characteristics of the toe region can be readily measured with minimal disruption to the tendon tissue in culture. With musculoskeletal tissues it often happens to be the case that the electromechanical signals that are required for nondestructive quantitative assessment of the tissues in culture are essentially the same as those that would be applied chronically to the tissues in culture to guide and promote development. For example, electrically elicited contractions of skeletal and cardiac muscle are currently in use in an attempt to promote development, and the application of mechanical strain has been used since the 1980s on many musculoskeletal (muscle, bone, tendon, cartilage, ligament) and cardiovascular tissues to promote development in culture.

An important future challenge is to develop bioreactor systems that permit the application of the stimulus signals, while simultaneously allowing the functional properties of the tissues to be nondestructively measured and recorded. If the functional properties of the developing tissue are measured in real-time in a bioreactor system, it then becomes possible to assess the current developmental status of each tissue specimen and to use this information to modify the stimulus parameters accordingly. This permits stimulus feedback control of the tissue during development and represents a significant increase in the level of sophistication and effectiveness of functional tissue engineering technology. This constitutes an important aspect of current research in the areas of both musculoskeletal and cardiovascular functional tissue engineering.

6.6.3 Bioartificial Liver Specifics

The development of bioartificial liver (BAL) devices arose from the fact that "backup" systems to replace deficient liver functions are nonexistent, in contrast to other tissue in which duplications exist (e.g., as dual lung lobes, two kidneys, fibular crutches). The liver

provides multiple functions to assist with body homeostasis and needs replacement systems when confronted with organ failures such as fulminant hepatic failure (FHF).

Reports of liver treatment date back to the 1950s when low-protein diets were recommended to improve mental impairment and hepatic encephalopathy, and in the 1960s novel concepts of liver assist devices began to emerge. Some of these precedent artificial assist systems (Artif-S) currently remain on the research bench or have entered into preliminary FDA trials due to their intrinsic capabilities of treating patients suffering from FHF or other liver-specific malfunctions. A few of these systems include charcoal filters for ammonia detoxification, mechanical dialysis permitting toxin transfers, and plasmapheresis for removal of diseased circulating substances. Investigations have shown that many Artif-S are successful in their focused purpose, but they are not complete solutions to replace organ function. Although Artif-S are continually being improved, the multitude of tasks performed by the healthy in vivo liver continues to be insufficiently replicated through mechanical mimics of liver cell function.

One successful hurdle in the treatment of patients with FHF is the process of tissue transplantation. This technique exchanges a nonfunctioning liver with a healthy organ capable of performing all metabolic reactions. In this way, successful replacement surgeries alleviate the burden of using mechanical devices in concert with cellular activity. The drawback is that patients must remain on potent immunosuppressants to lessen tissue rejection responses, which decreases quality of life. Even though transplantation options are successful, the limited supply of donor liver organs along with tissue matching requirements illustrate the demand is approximately 300 percent greater than the supply Ultimately, an alternative to bridge or dissolve the *waiting* gap for patients expiring while on liver donor lists must be resolved. This is a clear opportunity for the field of tissue engineering to improve the standard of care and quality of life of a patient population.

6.7 CONCLUSIONS

This chapter outlined some of the key challenges and potential solutions in the field of tissue engineering. To successfully understand tissue function, it must be possible to quantitatively describe the underlying cellular fate processes. Such understanding will allow the tissue engineering to design and control these processes. One aspect of this approach is designing the physicochemical rate processes so they match the requirements of the cellular processes that underlie tissue function. Tissue engineering is an effort that is still in an embryonic stage, but the use of order-of-magnitude and dimensional analysis is proving to be valuable in designing and reconstituting tissue function.

6.8 EXERCISES

1. Given the following data, assess whether human hematopoietic stem cells can truly self-renew in vivo:

- About 400 billion mature hematopoietic cells are produced daily.

Continued

- Best estimate of the Hayflick limit is 44 to 50.
- About 1:5,000 progenitors do *not* apoptose.
- About 50 to 1,000 mature cells are made per progenitor (6 to 10 doublings).
- The entire differentiation pathway may be 17 to 20 doublings (soft fact).
- In vitro, about 10 to 30 million cells maximum can be made from a highly purified single hematopoietic stem/preprogenitor cell.

Present order of magnitude calculations in constructing your decision. Also perform parameter sensitivity analysis of each of the parameters that govern the hematopoietic process. How important are the parameters that govern telomerase activity in determining the total number of mature progeny produced over a person's lifetime?

2. At 1 pM concentration, how many molecules are found in a volume of liquid that is equal to the volume of one cell (use a radius of 5 microns)?
3. Use the continuum approach (Eq. (4.3)) to show that in a steady state the number of cells produced during a differentiation process that involves replication but no apoptosis ($\alpha = 0$) is

$$\frac{X_{out}}{X_{in}} = e^{\mu/\delta}$$

and is thus primarily a function of the ratio δ/μ. $a = 0$ is the completely undifferentiated state, and $a = 1$ is the completely differentiated state. What is X_{out} if μ and δ are the same orders of magnitude and if δ is 10 times slower? Which scenario is a more reasonable possibility in a physiological situation?

(Note that if the rates are comparable, only two mature progeny will be produced. On the other hand, if the differentiation rate is 10 times slower than the replication rate (probably close to many physiologic situations—i.e., 20 hr doubling time, and 200 hr = 8 days differentiation time), then about 1 million cells will be produced. Thus, the overall dynamic state tissue is strongly dependent on the relative rates of the cellular fate processes.)

4. Kinetics of differentiation/continuous model.
   - **i.** Derive the first-order PDE that describes the population balance.
   - **ii.** Make time dimensionless relative to the rate of differentiation.
   - **iii.** Describe the two resulting dimensionless groups (call the dimensionless group for apoptosis A, and the one for the cell cycle B).
   - **iv.** Solve the equation in steady state for $A = 0$.
   - **v.** Solve the equation(s) where A is nonzero for a portion of the differentiation process—that is, between a_1 and a_2.
   - **vi.** Solve the transient equation for $A = 0$.
5. Consider two cells on a flat surface. One cell secretes a chemokine to which the other responds. Show that the steady-state concentration profile of chemokine emanating from the first cell is

$$C(r) = R^2F/D \bullet 1/r$$

where R is the radius of the cell, F is the secretion rate (molecules/area time), and D is the diffusion coefficient of the chemokine. The distance from the cell surface is r. Use the cell flux equation to calculate the time it would take for the responding cell to migrate to the signaling cell if there is no random motion ($\mu = 0$) given the following values:

$\chi(C) = 20\ cm^2/sec\text{-}M$
$D = 10^{-6}\ cm^2/sec$
$R = 5\ \mu m$
Production rate = 5,000 molecules/cell/sec

Hint: Show that the constitutive equation for J reduces to $v = \chi\ dC/dr$, where v is the velocity of the cell.

6. The flux (F) of a molecule present at concentration C through a circular hole of diameter d on a surface that is adjacent to a fluid that it is diffusing in is given by

$$F = 4DdC$$

The total that can be transferred is the per pore capacity times the number of pores formed.

Calculate the flux allowed though each pore if the diffusion coefficient is $10^{-6}\ cm^2/sec$, the concentration is 1 mM, and the pore diameter is 1.5 nm. Discuss your results, and try to estimate how many pores are needed to reach meaningful cell-to-cell communications. With the per-pore flux just estimated, derive the time constant for transfer of a metabolite from a particular cell to a neighboring cell. Assume that these are two epithelial cells whose geometry can be approximated as a box and that the two adjacent boxes are connected with transfer occurring through *n* pores.

7. If the cellularity in cartilage is about 1 million cells per cc, estimate the average distance between the cells. Discuss the characteristics of this microenvironment.

8. Use a one-dimensional analysis of the diffusion of oxygen into a layer of adherent cells to show that the maximum oxygen delivery per unit area (N_{ox}^{max}) in Example Problem 6.8 is given by

$$N_{ox}^{max} = DC^*/R$$

where C^* is the saturation concentration of oxygen and R is the thickness of the liquid layer.

9. Consider a neuron growth cone that is being influenced by a chemoattractant produced by a target cell. The geometry of the model system is shown in the attached figure. The target cell secretes a chemoattractant at a rate, P_r, which diffuses into a three-dimensional volume, with a diffusivity D. The governing equation for mass transport for a spherical source is

$$\frac{\partial C}{\partial t} = D\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial C}{\partial r}\right)$$

a. What are the boundary conditions for the system?

b. Considering a steady state, derive the concentration profile as a function of *r*. It has been found that the growth cone senses a target cell when the concentration difference across the growth cone is higher than 2 percent. It is believed that growth cones develop filopodia that extend radially out of the cone as a means to enhance their chemosensing ability. Let β_1 be the angle that a filopodia makes with the center radius line, R, and β_2 the angle made by a filopodia extending in the diametrically opposite direction. Filopodia can extend radially from the growth cone surface, except where the cone connects to the axon as defined by α.

c. What are the appropriate limits of β_1 and β_2?

d. What are r_1 and r_2 as a function of β_1?

Continued

e. What is the percentage change in concentration, $\Delta C = \frac{|C_1 - C_2|}{C_0}$, across the effective growth cone radius—for example, $R_{gc} + l$—as a function of r_1 and r_2 and consequently β_1?

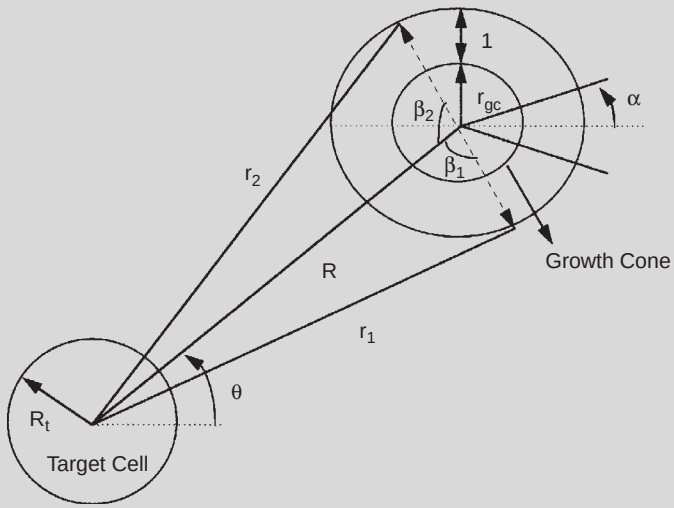
f. What is $\beta_1{}^{max}$, for which ΔC is maximum? Plot the gradient change for the entire range of β_1 for filopodial lengths of 1, 5, and 8 μm compared to no filopodia.

g. The limit for chemosensing ability is a concentration difference of 2 percent. For a target cell 500 μm away, what is the effect of filopodial length on enhancing chemosensing ability? Calculate for $\beta_1{}^{max}$.

$$C_0 = \frac{P_r R_t^2}{DR_o}, P_r = \frac{S_r}{4\pi R_t^2}$$

$$D = 10^{-6}\text{cm}^2/\text{s} \text{ and } S_r = 5{,}000 \text{ molecules/cell/s. } R_0 = R - R_{gc}$$

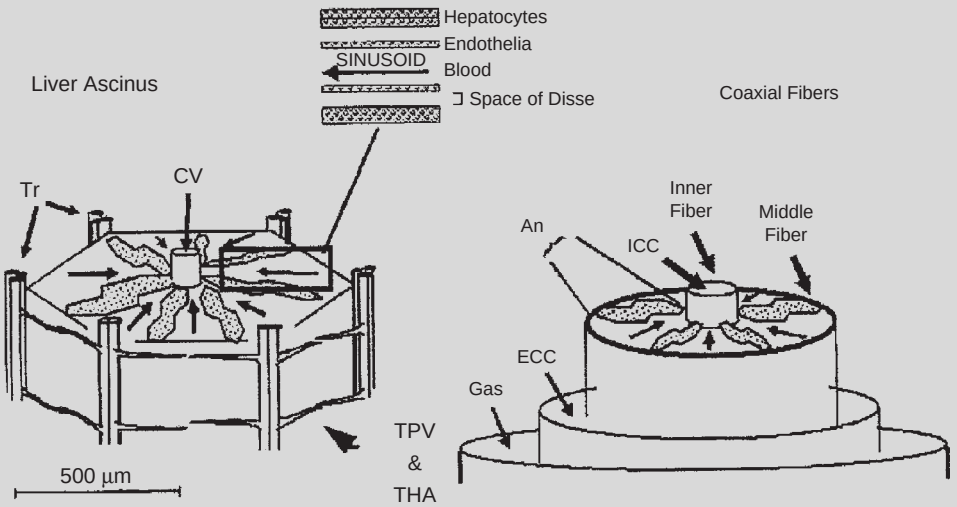
$$R_t = 20\ \mu\text{m}, R_{gc} = 2.5\ \mu\text{m}, \alpha = 30^\circ.$$



10. A coaxial bioreactor has been developed to mimic the liver acinus (see the following figure) .

a. Knowing the dimensions of the acinus (shown in figure above), what should be the distance between the two fibers?

b. Knowing what you've learned about biological aspects, what types of cells and soluble and insoluble factors would you choose to induce an environment of cell growth? What types of cells and soluble and insoluble factors would you choose to induce an environment of cell differentiation?



c. From the figure of the liver acinus shown above and in Table 6.9, the flow rate is 10^{-3} cm/s in the liver acinus. Given Darcy's law:

$$Q_r = AL_p[\Delta P - \Delta\pi]$$

where Q_r is the radial flow rate, A is the surface area, L_p is the hydraulic permeability (L_{p1} and L_{p2} are the inner and middle fiber hydraulic permeability, respectively), P is pressure, and π is osmotic pressure. If the bioreactor was filled with water and no cells, and flow is directed outward, flowing from the inner capillary compartment (ICC) outward, through the cell compartment to the extracapillary compartment (ECC), what would be the pressure in the three compartments if Q_r is 10^{-2} cm/s, L_{p1} is 10^{-4} mL/min/mmHg/mm^2, and L_{p2} is 10^{-6} mL/min/mmHg/mm^2? What are the pressures if each compartment flow went in the opposite direction from ECC into the ICC?

d. Assuming a slab configuration in the cell compartment, write a MATLAB program to give the oxygen profile if the diffusion coefficient in the cell matrix is 10^{-6}cm/s^2.

11. Use a compartmental model to calculate the number of mature cells produced from a single cell in a particular compartment. Use a doubling time of 24 hours ($\ln(2)/\mu$) and a mature cell half-life of 8 hours ($\ln(2)/k_d$). Assume that self-renewal can only take place in the first compartment with a probability of 0.5. Use a total of 10 compartments and calculate the number of cells as a function of time with the initial conditions

$$X_i(0) = 0, \text{ except } X_j(0) = 1$$

and vary j between 1 and 8. Plot all curves on the same plot. Discuss the implications of your results for transplantation.

12. Kinetics of differentiation/feedback control in a compartmental model.

Continued

i. Consider a six-stage differentiation process (N = 6), in which the last population, X_6, produces a cytokine, G, at a per-cell rate of q_G. This cytokine has a half-life of $t_{0.5}$ (= 2 hrs) and influences the growth rate of the stem cells—that is, $u_1 = fn([G])$, where [G] is the concentration of the growth factor G. Extend the base set of differential equations to describe the dynamics of [G].

ii. Incorporate into the equations

$$u_1([G]) = u/(1 + K[G])$$

where K is the binding constant for the growth factor. What does the function f([G]) describe physiologically?

iii. Make the equations dimensionless using the growth rate as the scaling factor for time, and K for the cell concentration.

iv. Describe the meaning of the dimensionless groups and estimate their numerical values.

v. Obtain the numerical values for a simulation starting from a single stem cell. Examine the effect of varying the numerical values of the parameters.

vi. Obtain the numerical values for a simulation starting from the steady-state solution and perturb the value of X_3 by 20 percent. Discuss your results.

References

[1] B. Alberts, D. Bray, et al., Molecular Biology of the Cell, Garland Publishing, New York, 1995.
[2] M. Brittberg, Treatment of deep cartilage defects in the knee with autologous chondrocyte transplantation, N. Engl. J. Med. 331 (1994) 889–895.
[3] L. Garby, J. Meldon, The Respiratory Functions of Blood, Plenum Medical Book Co, New York, 1977.
[4] S. Granick, R.D. Levere, Heme Synthesis in Erythroid Cells, Prog. Hematol. 4 (1964) 1–47.
[5] C.B. Harley, A.B. Futcher, C.W. Greider, Telomeres shorten during ageing of human fibroblasts, Nature 345 (1990) 458–460.
[6] A.V. Hoffbrand, J.E. Pettit, Sandoz Atlas: Clinical Haematology, Cower Medical Publishing, London, 1988.
[7] D.J. Knauer, H.S. Wiley, D.D. Cunningham, Relationship between epidermal growth factor receptor occupancy and mitogenic response. Quantitative analysis using a steady state model system, J. Biol. Chem. 259 (9) (1984) 5623–5631.
[8] M.R. Koller, B.O. Palsson, Review: tissue engineering: reconstitution of human hematopoiesis ex vivo, Biotechnol. Bioeng. 42 (8) (1993) 909–930.
[9] Lauffenburger, Linderman, Receptors: Models for Binding, Trafficking, and Signaling, Oxford University Press, New York, 1993.
[10] H. Lodish, J. Darnell, D. Baltimore, Molecular Cell Biology, W.H. Freeman & Company, New York, 1995.
[11] E. Lightfoot, Transport Phenomenon and Living Systems, Wiley, New York, 1974.
[12] J.M. Macdonald, J. Griffin, H. Kubota, L. Griffith, J. Fair, L.M. Reid, Bioartificial Livers, in: W. Kuhtreiber, R.P. Lanza, W.L. Chick (Eds.), Cell Encapsulation Technology and Therapeutics, Birkhauser Boston, Cambridge, MA, 1999, pp. 252–286.
[13] S.E. Mutsaers, J.E. Bishop, G. McGrouther, G.J. Laurent, Mechanisms of tissue repair: from wound healing to fibrosis 29 (1) (1997) 5–17.
[14] C.A. Peng, B.O. Palsson, Determination of specific oxygen uptake rates in human hematopoietic cultures and implications for bioreactor design, Ann. Biomed. Eng. 24 (3) (1996) 373–381.
[15] C. Peng, B. Palsson, Cell growth and differentiation on feeder layers is predicted to be influenced by bioreactor geometry, Biotechnol. Bioeng. 50 (1996) 479–492.
[16] J.P. Iredale, Tissue inhibitors of metalloproteinases in liver fibrosis, Intl J Biochem Cell Biol 29 (1) (1997) 43–54.
[17] A. Vander, J. Sherman, et al., Human Physiology: The Mechanisms of Body Function, McGraw-Hill, NY, 1994.

Suggested Readings

A. Atala, Future Perspective in Bladder Reconstruction, Adv. Exp. Med. Biol. 539 (Pt B) (2003) 921–940.
A. Atala, Tissue Engineering and Regenerative Medicine: Concept for Clinical Application, Rejuvenation Res. 7 (1) (2004) 15–31.
B.M. Baker, A.M. Handorf, L.C. Ionescu, W.J. Li, R.L. Mauck, New directions in nanofibrous scaffolds for soft tissue engineering and regeneration, Expert. Rev. Med. Devices 6 (5) (2009) 515–532.
S.P. Bruder, D.J. Fink, et al., Mesenchymal stem cells in bone development, bone repair, and skeletal regeneration therapy, J. Cell. Biochem. 56 (3) (1994) 283–294.
T. Burg, C.A. Cass, R. Groff, M. Pepper, K.J. Burg, Building off-the-shelf tissue-engineered composites, Philos. Transact. A Math. Phys. Eng. Sci. 368 (1917) (2010) 1839–1862.
D.L. Butler, S.A. Goldstein, R.E. Guldberg, X.E. Guo, R. Kamm, C.T. Laurencin, et al., The impact of biomechanics in tissue engineering and regenerative medicine, Tissue Eng Part B Rev 15 (4) (2009) 477–484.
E. Cattaneo, R. McKay, Identifying and manipulating neuronal stem cells, Trends Neurosci. 14 (8) (1991) 338–340.
E.M. Christenson, K.S. Anseth, J.J. van den Beucken, C.K. Chan, B. Ercan, J.A. Jansen, et al., Nanobiomaterial applications in orthopedics, J. Orthop. Res. 25 (1) (2007) 11–22.
J. Chute, J. Fung, et al., Ex Vivo Culture Rescues Hematopoietic Stem Cells with Long-Term Repopulating Capacity Following Harvest from Lethally Irradiated Mice, Exp. Hematol. 32 (3) (2004) 308–316.
W.P. Daley, S.B. Peters, M. Larsen, Extracellular matrix dynamics in development and regenerative medicine, J. Cell Sci. 121 (Pt 3) (2008) 255–264.
D. Fawcett, A Textbook of Histology, Saunders, PA, 1986.
M. Fröhlich, W.L. Grayson, L.Q. Wan, D. Marolt, M. Drobnic, G. Vunjak-Novakovic, Tissue engineered bone grafts: biological requirements, tissue culture and clinical relevance, Curr. Stem Cell Res. Ther. 3 (4) (2008) 254–264.
F.H. Gage, Neuronal stem cells: their characterization and utilization, Neurobiol. Aging 15 (Suppl. 2) (1994) S191.
D. Greber, M. Fussenegger, Mammalian synthetic biology: engineering of sophisticated gene networks, J. Biotechnol. 130 (4) (2007) 329–345.
R. Langer, J.P. Vacanti, Tissue engineering, Science 260 (5110) (1993) 920–926.
R. Langer, J.P. Vacanti, Artificial organs, Sci. Am. 273 (3) (1995) 130–133.
R.S. Langer, J.P. Vacanti, Tissue engineering: the challenges ahead, Sci. Am. 280 (4) (1999) 86–89.
R. Lanza, R. Langer, et al., Principles of Tissue Engineering, Academic Press, 2000.
K. Lee, C.K. Chan, N. Patil, S.B. Goodman, Cell therapy for bone regeneration—bench to bedside, J. Biomed. Mater. Res. B Appl. Biomater. 89 (1) (2009) 252–263.
H. Liu, F. DiCunto, et al., Citron kinase is a cell cycle-dependent, nuclear protein required for G2/M transition of hepatocytes, J. Biol. Chem. 278 (4) (2003) 2541–2548.
A.W. Lund, B. Yener, J.P. Stegemann, G.E. Plopper, The natural and engineered 3D microenvironment as a regulatory cue during stem cell fate determination, Tissue Eng. Part B Rev. 15 (3) (2009) 371–380.
B. Machalinski, M. Paczkowski, et al., An Optimization of Isolation of Early Hematopoietic Cells from Heparinized Cadaveric Organ Donors, Transplant. Proc. 38 (8) (2003) 3096–3100.
A. Mautes, J. Liu, et al., Regional Energy Metabolism Following Short-Term Neural Stem Cell Transplantation into the Injured Spinal Cord, J. Mol. Neurosci. 24 (2) (2004) 227–236.
T.C. McDevitt, S.P. Palecek, Innovation in the culture and derivation of pluripotent human stem cells, Curr. Opin. Biotechnol. 19 (5) (2008) 527–533.
C.M. Metallo, S.M. Azarin, L. Ji, J.J. de Pablo, S.P. Palecek, Engineering tissue from human embryonic stem cells, J. Cell. Mol. Med. 12 (3) (2008) 709–729.
A.G. Mikos, S.W. Herring, P. Ochareon, J. Elisseeff, H.H. Lu, R. Kandel, et al., Engineering complex tissues, Tissue Eng. 12 (12) (2006) 3307–3339.
B. Naughton, The importance of stroma cells, in: J. Bronzino (Ed.), CRC Handbook on Biomedical Engineering, CRC Press, Boca Raton, FL, 1995, pp. 1710–1726.
B. Palsson, S. Bhatia, Tissue Engineering, Pearson Prentice Hall, Upper Saddle River, NJ, 2004.
C. Paquet, D. Larouche, F. Bisson, S. Proulx, C. Simard-Bisson, M. Gaudreault, et al., Tissue engineering of skin and cornea: Development of new models for in vitro studies, Ann. N. Y. Acad. Sci. 1197 (2010) 166–177.
T. Petersen, L. Niklason, Cellular lifespan and regenerative medicine, Biomaterials 28 (26) (2007) 3751–3756.
J. Polak, L. Hench, Gene therapy progress and prospects: in tissue engineering, Gene Ther. 12 (24) (2005) 1725–1733.

L.M. Reid, Defining hormone and matrix requirements for differntiated epithelia, in: J.W. Pollard, J.M. Walker (Eds.), Basic Cell Culture Protocols, vol. 75, Humana Press, Inc., Totowa, New Jersey, 1990 (Chapter 21), pp. 237–262.
L.S. Sefcik, J.L. Wilson, J.A. Papin, E.A. Botchwey, Harnessing systems biology approaches to engineer functional microvascular networks, Tissue Eng. Part B Rev. 16 (3) (2010) 361–370.
M. Singh, C. Berkland, M.S. Detamore, Strategies and applications for incorporating physical and chemical signal gradients in tissue engineering, Tissue Eng. Part B Rev. 14 (4) (2008) 341–366.
J.P. Stegemann, S.N. Kaszuba, S.L. Rowe, Review: advances in vascular tissue engineering using protein-based biomaterials, Tissue Eng. 13 (11) (2007) 2601–2613.
J.P. Stegemann, H. Hong, R.M. Nerem, Mechanical, biochemical, and extracellular matrix effects on vascular smooth muscle cell phenotype, J. Appl. Physiol. 98 (6) (2005) 2321–2326.
D.L. Stemple, D.J. Anderson, Isolation of a stem cell for neurons and glia from the mammalian neural crest, Cell 71 (6) (1992) 973–985.
S. Stephan, S.G. Ball, M. Williamson, D.V. Bax, A. Lomas, C.A. Shuttleworth, et al., Cell-matrix biology in vascular tissue engineering, J. Anat. 209 (4) (2006) 495–502.
R. Susick, N. Moss, et al., Hepatic progenitors and strategies for liver cell therapies, Ann. N. Y. Acad. Sci. (2001), In Press.
C.T.-K.K. Tateno, C. Yamasaki, H. Sato, K. Yoshizato, Heterogeneity of growth potential of adult rat hepatocytes in vitro, Hepatology 31 (2000) 65–74.
C. Tateno, K. Yoshizato, Growth potential and differentiation capacity of adult rat hepatocytes in vitro, Wound Repair Regeneration 7 (1) (1999) 36–44.
D.A. Taylor, From stem cells and cadaveric matrix to engineered organs, Curr. Opin. Biotechnol. 20 (5) (2009) 598–605.
M.W. Tibbitt, K.S. Anseth, Hydrogels as extracellular matrix mimics for 3D cell culture, Biotechnol. Bioeng. 103 (4) (2009) 655–663.
A. Xu, T. Luntz, et al., Liver stem cells and lineage biology, in: R. Lanza, R. Langer, J. Vacanti (Eds.), Principles of Tissue Engineering, Academic Press, San Diego, 2000.

Glossary

Allogeneic Transplantation of cells or tissues from a donor into a recipient of the same species but a different strain.
Apoptosis A cellular process of aging and that leads to cell death. This process is initiated by the cell itself.
Autologous transplant Cells or tissue removed from a donor and then given back to the donor.
BMT Bone marrow transplantation.
Cellular therapies The use of grafted or transfused primary human cells into a patient to affect a pathological condition.
Chondrocytes Cells found in cartilage.
Colony-forming assay Assay carried out in semisolid medium under growth factor stimulation. Progenitor cells divide, and progeny are held in place so that a microscopically identifiable colony results after two weeks.
Cytokinesis The process occurring after DNA synthesis and resulting in completion of cell division from one cell into two.
Differentiation The irreversible progression of a cell or cell population to a more mature state. Distinct sets of genes are expressed and at varying levels in the cells during the differentiation process.
Engraftment The attainment of a safe number of circulating mature blood cells after a BMT.
Extracellular matrix (ECM) An insoluble complex of proteins and carbohydrates found between cells. These proteins serve to physically connect cell populations and to persistently signal cells to behave in specific ways.
Flow cytometry Technique for cell analysis using fluorescently conjugated monoclonal antibodies which identify certain cell types. More sophisticated instruments are capable of sorting cells into different populations as they are analyzed.
Functional subunits The irreducible unit in organs that gives tissue function—that is, alveoli in lung and nephron in kidney.

Graft-vs-Host disease The immunologic response of transplanted cells against the tissue of their new host. This response is often a severe consequence of allogenic BMT and can lead to death (acute GVHD) or long-term disability (chronic GVHD).

Hematopoiesis The regulated production of mature blood cells through a scheme of multilineage proliferation and differentiation.

Hyperplasia Growth process involving complete cell division, both DNA synthesis and cytokinesis

Hypertrophy Growth process involving DNA synthesis but absence of cytokinesis and resulting in cells of higher ploidy which, secondarily, causes cells to become larger. The late stages of many, if not most, tissue lineages have cells that undergo hypertrophy in response to regenerative stimuli.

Lineage Refers to cells at all stages of differentiation leading to a particular mature cell type.

Long-term culture-initiating cell Cell that is measured by a 7- to 12-week in vitro assay. LTC-IC are thought to be very primitive, and the population contains stem cells. However, the population is heterogeneous so not every LTC-IC is a stem cell.

Mesenchymal cells Cells of connective type tissue, such as fibroblasts, osteoblasts (bone), chondrocytes (cartilage), adipocytes (fat), and so on.

Microenvironment Refers to the environment surrounding a given cell in vivo.

Mitosis The cellular process that leads to cell division.

Mononuclear cell Refers to the cell population obtained after density centrifugation of whole bone marrow. This population excludes cells without a nucleus (erythrocytes) and polymorphonuclear cells (granulocytes).

Myoablatron The death of all myeloid (red, white, and platelet) cells, as occurs in a patient undergoing high dose chemotherapy.

Parenchymal cells The essential and distinctive cells of a particular organ (i.e., hepatocytes in the liver or myocytes in muscle).

Progenitor cells Unipotent cells that derive from stem cells and will differentiate into mature cells.

Self-renewal Generation of a daughter cell with identical characteristics to the parent cells. Most often used to refer to stem cell division, which results in the formation of new stem cells.

Stem cells Pluripotent cells that are capable of self-replication (and, therefore, unlimited proliferative potential). Malignant tumor cells are aberrant forms of stem cells.

Stromal cells Mesenchymal cells that partner with epithelial cells. They are age- and tissue-specific. Their roles in regulating the expansion and/or differentiation of epithelia have long been known. For example, heterogeneous mixture of support or accessory cells of the BM, also referred to as the adherent layer, is requisite for BM cultures.

Syngeneic Transplantation of cells or tissues from a donor into a genetically identical recipient.

Xenogeneic Transplantation of cells or tissues from a donor of one species into a recipient of a different species.

This page intentionally left blank

CHAPTER

7

Compartmental Modeling

John D. Enderle, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Explain the concept of a compartment.
- Quantitatively describe the transfer of substances separated by a membrane.
- Describe how osmosis affects cell volume.
- Analyze a physiological system using a one-compartment model.
- Describe and calculate the half-life of a substance in a one-compartment model.
- Use the washout curve to estimate parameters.
- Describe a pharmacokinetic model and its use.
- Describe how to maintain the concentration of a drug in the body by repeat dosages.
- Use a two-compartment model with sources and sinks.
- Use a three-compartment model with sources and sinks in a general, mammillary, and unilateral configuration.
- Use a multicompartment model with sources and sinks in a general, mammillary, unilateral, and catenary configuration.

7.1 INTRODUCTION

Compartmental modeling describes the movement of a substance from one compartment to another. Its origins are based on the metabolism of tracer-labeled compound studies in the 1920s. As we will see, compartmental modeling is a special case of physiological modeling, which is covered elsewhere in this book. Primarily, it is concerned with maintaining correct chemical levels in the body and their correct fluid volumes. A compartment can be a volume (or space) or the amount of a substance in a volume. Both representations are commonly used. Here we use the amount of substance in a volume as a compartment, with each compartment assumed to be homogeneous, as described later in the chapter. The process of transfer of substance from one compartment to another is based on diffusion and mass conservation. As shown, compartmental analysis provides a uniform theory that can be systematically applied to many linear and nonlinear systems. While interest in compartmental analysis here is focused on the human body, other engineers and scientists use this technique in studying evolution, carcinogenesis, chemical reactions, infectious disease models, and even semiconductor design and fabrication.

Before investigating compartmental modeling, Fick's Law of diffusion and osmosis is presented first from basic principles. Next, the volume of a cell and capillary diffusion are discussed. The basics of compartmental modeling are described from simple to more complex models that are increasingly more realistic. Much of the material in this chapter is based on the book by Godfrey.

7.2 SOLUTES, COMPARTMENTS, AND VOLUMES

When analyzing systems of the body characterized by a transfer of a solution from one compartment to another, such as the respiratory and circulatory systems, it is convenient to describe the system as a finite series of interconnected compartments. A *solution* is defined as a homogeneous mixture of two or more substances in any of the three states of matter: gas, liquid, or solid. Within a solution, we may have a mixture of matter—for instance, solid within a liquid, gas within a liquid, and so on. A solution is described by a component called a solute and another called the solvent. While there are no absolute rules regarding which component is the solute and which is the solvent, we typically call the component in the lesser amount the solute and the other the solvent. For instance, blood is a fluid that consists of 90 percent water with suspended cells such as red blood cells (erythrocytes), white blood cells (leukocytes), platelets, small molecules (i.e., glucose and carbon dioxide), large proteins, and electrolytes (i.e., sodium, calcium, magnesium, potassium). The solute could be a particular protein, and the solvent is the blood minus the protein.

The following are some of the readily identifiable compartments in the human body:

- Cell nucleus that is separated from the cytoplasm of the cell
- Internal organelle volumes, such as the mitochondria, endoplasmic reticulum, and so on that are separated from the cytoplasm
- Cell volume that is separated from the extracellular space by the cell membrane

- Interstitium or interstitial volume[1] that is separated from the plasma[2] by the capillary walls
- Plasma that is separated from the blood

Variables tracked in compartmental analysis are typically quantity or concentration of a solute, temperature, and pressure. Substances of interest are exogenous, such as a drug or tracer, or endogenous, such as glucose or an enzyme or hormone like insulin. Radioactive and stable isotopes[3] are used to track the dispersion in a compartmental system and are easily measured. Often a tracer dose of a radioactive isotope is used so the radiation emitted is small and does not interfere with the system. Typically, a tracer dose is less than 1 percent of the total amount of solute in the compartment. High-performance liquid chromatography (HPLC) is used to measure proteins and other macromolecules. Radioimmunoassay is used to measure hormones or proteins, a method based on the immune response of the body to an antigen, which is then bound to an antibody. Other modes of tracking involve injecting a dye (e.g., Evans blue) at one site in the cardiovascular system and measuring the concentration at one or more sites as a function of time.

Fluid in the body is separated into intracellular and extracellular fluid. A typical 70 kg adult contains 42 L of fluid, which is approximately 60 percent of the total body weight. Fluid in the body is tightly regulated so a relatively constant fluid volume is maintained. Intracellular fluid consists of the fluid inside the approximately 75 trillion cells in the body, totaling about 28 L and 40 percent of total body weight. Extracellular fluid consists of two major compartments, the interstitial fluid compartment and the plasma, and two minor compartments, the transcellular fluid and the lymph. The interstitial fluid is 11 L and the plasma is 3 L. The transcellular fluid includes fluids from the synovial, peritoneal, pericardial, intraocular spaces, and the cerebrospinal fluid. These compartments contain approximately 1 to 2 L of fluid. Lymph is the fluid that originates in the interstitial fluid that diffuses into the lymphatic system through lymph capillaries. It returns to the venous plasma after passing through the lymph nodes and has a volume of approximately 1 L. The blood in the circulatory system is a mixture of intracellular and extracellular fluid totaling 5 L of fluid and is 7 percent of total body weight. It consists of the plasma (3 L) and the fluid in the red blood cells (2 L). Table 7.1 summarizes various volumes in the body.

Typically, we work with the plasma compartment rather than the blood compartment, except when dealing with the cardiovascular system. Fluids in the body continually flow from one compartment to another without much change in fluid volume. As will be described more thoroughly in Section 7.3.5, fluid moves from the plasma to the interstitial fluid through the arterial side of the capillary bed and returns from the interstitial fluid to the plasma from the venous side of the capillary bed. Approximately 10 percent of the interstitial fluid does not immediately return to the plasma but moves into the lymphatic system through lymph capillaries by diffusion. The lymphatic system acts like a second parallel circulatory system, with the lymph returning to the plasma after traveling through

[1]Interstitial volume is the fluid that bathes the cells and is separated from the blood by capillaries.

[2]Plasma is the noncellular part of the blood. The plasma volume bathes the blood cells.

[3]Isotopes of an atom have the same atomic number but differ in mass. Radioactive isotopes disintegrate with time and emit ionizing radiation.

TABLE 7.1 Fluid Volumes in a 70 Kg Adult

Compartments	Volume in L
Total Fluid (60% of total body weight)	42
Intracellular	28
Extracellular	14
Interstitial	11
Blood	5
Plasma	3
Transcellular	1–2
Lymph	1

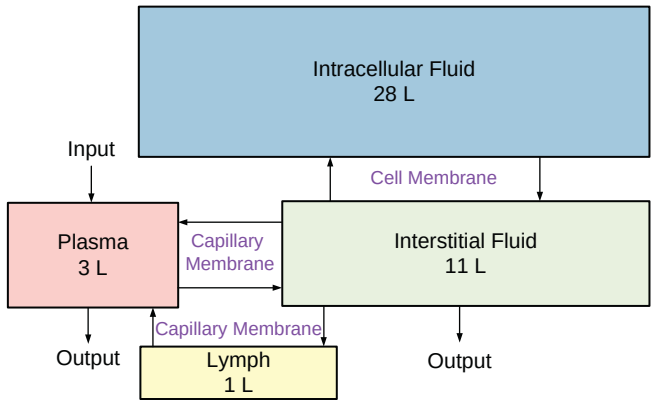


FIGURE 7.1 The compartment volumes of the body. A box depicts the volume. These volumes are tightly controlled by the body through mechanisms that will be described in this chapter. The arrows indicate a flow from one compartment into another (next to the arrows are the types of membranes the fluid must pass through). The rate of flow through a membrane depends on the properties of the membrane. The input includes the fluid ingested. Output is fluid lost from the kidneys, lungs, skin, and sweat, with a small amount lost in the feces. Not shown is the transcellular fluid.

the lymph nodes. The white blood cells in the lymph nodes monitor the lymph and destroy foreign substances to protect the body from disease. Figure 7.1 illustrates the relationships among the compartments.

7.3 TRANSFER OF SUBSTANCES BETWEEN TWO COMPARTMENTS SEPARATED BY A MEMBRANE

7.3.1 Diffusion and Membranes

Molecules and atoms randomly move due to thermal energy. They are never at rest even though the medium in which they are present is still. The speed of the molecules depends

on the temperature—the higher the temperature, the greater the speed of the molecules. The motion of a molecule is completely random and equally likely to move in any direction. If the motion is restricted to one axis, the probability of the molecule moving left or right is equal—that is, $p = 0.5$. This motion is often referred to as a *random walk.* As the molecule is moving, it collides with other molecules or the compartment wall, and then changes direction. The collision of one molecule into another molecule involves a transfer of kinetic energy. Sometimes collisions between molecules cause chemical reactions to occur, as will be described in a later chapter.

Consider a tube with a thin permeable membrane separating molecules from the right and left side at $t = 0$, as illustrated in Figure 7.2. As shown, the number of molecules on the left side is greater than the right side at $t = 0$. After a period of time, half the molecules on the left side move to the right and half of those on the right side move to the left. Thus, there is a net movement of molecules from the left to the right side, since there are more molecules on the left at $t = 0$. The number of molecules on each side are equal at $t = \infty$. We call this process simple diffusion. The rate of simple diffusion is impacted by the velocity of the molecules (temperature) and the number of channels in the membrane. A second type of diffusion is called facilitated diffusion or carrier-mediated diffusion. This requires the binding of carrier proteins to the molecule to move through the membrane and may involve the movement of molecules at rates greater than predicted by simple diffusion.

One typically works with concentrations rather than the number or quantity of molecules or ions, since measurements are made in concentration rather than quantity. The following relationship is used in moving between quantity and concentration:

$$\text{Concentration} = \frac{\text{Quantity}}{\text{Volume}}$$

The flow of particles due to diffusion is along the concentration gradient, with particles moving from high concentration areas to low ones. Physiological compartments are surrounded by membranes. Membranes provide structure, filter molecules and ions entering and leaving the cell, and control the cell volume. A cell membrane readily allows water, oxygen, and carbon dioxide to move across it, but prohibits other molecules and ions from passing through, except through protein channels. For example, a neuron's cell membrane consists of a lipid bilayer made of phospholipids, cholesterol, and proteins, which separates ions from the inside and outside of the cell, as shown in Figure 7.3. Cell membranes are selective, allowing some ions to pass through and not others. Figure 7.3 illustrates the use of a passive channel that allows only a particular ion to pass through by diffusion and

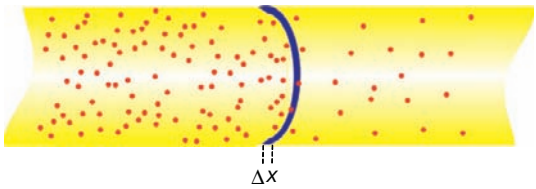


FIGURE 7.2 A tube filled with molecules and separated into two compartments by a thin membrane of width Δx at $t = 0$. The membrane has a number of channels that allow the molecules to pass through it. Assume that the left and right sides of the compartments are of equal volume.

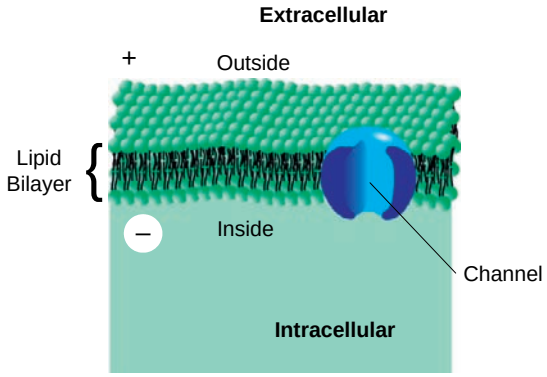


FIGURE 7.3 A typical cell membrane.

prevents other ions from passing through this channel. Passive channels are always open and driven by the concentration gradient.[4] Another type of channel is an active channel that uses energy from adenosine triphosphate (ATP) to move ions across the membrane against a concentration gradient. An active channel may be open all the time, or it may be open or closed based on the voltage across the membrane.

A third type of channel is called a pump, which exchanges ions from the inside to the outside using ATP. The neuron cell membrane uses an $Na^+ - K^+$ pump to remove Na^+ from inside the cell and replaces it with K^+ from the outside, as discussed in Sections 7.3.4 and 8.4.4. One of the most important functions of the pump is to control cell volume. If cell volume is not controlled, the cell will swell and burst. Approximately 60 percent of a neuron's energy requirements are used to maintain the $Na^+ - K^+$ pump.

The use of passive and active channels allows the concentration of ions to be different inside and outside the cell. Table 7.2 gives approximate concentrations across a mammalian cell membrane for some important positively charged ions (cations) and negatively charged ions (anions) in moles/liter.[5] The intracellular space (cytoplasm) has organelles with different ionic compositions. We will return to diffusion across the cell membrane in Section 7.3.4.

Another example of a membrane is the capillary wall. A capillary wall consists of endothelial cells that separate the interstitial volume from the blood volume. The typical length of a capillary is approximately 1 mm, with a diameter of 7 μm. Since the diameter of a

[4]Or an electric gradient as described in Chapter 12.

[5]A mole equals 6.02×10^{23} (Avogadro's number) particles and is defined as a unit in SI. The unit for the mole is M. For any ion or molecule, a mole's weight equals its atomic weight. So one mole of H_2O molecules weighs 18 g. Assuming a cell's volume is 1 nL with concentration of $Cl^- = 4\frac{\text{mM}}{\text{L}}$, the number of Cl^- ions inside the cell equals

$$0.004\ \frac{\text{M}}{\text{L}} \times 6.02 \times 10^{23}\ \frac{\text{molecules}}{\text{M}} \times 1 \times 10^{-9}\text{L} = 2.408 \times 10^{12}\ \text{molecules}$$

TABLE 7.2 Approximate Intracellular (Cytoplasm) and Extracellular Concentrations of the Important Ions across a Mammalian Cell Membrane

Ion	Intracellular (mM/L)	Extracellular (mM/L)
K^+	140	4
Na^+	10	142
Cl^-	4	103
Ca^{++}	0.0001	2.4

capillary is about the same size as a red blood cell, red blood cell movement through the capillary involves significant contact with the wall.

Pores in the capillary wall between cells allow movement between the interstitial volume and blood volume. Water freely moves through the capillary membrane. In fact, all other components of the plasma easily move through the capillary membrane except for some proteins such as albumin. While the interstitial fluid is similar to the plasma except for some proteins, movement of fluid through the interstitial volume is much slower than through the plasma. This is due to the structure of the interstitium that is maintained by collagen fiber bundles and proteoglycan filaments. Fluid flow through the network of filaments is mostly driven by diffusion. We will return to diffusion in capillaries in Section 7.3.5.

In this section we presented the cell as a static structure. In actuality, the cell is dynamic, and rapid changes in ionic concentrations are possible. The amount of water that moves in and out of the cell each second is 100 times the volume of the cell; this process is balanced so there is no net movement of water. Chapter 12 describes the neuron, which, when signaling, involves the rapid movement of sodium and potassium across the membrane. Chapter 4 describes the muscle and muscle contraction involving the rapid movement of calcium across the membrane.

7.3.2 Fick's Law of Diffusion

Fick's Law of diffusion describes the time course of the transfer of a solute between two compartments that are separated by a thin membrane, given by

$$\frac{dq}{dt} = -DA\frac{dc}{dx} \tag{7.1}$$

where

q = quantity of solute
A = membrane surface area
c = concentration
D = diffusion coefficient
dx = membrane thickness
$\frac{dc}{dx}$ = concentration gradient

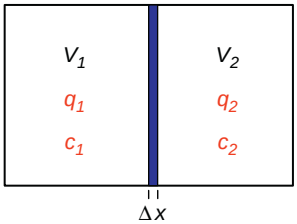


FIGURE 7.4 Two-compartment model with a membrane of width $\Delta x = dx$.

Consider the system of two compartments shown in Figure 7.4, where

V_1 and V_2 are the volumes of compartments 1 and 2
q_1 and q_2 are the quantities of solute in compartments 1 and 2
c_1 and c_2 are the concentrations of solute in compartments 1 and 2

and an initial amount of solute, Q_{10}, is dumped into compartment 1. After approximating the derivative $\frac{dc}{dx}$ as $\frac{c_1 - c_2}{\Delta x}$, the rate of change of solute in compartment 1 is given by

$$\dot{q}_1 = -DA\frac{dc}{dx} = -DA\frac{(c_1 - c_2)}{\Delta x} \tag{7.2}$$

Next, the quantity is converted into a concentration by

$$q_1 = V_1 c_1 \tag{7.3}$$

and after differentiating Eq. (7.3), gives

$$\dot{q}_1 = V_1 \dot{c}_1 \tag{7.4}$$

Substituting Eq. (7.4) into Eq. (7.2) yields

$$V_1 \dot{c}_1 = \frac{-DA}{\Delta x}(c_1 - c_2) \tag{7.5}$$

With the transfer rate K defined as

$$K = \frac{DA}{\Delta x}$$

when substituted into Eq. (7.5) yields

$$\dot{c}_1 = -\frac{K}{V_1}(c_1 - c_2) \tag{7.6}$$

From conservation of mass, we have

$$Q_{10} = q_1 + q_2$$

which after converting to a concentration gives

$$V_1 C_{10} = V_1 c_1 + V_2 c_2 \tag{7.7}$$

where $C_{10} = \frac{Q_{10}}{V_1}$ is the initial concentration in compartment 1 due to the initial amount of solute dumped into the compartment.

The concentration in compartment 2 is found from Eq. (7.7) as

$$c_2 = \frac{C_{10}V_1 - V_1c_1}{V_2} \tag{7.8}$$

which when substituted into Eq. (7.6) gives

$$\dot{c}_1 = \frac{-K}{V_1V_2}(V_2c_1 - V_1C_{10} + V_1c_1) = \frac{KC_{10}}{V_2} - \frac{Kc_1}{V_1V_2}(V_1 + V_2)$$

or

$$\dot{c}_1 + K\left(\frac{V_1 + V_2}{V_1V_2}\right)c_1 = \frac{KC_{10}}{V_2} \tag{7.9}$$

This is a first-order linear differential equation with forcing function

$$f(t) = \frac{KC_{10}}{V_2} \tag{7.10}$$

and initial condition $c_1(0) = C_{10}$.

Assume for simplicity that $V_1 = V_2$. Then Eq. (7.9) becomes

$$\dot{c}_1 + \frac{2K}{V_1}c_1 = \frac{KC_{10}}{V_1} \tag{7.11}$$

To solve Eq. (7.11), note that the root is $-\dfrac{2K}{V_1}$ and the natural solution is

$$c_{1_n} = B_1e^{\frac{-2Kt}{V_1}} \tag{7.12}$$

where B_1 is a constant to be determined from the initial condition. The forced response has the same form as the forcing function in Eq. (7.9), $c_{1_f} = B_2$, which when substituted into Eq. (7.11) yields

$$\frac{2K}{V_1}B_2 = \frac{KC_{10}}{V_1}$$

or

$$B_2 = \frac{C_{10}}{2}$$

Thus, the complete response is

$$c_1 = c_{1_n} + c_{1_f} = B_1e^{-\frac{2Kt}{V_1}} + \frac{C_{10}}{2}$$

To find B_1, the initial condition is used

$$c_1(0) = C_{10} = B_1e^{-\frac{2Kt}{V_1}}\Big|_{t=0} + \frac{C_{10}}{2} = B_1 + \frac{C_{10}}{2}$$

or

$$B_1 = \frac{C_{10}}{2}$$

The complete solution is

$$c_1 = \frac{C_{10}}{2}\left(e^{-\frac{2Kt}{V_1}} + 1\right)$$

for $t \geq 0$. Note that the concentration in compartment 2 is found using Eq. (7.8) as

$$c_2 = \frac{V_1 C_{10} - V_1 c_1}{V_2} = \frac{C_{10}}{2}\left(1 - e^{-\frac{2Kt}{V_1}}\right)$$

If $V_1 \neq V_2$, then

$$c_1 = \frac{C_{10}}{(V_1 + V_2)}\left(V_2 e^{-K\frac{(V_1+V_2)}{V_1 V_2}t} + V_1\right)u(t)$$

and

$$c_2 = \frac{V_1 C_{10}}{(V_1 + V_2)}\left(1 - e^{-K\frac{(V_1+V_2)}{V_1 V_2}t}\right)u(t)$$

At steady state, the concentrations on either side of the membrane are equal. In fact, if the volumes of the compartments are not equal, the concentrations at steady state are still equal. This should also be clear using Eq. (7.6); setting the derivative term equal to zero gives $c_1(\infty) = c_2(\infty)$. Note, however, that the number of moles of solute will be greater in the larger compartment.

7.3.3 Osmosis

Solutes and fluids must be maintained within a rigid tolerance in the body, both inside and outside the cell. Of all substances that move through the cell membrane, the most plentiful is water. Each second, the amount of water moving in and out of a cell is about 100 times the volume of the cell, with no net movement of water. Osmosis is the process that drives water across the membrane to maintain a zero water concentration gradient across the membrane. If the concentration gradient of water is not zero, then osmosis occurs to force it to zero. Obviously, any net movement of water through a cell membrane causes the cell to swell or shrink.

The distribution of water and solutes among the body's compartments is regulated by chemical and physical forces. The chemical potential is a function of concentration, pressure, and temperature. With regard to water concentration, pure water has a higher concentration than water mixed with a solute. Consider two compartments that are initially filled with water and different solute concentrations on either side of the membrane. Further assume that the membrane is not permeable to the solute. Osmosis causes a net movement of water along the concentration gradient, from the high concentration side to the low concentration side until the gradient is zero.

Pressure, defined as a force per unit area, is observed throughout the body for gases and liquids. The term *pressure* is reserved for gases and liquids, with *stress* used to describe the force per unit area for solids. When discussing pressure, we talk about a pressure difference

between two locations, measured in units of mm Hg.[6] Zero pressure is that observed in a vacuum, and a pressure relative to 0 pressure is called absolute pressure. Unless stated otherwise, pressure is given relative to standard atmospheric pressure. Systolic or peak blood pressure is typically around 120 mm Hg relative to standard atmospheric pressure, or 880 mm Hg in terms of absolute pressure. With regard to inspiration, the pressure within the lungs is -4 mm Hg, a negative pressure (with respect to standard atmospheric pressure) that allows air to move into the lungs.

Within the body, we identify relative pressures that drive solutes in or out of a compartment. For example, the pressure inside a capillary on the arterial end is approximately 30 mm Hg and inside a capillary on the venous end is approximately 10 mm Hg. Thus, a relative pressure difference of 20 mm Hg drives the blood from the arterial to the venous end of the capillary. On the arterial end of the capillary, the pressure in the interstitial fluid is approximately 17 mm Hg, with a relative pressure difference of 13 mm Hg that drives the plasma through the capillary walls into the interstitial fluid. On the venous end of the capillary, the pressure in the interstitial fluid is still approximately 17 mm Hg. Thus, a relative pressure of 7 mm Hg drives the interstitial fluid through the capillary walls into the plasma.

For the cell, the pressure difference between inside and outside is zero. Any pressure difference across the cell membrane causes a flow of water from high pressure to low pressure to equilibrate the pressure gradient.

As given by the ideal gas law, pressure is a function of temperature, volume, and the number of atoms and molecules. The ideal gas law assumes that there is no energy between atoms or molecules, such as attractive forces, and the only energy is kinetic. The motion of atoms and molecules create pressure as they collide with each other and the walls of the compartment—the faster the motion, the larger the pressure. Changes in temperature affect the motion of the particles. Increasing the temperature increases the speed of the atoms and molecules, which in turn, increases the pressure. Increasing the number of particles increases the pressure, since more collisions are possible. Pressure is inversely proportional to volume. Increasing the volume reduces the pressure, since there is more space for the particles to move, which reduces the number of collisions. Since temperature is highly regulated in the body, the affects of temperature changes are not a major consideration in osmosis for the body. To more fully appreciate osmosis, we first present two situations that are analyzed qualitatively. Following this, we quantitatively analyze osmosis.

Consider Figure 7.5, with water on both sides of the membrane. The two pistons allow pressures p_1 and p_2 to be applied to each compartment. If $p_1 > p_2$, Piston 1 drives water through the membrane from the left side to the right side. Suppose $p_1 = p_2$, and water is mixed with a small amount of solute on the left side and water is mixed with a large amount of solute on the right side. Osmosis causes water to move from the left side to the right side until the water concentrations are the same on both sides of the membrane.

[6]In physiology and medicine, the unit for measuring pressure is mm Hg. This unit is defined as the height of a column of mercury that can be supported for a given pressure. The pressure of atmosphere at sea level, called the standard atmospheric pressure, is a commonly used reference for pressure measurements and equals 760 mm Hg. The SI unit of pressure is the *Pascal* (Pa), where $1\text{Pa} = 1\frac{\text{N}}{\text{m}}$, and the U.S. common unit of pressure is $\frac{\text{lb}}{\text{in}^2}$(psi). At sea level, standard atmospheric pressure is 760 mm Hg $= 101.325$ k Pa $= 14.696\,\frac{\text{lb}}{\text{in}^2}$.

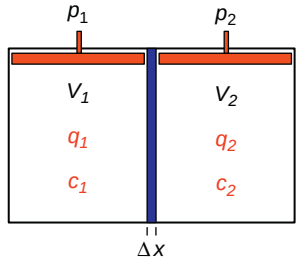


FIGURE 7.5 Two-compartment model with pistons and a membrane. Water can freely move through the membrane, but the solute cannot. The pistons create pressures on either side of the membrane. V_i is the volume, q_i is the quantity of solute, and c_i is the concentration of solute in compartment i. Δx is the width of the membrane.

Next, consider placing pure water into the left side and water mixed with a small amount of solute into the right side, initially with $p_1 = p_2$, and assume no change in volume is possible. Osmosis causes water to flow from the left to the right side. However, osmosis by itself does not allow the system to equalize water concentration, since there is no solute on the left side. As water moves across the membrane, it increases the number of water molecules on the right-hand side, which creates a higher pressure. As the pressure differential between right and left side increases, water is driven from the right to the left side. The pressure-driven water flow continues to increase until it is balanced by the osmosis-driven water flow. The pressure necessary to achieve this balance is called the osmotic pressure.

At this time, we wish to consider these situations from a quantitative point of view. Suppose each compartment in Figure 7.5 is filled with water. The relationship between flow of water through the membrane, Q, and the pressure difference is given by

$$p_1 - p_2 = R_m Q \tag{7.13}$$

where R_m is the resistance of the membrane to water.[7] If $p_1 > p_2$, then water flow is from left to right and equals $Q = \dfrac{p_1 - p_2}{R_m}$. It should be clear that if $p_1 = p_2$, then no net movement of water occurs through the membrane.

Next, consider that a small amount of solute is placed into the water on the left side and pure water is placed in the right side. The flow of water is now given by

$$(p_1 - p_2) - RTc_1 = R_m Q \tag{7.14}$$

where c_1 is the solute concentration, R is the ideal gas constant $\left(R = 62.3637 \ \dfrac{\text{L·mmHg}}{^\circ\text{K·mol}}\right)$, and T is the absolute temperature in degrees Kelvin (273° + centigrade°). Note that normal body temperature is 37°C or 310°K. The flow of water is now given by $Q = \dfrac{(p_1 - p_2) - RTc_1}{R_m}$. To have $Q = 0$ in Eq. (7.14) requires that

$$(p_1 - p_2) - RTc_1 = 0 \tag{7.15}$$

[7]This equation is similar to Ohm's law, where pressure is equivalent to voltage, flow is equivalent to current, and membrane resistance is equivalent to resistor resistance.

that is, a pressure difference is necessary to prevent a net flow of water from traveling to the left side. The pressure difference that causes $Q = 0$ is called the osmotic pressure, with $P_{osmotic} = (p_1 - p_2) = RTc_1$. The traditional symbol used to denote osmotic pressure is π, where $\pi = RTc_1$. If there is no pressure difference, then water is driven from the right side to the left side at a rate $Q = \dfrac{RTc_1}{R_m}$.

If there is solute on both sides of the membrane, then the osmotic pressure equals

$$\pi = RT(c_1 - c_2) \tag{7.16}$$

where c_1 and c_2 are the solute concentration on either side of the membrane. Equation (7.16) is called the van't Hoff law. If there are a number of different impermeable particles on either side of the membrane, then

$$(p_1 - p_2) - RT((c_{A_1} - c_{A_2}) + (c_{B_1} - c_{B_2}) + \cdots + (c_{Z_1} - c_{Z_2})) = R_m Q \tag{7.17}$$

and the total osmotic pressure equals the sum of the concentration differences for each particle (ion or molecule) on either side of the membrane

$$\pi = RT((c_{A_1} - c_{A_2}) + (c_{B_1} - c_{B_2}) + \cdots + (c_{Z_1} - c_{Z_2}))$$

EXAMPLE PROBLEM 7.1

Find the initial osmotic pressure at room temperature for a cell if the only ions present are *NaCl* on either side of the membrane. Assume the concentrations for Na^+ and Cl^- from Table 7.2 and that the ions cannot cross the membrane, and the cell volume is $V_i = 2$ nL.

Solution

Note first that the cell cannot withstand a pressure gradient across the membrane, so $p_i = p_o$. Therefore, we use Eq. (7.17) to find the initial osmotic pressure at room temperature as

$$\pi = RT\left([Na^+]_i - [Na^+]_o + [Cl^-]_i - [Cl^-]_o\right)$$

where $[Na^+]$ and $[Cl^-]$ are the concentrations of sodium and chlorine on the inside and outside of the membrane.[8] From Table 7.2, we substitute the concentrations into this equation, giving

$$\pi = 62.3637 \times 310((10 - 142) + (4 - 103)) = -4466 \text{ mmHg}$$

The osmotic pressure initially drives water out of the cell to equalize water concentration, which reduces the cell volume. This continues until the inside water concentration equals the outside water concentration. To compute the final cell volume size, note that the number of moles of Na^+ and Cl^- inside the cell remains constant, since it cannot pass through the membrane, and is

$$q_i = \left([Na^+]_i + [Cl^-]_i\right) \times V_i = (10 + 4) \times 2 \times 10^{-9} = 28 \times 10^{-9} \text{ M}$$

To have a zero osmotic pressure at steady state, we require that the inside steady state concentration, $c_{i_{ss}}$, equal the outside concentration, $c_{o_{ss}}$,

Continued

[8]We use the symbol $[I]$ to denote the concentration of ion "I." In this case, $[Na^+]$ and $[Cl^-]$.

$$c_{i_{ss}} = \left([Na^+]_o + [Cl^-]_o\right) = (142 + 103)\,\frac{\mathrm{M}}{\mathrm{L}} = c_{0_{ss}} = 245\,\frac{\mathrm{M}}{\mathrm{L}},$$

giving a new cell volume of $V_i = \dfrac{q_i}{c_{i_{ss}}} = \dfrac{28 \times 10^{-9}\mathrm{M}}{245\frac{\mathrm{M}}{\mathrm{L}}} = 0.114\,\mathrm{nL}$. Realistically, the cell cannot shrink to this degree, and there are other ions inside and outside the cell that move to maintain osmotic pressure at zero.

Osmotic pressure depends on the total number of particles per unit volume and not the size of the particle or its atomic weight. On average, all particles in the solution exert the same amount of pressure on the membrane; smaller particles move at higher velocities and larger particles move at lower velocities, providing about the same energy.

Since osmotic pressure does not depend on the weight of the particles and only the number of particles, the unit *osmole* (osm) is used. One osmole equals one mole of solute particles. If a molecule dissociates into two ions, then it contributes two osmoles; if a molecule dissociates into three ions, then it contributes three osmoles, and so on. For instance, one mole of glucose contributes 1 osm, since glucose does not dissociate into multiple ions; one mole of sodium chloride ($NaCl$) contributes 2 osm, since it dissociates into 2 ions; one mole of sodium sulfate (Na_2SO_4) contributes 3 osm, since it dissociates into 3 ions. Glucose and sucrose do not dissociate into ions, so their osmolarity equals the number of moles. Body fluids typically are observed in milliosmoles, so these units are expressed in mOsm.

Because concentrations are used in calculating osmotic pressures, we use the term *osmolarity* instead of osmoles. Osmolarity equals the number of osmoles per liter of solution. The osmolarity (in mOsm/L) inside a mammalian cell is approximately 301.2, plasma is 301.8, and the interstitial fluid is 300.8. The plasma has a slightly higher osmolarity than the interstitial fluid because of the proteins that do not diffuse through the capillary wall. For plasma and the interstitial fluid, 80 percent of the total osmolarity is due to $NaCl$. Within the cell, 50 percent of the total osmolarity is due to KCl. A solution containing one osmolarity contains one mole of undissociated particles per liter of solution. If a solute dissociates into more than one particle, then the osmolarity is given by

$$\text{Osmolarity} = \frac{\text{Moles} \times \text{Number of Dissociated Particles}}{\text{1L Solution}}$$

and osmotic pressure in terms of osmolarity is

$$\pi = c\left(\frac{\mathrm{osm}}{\mathrm{L}}\right) \times R\left(\frac{\mathrm{L}\cdot\mathrm{mmHg}}{^\circ\mathrm{K}\cdot\mathrm{mol}}\right) \times T(^\circ\mathrm{K}) = 62.3637 \times cT\ \mathrm{mmHg} \tag{7.18}$$

EXAMPLE PROBLEM 7.2

Find the osmolarity of 5 mM $NaCl$ in 1 L of solution.

Solution

Since $NaCl$ dissociates into Na^+ and Cl^- in solution, the osmolarity is equal to 2 times the moles of $NaCl$. Thus, a solution of $5\frac{\mathrm{mM}}{\mathrm{L}}$ $NaCl$ forms a $10\,\frac{\mathrm{mOsm}}{\mathrm{L}}$ solution.

EXAMPLE PROBLEM 7.3

Find the osmolarity and osmotic pressure of a 0.5% by weight solution of glucose at room temperature.

Solution

A 0.5% solution of glucose equals $\left(\frac{0.5}{100}\right)\frac{\text{g}}{\text{mL}} = 5\frac{\text{g}}{\text{L}}$. Since the molecular weight of glucose is $180\frac{\text{g}}{\text{mol}}$, we have a molarity of $\frac{5\frac{\text{g}}{\text{L}}}{180\frac{\text{g}}{\text{M}}} = 0.0278\ \frac{\text{M}}{\text{L}}$. Since glucose does not dissociate into separate ions in solution, the osmolarity is equal to the molarity of glucose, giving $27.8\ \frac{\text{mOsm}}{\text{L}}$. The osmotic pressure is therefore

$$\pi = cRT = 0.0278 \times 62.367 \times 310 = 537.5 \text{ mmHg}.$$

EXAMPLE PROBLEM 7.4

Consider a cell with an internal osmolarity of $300\ \frac{\text{mOsm}}{\text{L}}$ and volume of 2 nL, in a 30 nL solution of $300\ \frac{\text{mOsm}}{\text{L}}$. A 5 nL, 2% *NaCl* by weight solution is added to the extracellular space. Assuming that *NaCl* is impermeable and that the moles inside the cell do not change, describe the events that take place until steady state is achieved. What is the volume of the cell at steady state?

Solution

The first step in the solution is to determine the number of osmoles inside and outside the cell at $t = 0^-$. Inside the cell, we have $300\ \frac{\text{mOsm}}{\text{L}} \times 2\text{ nL} = 600 \times 10^{-9}$mOsm. Outside the cell, we have $300\ \frac{\text{mOsm}}{\text{L}} \times 30\text{ nL} = 9{,}000 \times 10^{-9}$mOsm.

Next, determine the number of osmoles outside the cell at $t = 0^+$. The 5 nL, 2% solution of *NaCl* added to the solution outside the cell has $\frac{2}{100}\ \frac{\text{g}}{\text{mL}} = 20\frac{\text{g}}{\text{L}}$, and with a molecular weight of $58.5\ \frac{\text{g}}{\text{M}}$, has a molarity of $\frac{20\frac{\text{g}}{\text{L}}}{58.5\frac{\text{g}}{\text{M}}} = 0.342\ \frac{\text{M}}{\text{L}}$. In 5 nL, there are $0.342{\times}5{\times}10^{-9} = 1.71{\times}10^{-9}$ moles of *NaCl*. Since *NaCl* separates into two ions, 3,420 $\times 10^{-9}$ mOsm are added to the solution outside the cell, giving a total of 12,420 mOsm. The outside osmolarity is then $\frac{12420}{35} = 355\ \frac{\text{mOsm}}{\text{L}}$. This difference in osmolarity causes an osmotic pressure of $\pi = \Delta cRT = (355 - 300) \times 62.367 \times 310 = 1{,}063$ mmHg, driving water out of the cell at room temperature.

At steady state, the total inside and outside cell volume is 37 nL, and total inside and outside osmoles equal $13{,}420{\times}10^{-9}$ mOsm. The total osmolarity then is $\frac{13420}{37} = 352\ \frac{\text{mOsm}}{\text{L}}$, which is the osmolarity inside and outside the cell. The final volume of the cell is therefore $\frac{600}{352} = 1.7$ nL.

7.3.4 Cell Volume and Osmosis

The cell regulates its volume by controlling the internal osmolarity through primary active transport mechanisms using ATP-driven pumps for sodium, potassium, calcium, chlorine, hydrogen, and other ions. The most important ATP-driven pump is the *Na-K* pump. Sixty to 70 percent of a cell's energy consumption is devoted to ATP pumps. Within the cell are anions (such as chlorine, proteins, nucleic acids, sulfate ions, and phosphate ions) and cations (such as potassium, calcium, and sodium). As we will see, some of these ions move freely through the membrane, some appear impermeable, and others are impermeable. The ATP pumps move ions through the membrane to maintain nonzero concentration gradients for many ions at steady-state.[9] Transport of ions is achieved using carrier proteins, which differs from diffusion. These pumps are used to keep the intracellular osmolarity of the cell equal to extracellular osmolarity and to maintain cell volume. A secondary active transport of ions not involved in ATP-driven pumps is a by-product of ion concentration gradients created through the primary active transport.

Any change in the steady-state ion concentrations causes water to be drawn into or be withdrawn from the cell by osmosis. Most mammalian cells have an osmolarity equal to plasma, approximately 300 mOsm. Because of its frailty, the cell membrane cannot survive even minimal pressure differences between the intracellular and extracellular fluids. The ATP pumps keep the pressure difference equal to zero and control the cell volume.

Two solutions with the same osmolarity are called isotonic. A solution that has a lower osmolarity to another is called hypotonic, and a solution that has a higher osmolarity to another is called hypertonic. If the osmolarities are not equal, water moves from the lower osmolarity side to the higher osmolarity side until the osmolarities are equal. Water will leave a cell that is hypotonic to the extracellular solution, which decreases the volume until an isotonic condition is achieved. Water will enter a cell if it is hypertonic to the extracellular solution, which increases the volume until an isotonic condition is achieved. ATP pumps also restore the cell volume to its original state.

Cell volume is typically maintained during extracellular osmolarity changes by adjusting the intracellular osmolarity. When the extracellular osmolarity is hypotonic with respect to the intracellular osmolarity, the following occur:

1. Cell volume immediately increases due to osmosis.
2. The ATP pumps drive ions out of the cell to restore cell volume while maintaining the osmotic balance.

When the extracellular osmolarity is hypertonic with respect to the intracellular osmolarity, the following occur:

1. Cell volume immediately decreases due to osmosis.
2. The ATP pumps drive ions into the cell to restore cell volume while maintaining the osmotic balance.

At steady-state, the ionic concentrations across the cell membrane are maintained by ATP pumps that create electrical and concentration gradients. The electrical gradient will be

[9]Note that the system is not in equilibrium, since the pump uses energy to maintain the concentrations. We refer to this situation as steady-state.

discussed in Chapter 12, and here we focus on the concentration gradient. For a typical mammalian cell, the ions of interest are K^+ and Na^+, with Table 7.2 listing the typical ion concentrations inside and outside the cell. To maintain ion concentrations at steady-state, the flow of each ion into the cell must be balanced by the flow of that ion out of the cell. First consider Na^+. The concentration gradient drives Na^+ into the cell. For K^+, the concentration gradient drives K^+ out of the cell. Osmotically, as long as the K^+ loss is balanced by the Na^+ gain, the cell remains isotropic. However, if the loss/gain were allowed to happen, the ionic concentrations across the cell membrane could not be maintained. To maintain the ionic concentrations of K^+ and Na^+, the cell membrane uses the ATP *Na-K* pump, driving Na^+ out of the cell and K^+ into the cell. Due to the action of the *Na-K* pump, Na^+ behaves as if it is impermeable to the membrane.

The *Na-K* pump also controls the volume of the cell. The cell contains a large number of impermeable proteins and molecules that have a negative charge, which attracts positively charged K^+ and Na^+, and water driven into the cell by osmosis. Without the *Na-K* pump, the cell would swell and eventually burst. The *Na-K* pump removes 3 Na^+ ions for every 2 K^+ ions pumped into the cell, thus creating a hypotonic condition that drives water from the cell by osmosis. If the cell changes volume, the *Na-K* pump operates to restore it by fine-tuning the flow of water into or out of the cell.

Consider the membrane illustrated in Figure 7.6, with two passive channels for K^+ and Na^+, an *Na-K* pump and an impermeable anion A^-. Assume that there is no pressure difference between the inside and outside, which is required for a mammalian cell. Thus, the flow equation using Eq. (7.17) is given by

$$-RT\left([K^+]_i - [K^+]_o + [Na^+]_i - [Na^+]_o + [A^-]\right) = R_m Q \tag{7.19}$$

In addition to the flow equation, the flux[10] equation for each of the permeable ions is given as

$$\begin{aligned} \vec{J}_K &= P_K\left([K^+]_i - [K^+]_o\right) - \vec{J}_p \\ \vec{J}_{Na} &= P_{Na}\left([Na^+]_i - [Na^+]_o\right) + \vec{J}_p \end{aligned} \tag{7.20}$$

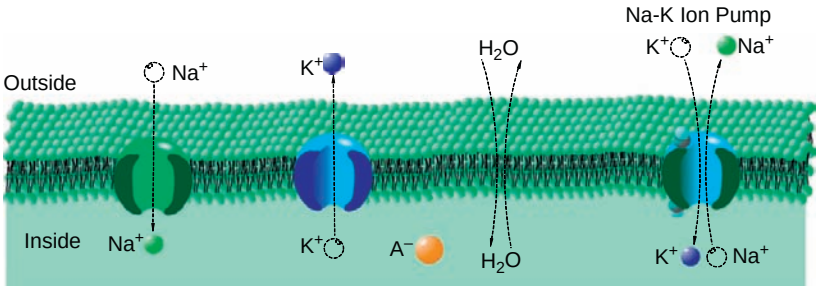


FIGURE 7.6 Cell membrane with two passive channels and a pump. Ions K^+ and Na^+ diffuse through the membrane, while anion A^- is impermeable. As indicated, water easily moves through the cell membrane.

[10]Flux is the number of particles (ions or molecules) that flow through a unit area per unit time. Flux is a vector that is given by the symbol $\vec{J}$.

where P_K and P_{Na} are permeabilities for K^+ and Na^+, and $\vec{J}_p$ is the pump flow rate. Permeability is a function of the membrane that describes the ease with which a particle moves through the cell membrane and is discussed more fully in Chapter 12. Potassium has a high permeability and sodium has a low permeability. To determine the cell volume, we replace $[A^-]$ in Eq. (7.19) with $\frac{a}{V}$, where a is the number of anions and V is the volume of the cell, giving

$$-RT\left([K^+]_i - [K^+]_o + [Na^+]_i - [Na^+]_o + \frac{a}{V}\right) = R_m Q \tag{7.21}$$

At steady state, the net flow of ions, Q, and water is zero, so Eqs. (7.20) and (7.21) reduce to

$$\begin{aligned} 0 &= P_K\left([K^+]_i - [K^+]_o\right) - \vec{J}_p \\ 0 &= P_{Na}\left([Na^+]_i - [Na^+]_o\right) + \vec{J}_p \end{aligned} \tag{7.22}$$

and

$$[K^+]_i - [K^+]_o + [Na^+]_i - [Na^+]_o + \frac{a}{V} = 0 \tag{7.23}$$

Equation (7.22) is solved for the concentrations, $[K^+]_i - [K^+]_o = \frac{\vec{J}_p}{P_K}$ and $[Na^+]_i - [Na^+]_o = -\frac{\vec{J}_p}{P_{Na}}$, and then substituted into Eq. (7.23), giving

$$\frac{\vec{J}_p}{P_K} - \frac{\vec{J}_p}{P_{Na}} + \frac{a}{V} = \vec{J}_p\left(\frac{P_{Na} - P_K}{P_{Na}P_K}\right) + \frac{a}{V} = 0 \tag{7.24}$$

Equation (7.24) is easily solved for V, yielding

$$V = \frac{aP_{Na}P_K}{\vec{J}_p(P_K - P_{Na})} \tag{7.25}$$

Positive cell volumes are possible in Eq. (7.25) when $P_K > P_{Na}$, which is the case for the mammalian cell membrane. Notice that the cell volume is inversely related to the pump rate, so as the pump rate increases, cell volume decreases. The cell carefully controls the pump rate so cell volume is maintained. Also, note that as the cell grows, evidenced by increasing the number of impermeable proteins and molecules (a in Eq. (7.25)), the volume of the cell increases according to Eq. (7.25). Finally, note that when the pump rate goes to zero (death), the cell volume heads to infinity, but before getting there, the cell membrane bursts.

7.3.5 Capillary Diffusion

The movement of water, nutrients, electrolytes, and other particles through the capillary wall is driven by osmotic and hydrostatic pressure. As we will see, these pressures cause fluid to flow out of the capillary at the arterial end and flow into the capillaries at the venous end.

Figure 7.7 illustrates a capillary network. To reach the capillaries, blood first flows from the heart to the aorta under high pressure. The blood leaves the aorta and flows into other arteries until it reaches the arterioles, the smallest branch of the arterial system. From the arterioles, blood flows through the capillaries, where diffusion into the interstitial volume occurs in the lower portion of Figure 7.7. In the upper portion of Figure 7.7, diffusion from the interstitial volume into the capillaries occurs, with the plasma then flowing into the venules, the smallest branch of the venous system. From the venules, blood flows through

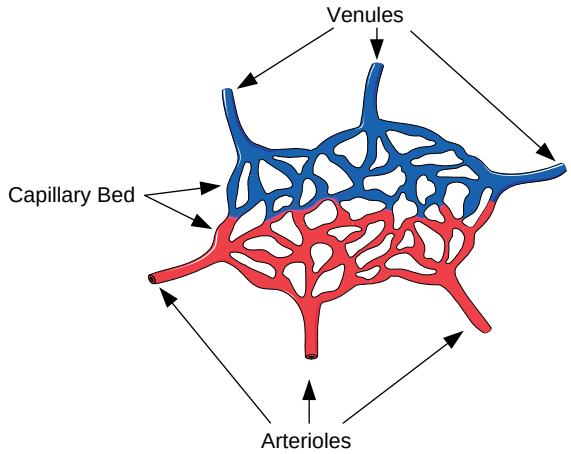


FIGURE 7.7 A capillary network. A sphincter is usually located at the arteriole's entry to the capillary bed that restricts the blood flow. Not all capillaries carry blood all the time, and blood flow is based on the needs of the tissue. For instance, less than 10 percent of the capillaries in muscle carry blood during rest.

TABLE 7.3 Percentage of Plasma Volumes in the Circulatory System

Volume	Percentage
Heart	7%
Pulmonary System	9%
Arteries	13%
Arterioles and Capillaries	7%
Veins and Venules	64%

larger and larger veins until it reaches the heart. Table 7.3 provides the volume percentages in each of the circulatory system compartments. Total plasma volume is approximately 3 L, and total interstitial volume is 11 L.

Plasma proteins are the only substances that are impermeable[11] to the capillary wall, which creates an osmotic pressure. The osmotic pressure created by plasma proteins is called the colloid osmotic pressure, named to differentiate it from the osmotic pressure across the cell membrane. The concentration of proteins in the plasma is 7.5 gm/dl and in the interstitial volume is 3 gm/dl. By far, the protein with the greatest contribution to the colloid osmotic pressure in the plasma/capillary space is albumin (21.8 mm Hg), followed by globulins (6 mm Hg), and then fibrinogen (0.2 mm Hg), giving a total colloid pressure of 28 mm Hg to move fluids inward. The interstitial colloid osmotic pressure due to all proteins within this space is 8 mm Hg. Thus, the total colloid osmotic pressure (difference between plasma/capillary and interstitial volume) is 20 mm Hg, a resultant force that moves fluid inward.

[11]Actually, some plasma proteins slowly leak through the capillary wall into the interstitial compartment, which are then returned to the plasma from lymphatic circulation.

Unlike the cell membrane, hydrostatic pressure exists across the capillary wall that changes as we move from the arteriole to the venule ends. The hydrostatic pressure at the arteriole end is 30 mm Hg and at the venule end is 10 mm Hg. The pressure drop occurs approximately linearly from the arteriole to the venule end of the capillary.

Summing all the pressures at the arteriole end gives the following:

Filtration Pressure	mm Hg
Hydrostatic Pressure	+30
Interstitial Colloid Osmotic Pressure	+ 8
Plasma Colloid Osmotic Pressure	−28
Net Outward Pressure	***+10***

Thus, a net outward pressure of 10 mm Hg exists that moves fluids out of the capillaries into the interstitial volume. Some refer to this pressure as the filtration pressure. The amount of fluid moved is approximately 5 percent of the total plasma.

Summing all the pressures at the venule end gives the following:

Reabsorption Pressure	mm Hg
Hydrostatic Pressure	+10
Interstitial Colloid Osmotic Pressure	+ 8
Plasma Colloid Osmotic Pressure	−28
Net Outward Pressure	***−10***

Thus, a net inward pressure of 10 mm Hg exists that moves fluids out of the interstitial volume into the capillaries. Some refer to this pressure as the reabsorption pressure. Almost all of the fluid that flows into the interstitial volume from the capillaries flows back into the capillaries from the interstitial volume. The remainder left in the interstitial volume flows into the lymphatic fluid, from which it returns to the plasma.

Mathematically, at either end of the capillary, the flow rate of fluid is given by

$$Q_A = \frac{\Delta p_A - RTc_C + RTc_I}{R_m} \tag{7.26}$$

$$Q_V = \frac{\Delta p_V - RTc_C + RTc_I}{R_m} \tag{7.27}$$

where Q_A is the flow rate at the arteriole end into the interstitial fluid, Q_V is the flow rate at the venule end into the capillary, Δp_A is the hydrostatic pressure at the arteriole end, Δp_V is the hydrostatic pressure at the venule end, c_C is the concentration of proteins in the plasma/capillary, and c_I is the concentration of proteins in the interstitial fluid.

With the rate of plasma protein leakage into the interstitial volume given by $\dot{V}_P$, and the rate of fluid that moves from the interstitial volume into lymphatic volume given by $\dot{V}_L$, the change in interstitial volume, denoted $\dot{V}_I$, is given by

$$\begin{aligned}\dot{V}_I &= Q_A - Q_V + \dot{V}_P - \dot{V}_L \\ &= \frac{\Delta p_A - RTc_C + RTc_I}{R_m} - \frac{\Delta p_V - RTc_C + RTc_I}{R_m} + \dot{V}_P - \dot{V}_L\end{aligned} \tag{7.28}$$

Equation (7.28) is useful in studying the effects of pathological conditions. In some types of trauma, Δp_A increases, causing an interstitial volume increase (swelling). If plasma protein concentration, c_C, increases, then the flow rate into the interstitial volume decreases.

7.4 COMPARTMENTAL MODELING BASICS

In the previous section, we worked from basic principles to examine diffusion and osmosis. Here we use a systematic approach called compartmental modeling to describe the movement of a solute through a system. Compartmental modeling involves describing a system with a finite number of compartments, each connected by a flow of solute from one compartment to another. Movement of the solute can be any of the following:

1. Among organelles in a cell
2. A cell and the extracellular space
3. An organ and the interstitial space
4. Among organs through the circulatory system

In modeling a system with compartments in this chapter, we consider a lumped parameter system rather than a distributed system. Therefore, we work with ordinary differential equations. More accurate distributed system models have solutions that involve partial differential equations. In many cases, however, we will find that the more complex distributed system models are quite accurately described by compartmental models using a large number of compartments.

Compartmental modeling applied to the human body is usually a gross simplification of the inherent underlying processes, and there may be a limited anatomical relationship between the true system and the model. For example, we may define the compartments of the body as the plasma compartment (that includes all noncellular fluids in all the blood vessels), tissue compartment (that includes the fluid in 75 trillion cells), interstitial compartment, and the lymph compartment, as illustrated in Figure 7.1. Naturally, the tissue compartment can be further divided into organs and so forth, and the plasma compartment can be further divided into arteries, veins, and progressively smaller vessels. Nevertheless, we will see that compartmental modeling is very useful in describing the movement of a solute through the body, especially when small perturbations from steady-state are considered. Many important applications of compartmental analysis are found in pharmacokinetics.

The alternative to compartmental modeling involves modeling the system via a fluid model of the blood and lymph systems, and the transport phenomena of the solute in each organ. The problem with this approach is finding a model that is time dependent, 3-D, and distributed. The solute is not homogeneously distributed in the organs, and we do not have detailed information about the parameters that describe the model.

As usual in physiological modeling, identifying the number of compartments and the connections between compartments is the most difficult challenge, along with collecting appropriate data. Regardless of the simplicity of the model, there should be some relationship between the model and the system being modeled based on a priori knowledge. Furthermore, it should be possible to test the model after collecting data and compare its performance to the real system. After testing, the model is usually modified, typically

making it more complex. By adding complexity, we add additional parameters, which must be estimated based on the data collected using parameter estimation techniques.

Given a system described by a group of compartments, some exchange of solute (i.e., a radioactive tracer, a molecule like glucose or insulin, a gas like oxygen or carbon dioxide) is expected between compartments by diffusion. Compartmental analysis predicts the quantity or concentration of solutes under consideration in each compartment as a function of time using conservation of mass—that is, accumulation equals input minus output. The model may be linear, nonlinear, continuous, or discrete, and may even have time-varying or stochastic parameters. If the model is continuous and linear, then the change in solute concentration is described as a sum of exponential and sinusoidal terms.

The following assumptions are made when describing the transfer of a solute by diffusion between any two compartments:

1. The volume of each compartment remains constant.
2. Any solute q entering a compartment is instantaneously mixed throughout the entire compartment.
3. The rate of loss of a solute from a compartment is proportional to the amount of solute in the compartment times the transfer rate, K, given by Kq. The transfer rate typically has units of liters per minute.

If two solutes are being tracked in a system, the overall model can be described using two parallel models. For instance, if each solute flows in and out of the plasma, each model can have its own plasma compartment separate from the other. It follows that we can track n solutes, with each solute having its own model, all separately sharing the plasma. This concept follows with additional tissues and blood vessels.

From a modeling perspective, identifying compartments and the number of compartments to describe a system is a difficult step. Acquiring measurement data for model facilitation is another difficult step because some compartments are inaccessible. Both of these steps are beyond the scope of this book; interested readers can examine books listed at the end of this chapter for more information.

7.4.1 Inputs to a Compartmental System

The inputs to a compartmental system are discussed following.

Bolus Injection

A bolus injection is an immediate injection of a solute into a compartment. It is assumed that the injected solute instantaneously mixes with the solution in the compartment. Mathematically, a bolus is approximated as either a change in initial conditions or as an impulse function, $\delta(t)$.

Constant Continuous Infusion

A constant continuous infusion input is delivered by an infusion pump or an intravenous drip into a compartment. It is assumed that the injected solute instantaneously mixes with the solution in the compartment. Mathematically, a constant continuous infusion input is approximated as a unit step function, $u(t)$.

Encapsulated Pill or Hypodermic Needle Injection

An encapsulated pill or hypodermic needle injection provides a constant continuous input over a period of time. It is assumed that the diffused or injected solute instantaneously mixes with the solution in the compartment. Mathematically, an encapsulated pill or hypodermic needle injection is approximated by a pulse function, $u(t)-u(t-t_1)$, where t_1 is the duration of the pulse.

Depending on the type of input, the solute moves through the body via the circulatory system. We assume instantaneous mixing as the solute enters the system by action of the heart. In some situations, the solute enters the body through the digestive system and then through the plasma. The solute diffuses out of the circulatory system into the other compartments of the body. In general, the elimination of a solute from the body occurs through the kidneys; intestines; lungs, skin, and sweat; biotransformation (converted to another form) in the liver and other organs; and metabolized in tissues.

7.5 ONE-COMPARTMENT MODELING

The simplest compartment model consists of only one compartment. A one-compartment model is shown in Figure 7.8, where a box is used to define the compartment, and the flow of solute is defined by arrows. The input to the compartment can be any of those described in Section 7.4.1 or other types of functions. The output transfer rate, K_{10}, depicts the flow of solute from compartment 1 to the environment space, 0. The convention used in writing the transfer rate, K_{ij}, describes the flow of solute leaving compartment i, and entering compartment j. All transfer rates are given by $K_{ij} \geq 0$. While only one output is shown in Figure 7.8, there can be multiple outputs to different spaces, such as the urine, liver, and so on, and multiple inputs. As we will see, all of the outputs in this case can be combined into a single output by summing the transfer rates into a single transfer rate if convenient. There can be more than one input to the compartment, and if so, each input can be solved for separately using superposition, with zero initial conditions, and the natural response to the initial conditions.

To analyze the system in Figure 7.8, we use conservation of mass to write the differential equation describing the rate of change of the quantity of solute in the compartment, given as accumulation = input – output, where

$$\text{Accumulation} = \dot{q}_1$$
$$\text{Input} = f(t)$$
$$\text{Output} = K_{10}q_1$$

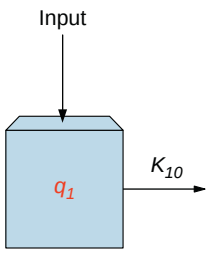


FIGURE 7.8 A one-compartment model. Assume the volume of the compartment is V_1. The input is $f(t)$.

Thus,

$$\dot{q}_1 = f(t) - K_{10}q_1 \tag{7.29}$$

The solution of Eq. (7.29) consists of the natural and forced responses. The natural response has the form $q_{1_n} = B_1 e^{-K_{10}t}$, where the root is $-K_{10}$. The forced response takes the form of the input. For instance, if the input is a constant continuous infusion, $u(t)$, the forced response is $q_{1_f} = \dfrac{1}{K_{10}}$. Thus, the complete unit step response is $q_{1_u} = q_{1_n} + q_{1_f} = \left(B_1 e^{-K_{10}t} + \dfrac{1}{K_{10}}\right)u(t)$, where B_1 is determined from the initial condition of the system. If the initial condition is zero, then $q_{1_u} = \dfrac{1}{K_{10}}\left(1 - e^{-K_{10}t}\right)u(t)$. If the initial condition is not zero, then $B_1 = q_1(0) - \dfrac{1}{K_{10}}$. If the magnitude of the unit step input is ζ, then the response with zero initial conditions is $q_{1_u} = \dfrac{\zeta}{K_{10}}\left(1 - e^{-K_{10}t}\right)u(t)$, and with initial condition $q_1(0)$ is $q_{1_u} = \left(\dfrac{\zeta}{K_{10}} - \left(\dfrac{\zeta}{K_{10}} - q_1(0)\right)e^{-K_{10}t}\right)u(t)$.

If the input is a bolus injection, $\delta(t)$, the solution method begins the same as the unit step response with zero initial conditions, and then the bolus response is $q_{1_\delta} = \dfrac{dq_{1_u}}{dt} = e^{-K_{10}t}u(t)$. If the input is $\zeta\delta(t)$, then the response is $q_{1_\delta} = \zeta e^{-K_{10}t}u(t)$.

If the input is hypodermic needle injection, $u(t) - u(t - t_1)$ with zero initial conditions, then the response based on superposition is $q_{1_p} = \dfrac{1}{K_{10}}\left(1 - e^{-K_{10}t}\right)u(t) - \dfrac{1}{K_{10}}\left(1 - e^{-K_{10}(t-t_1)}\right)u(t - t_1)$. If the input is $\zeta(u(t) - u(t - t_1))$, then the response is

$$q_{1_p} = \frac{\zeta}{K_{10}}\left(1 - e^{-K_{10}t}\right)u(t) - \frac{\zeta}{K_{10}}\left(1 - e^{-K_{10}(t-t_1)}\right)u(t - t_1).$$

7.5.1 Half-Life

When tracking a solute in the body, the half-life of the concentration is an important metric, where the half-life is the time required to reduce the concentration by 50 percent from maximum. Consider the case in which a bolus of $\zeta\delta(t)$ is injected into the system in Figure 7.8, with solution $q_{1_\delta} = \zeta e^{-K_{10}t}u(t)$. To find the half-life, $t_{\frac{1}{2}}$, we convert to concentration by letting $c_1 = \dfrac{q_{1_\delta}}{V_1} = \dfrac{\zeta}{V_1}e^{-K_{10}t}u(t)$, then set $c_1\left(t_{\frac{1}{2}}\right) = \dfrac{\zeta}{2V_1}$, and solve for $t_{\frac{1}{2}}$. Thus, we have

$$\frac{\zeta}{2V_1} = \frac{\zeta}{V_1}e^{-K_{10}t_{\frac{1}{2}}}$$

Taking the natural logarithm gives

$$\ln\left(\frac{1}{2}\right) = -K_{10}t_{\frac{1}{2}}$$

or

$$t_{\frac{1}{2}} = \frac{\ln(2)}{K_{10}} = \frac{.693}{K_{10}} \tag{7.30}$$

7.5.2 Washout Curve

A washout curve is a useful experimental technique for parameter estimation. The experiment begins with a continuous infusion of solute until the concentration reaches steady state. After reaching steady state, the input is stopped; the subsequent decay of solute from its maximum back to zero is called the washout curve. In essence, the input is a pulse input, $\zeta(u(t)-u(t-t_1))$, with a sufficiently large t_1 $\left(\text{much larger than } \frac{5}{K_{10}}\right)$, which has the solution

$$c_1 = \frac{q_{1_p}}{V_1} = \frac{\zeta}{V_1 K_{10}}\left(1 - e^{-K_{10}t}\right)u(t) - \frac{\zeta}{V_1 K_{10}}\left(1 - e^{-K_{10}(t-t_1)}\right)u(t-t_1). \tag{7.31}$$

The washout curve is the portion of the response from t_1 to ∞, which is essentially an impulse response beginning at t_1; ignoring the solution from 0 to t_1, and resetting time t_1 to 0, the washout curve is rewritten as

$$c_{1_{wo}} = \frac{\zeta}{V_1 K_{10}} e^{-K_{10}t}u(t) \tag{7.32}$$

From the experimental data and using Eq. (7.32), the time constant from the data gives K_{10}, and the initial value and time constant gives V_1. Note that the washout curve technique described here only works for one-compartment models.

The following example considers the uptake of radioactive iodine (I^{131}) by the thyroid gland in a very simplified model as depicted in Figure 7.9. We will return to modeling the thyroid system in Section 7.7.4. The compartment represents the plasma. The thyroid gland, located in the neck, absorbs iodine from blood and uses it to produce an iodine-containing hormone, primarily thyroxine, which is then secreted into the blood. Thyroxine increases the chemical reaction rate in the cell and overall metabolic rate of the body. The level of thyroxine is closely regulated by the hypothalamus and the pituitary gland. The hypothalamus releases thyroid releasing hormone (TRH) in reaction to thyroid levels in the blood. TRH stimulates the pituitary gland to release of thyroid-stimulating hormone (TSH), which causes a reaction in the thyroid. Thus, the thyroid gland is used to regulate the body's

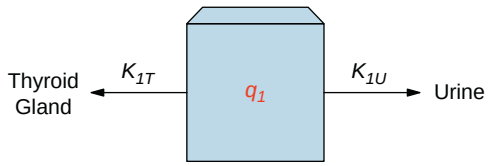


FIGURE 7.9 Illustration for Example Problem 7.5, where q_1 equals quantity of I^{131} in the plasma compartment.

metabolism. If the level of thyroxine is too low, TSH is released by the pituitary gland, which increases the release of thyroxine by the thyroid gland. If the level of thyroxine is too high, then less TSH is released, which reduces the production of thyroxine by the thyroid gland. To examine thyroid disorders, radioactive iodine (I^{131}) is given to the patient, and its uptake by the thyroid gland is measured. Simply put, too much iodine uptake is indicative of hyperthyroidism (too much thyroxine is produced) or too little uptake is indicative of hypothyroidism (too little thyroxine is produced). Removal of I^{131} from the plasma is via the urine by the kidneys or taken up by the thyroid gland.

EXAMPLE PROBLEM 7.5

Consider the removal of I^{131} for the plasma compartment shown in Figure 7.9. Assume that a bolus of I^{131} is injected into the plasma. I^{131} then moves from the plasma into the thyroid gland or is excreted into the urine. Assume the injected quantity of I^{131} in the plasma is $q_1(0)$. Find the response $q_1(t)$.

Solution

Our solution is easily considered by modeling the bolus as a change in initial condition at $t = 0$, and a zero input. Using conservation of mass, the differential equation describing the rate of change of the quantity of I^{131} in the plasma compartment is

$$\dot{q}_1 = -(K_{1T} + K_{1U})q_1$$

The solution is

$$q_1 = q_1(0)e^{-(K_{1T}+K_{1U})t}u(t)$$

Notice that the effect of the removal of I^{131} from the plasma is easily carried out with a single transfer rate, $K_{10} = K_{1T} + K_{1U}$, with no change in the outcome. However, if we are interested in tracking the uptake of I^{131} in the thyroid gland, separating the removals with two different transfer rates allow us to track it.

This example is also solved by applying an input of $q_1(0)\delta(t)$ with a zero initial condition. We start by solving for the response to $u(t)$, $q_{1_u}(t)$, differentiating the unit step response to yield the $\delta(t)$ response, $q_{1_\delta}(t)$, and scale by $q_1(0)$ to give $q_1(t)$.

Conservation of mass yields the following differential equation,

$$\dot{q}_1 = u(t) - (K_{1T} + K_{1U})q_1$$

which has the solution

$$q_{1_u} = \left(\frac{1 - e^{-(K_{1T}+K_{1U})t}}{K_{1T} + K_{1U}}\right)u(t)$$

The impulse response is

$$q_{1_\delta} = \dot{q}_{1_u} = e^{-(K_{1T}+K_{1U})t}u(t)$$

and scaled by $q_1(0)$, gives the response $q_1 = q_1(0)e^{-(K_{1T}+K_{1U})t}u(t)$. Note that treating a bolus as a change in initial conditions yields a much easier solution.

7.5.3 Pharmacokinetic Models

A pharmacokinetic model describes the movement of a drug or anesthetic agent to sites throughout the body, as shown in Figure 7.10. Much of the model focuses on drug absorption, distribution, metabolism, and excretion. Pharmacokinetic models usually have three or fewer compartments. Metabolic models have many more compartments than pharmacokinetic models, typically six or more compartments. Pharmacokinetic models usually are linear, whereas metabolic models are nonlinear in order to capture an adequate representation of the system. Pharmaceutical drugs do not appear naturally in the body and can be traced easily. Since metabolic solutes occur naturally in the body, following the movement of the solute is possible only with the use of isotopes.

Assume that the drug absorption site is the GI tract, where the drug passes to the blood via the liver. Since drugs are relatively small, they easily move through the capillary walls and into the interstitial fluid. From there, the drug then moves into the targeted organs and other tissues. As shown, the drug is removed via the kidneys, GI tract, and biotransformation in the liver and other sites. Transfer from compartment to compartment is through diffusion. The rate of movement from each compartment depends on the transfer rate. Typically, the removal of the drug from the body occurs more slowly than movement among the plasma, body fluids, and tissues. For example, tissues and organs that are highly perfused, such as the thyroid gland, liver, and kidney, have a large transfer rate, and tissues with low perfusion have a small transfer rate. If the removal transfer rates

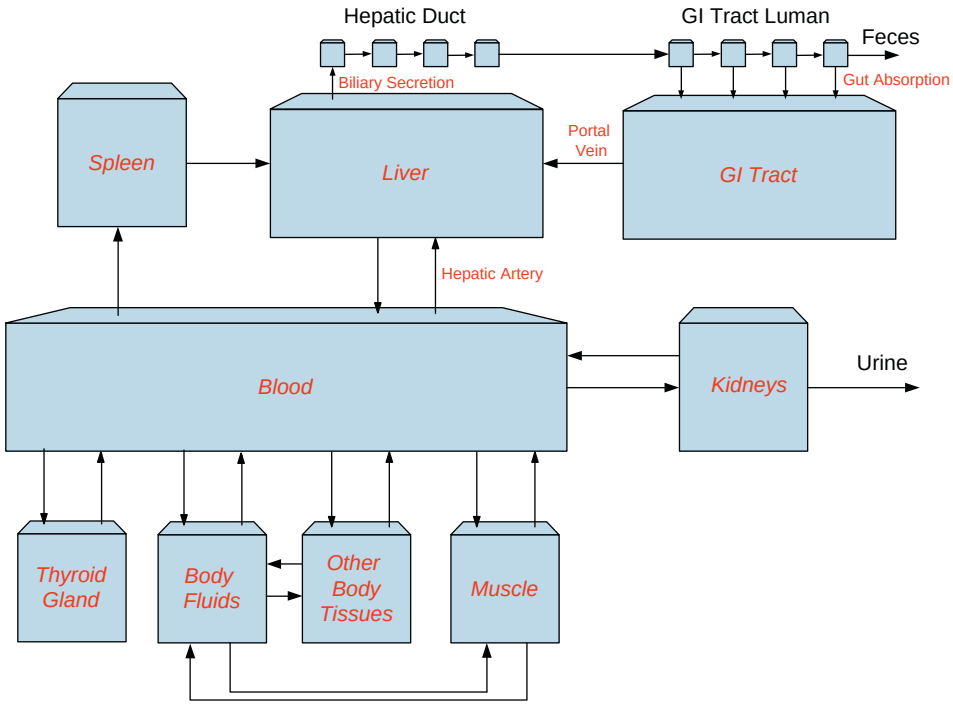


FIGURE 7.10 A pharmacokinetic model describing the movement of a solute in the body.

are considerably smaller than those associated with plasma, body fluids, and tissues, then the removal sites can be eliminated.

By eliminating the removal sites and combining the other compartments in Figure 7.10, we can reduce the model to a one-compartment model, as shown in Figure 7.8, with the drug absorption site as the input to the system. The removal rates can be combined into a single removal rate. While simplistic, this model allows one to judge the therapeutic and toxicological effects of a drug, persistence of the drug in the body, and under- and overdosing of the drug.

To represent the output from the absorption site, an exponential function is used as the input of the one-compartment model, $f(t) = q_2(0)e^{-\gamma t}u(t)$. For $t \geq 0$, the model is described as

$$\dot{q}_1 = q_2(0)e^{-\gamma t} - K_{10}q_1 \tag{7.33}$$

The natural response is the same as before, $q_{1_n} = B_1 e^{-K_{10}t}$. The forced response takes the form of the input, with $q_{1_f} = B_2 e^{-\gamma t}$. Substituting q_{1_f} and $\dot{q}_{1_f} = -\gamma B_2 e^{-\gamma t}$ into Eq. (7.33) gives $B_2 = \dfrac{q_2(0)}{K_{10} - \gamma}$, and $q_{1_f} = \dfrac{q_2(0)}{K_{10} - \gamma} e^{-\gamma t}$. The complete solution is

$$q_1 = q_{1_n} + q_{1_f} = \left(B_1 e^{-K_{10}t} + \frac{q_2(0)}{K_{10} - \gamma} e^{-\gamma t}\right) u(t),$$

where B_1 is determined from the initial condition of the system. With an initial condition of zero, then

$$B_1 = -\frac{q_2(0)}{K_{10} - \gamma},$$

and

$$q_1 = \frac{q_2(0)}{K_{10} - \gamma}\left(e^{-\gamma t} - e^{-K_{10}t}\right) u(t) \tag{7.34}$$

The time course for an exponentially drug administered is illustrated in Figure 7.11.

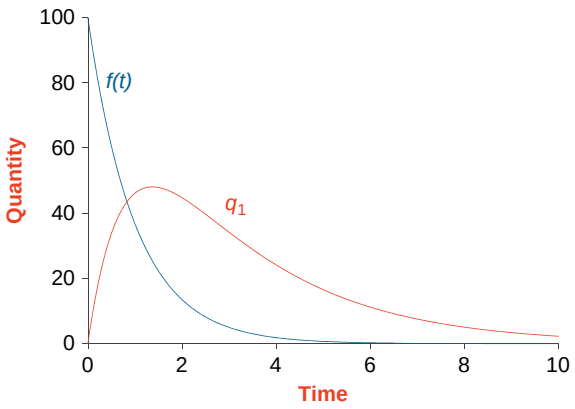


FIGURE 7.11 The time course for an exponentially administered drug. Values used: $q_2(0) = 100$, $K_{10} = 0.4$, and $\gamma = 1.2$.

To determine the time when the maximum drug is in the compartment, Eq. (7.34) is differentiated with respect to t, set equal to zero, and solved as follows:

$$\begin{aligned}\dot{q}_1 &= \frac{d}{dt}\left(\frac{q_2(0)}{K_{10}-\gamma}\left(e^{-\gamma t}-e^{-K_{10}t}\right)\right)\\ &= \frac{q_2(0)}{K_{10}-\gamma}\left(-\gamma e^{-\gamma t}+K_{10}e^{-K_{10}t}\right)\end{aligned} \tag{7.35}$$

Setting Eq. (7.35) equal to zero and $t = t_{\max}$ gives

$$\frac{q_2(0)}{K_{10}-\gamma}\left(-\gamma e^{-\gamma t_{\max}}+K_{10}e^{-K_{10}t_{\max}}\right)=0$$

or

$$\gamma e^{-\gamma t_{\max}} = K_{10}e^{-K_{10}t_{\max}} \tag{7.36}$$

Multiplying both sides of Eq. (7.36) by $e^{K_{10}t_{\max}}$ and dividing by γ yields

$$e^{K_{10}t_{\max}}e^{-\gamma t_{\max}} = e^{(K_{10}-\gamma)t_{\max}} = \frac{K_{10}}{\gamma} \tag{7.37}$$

Taking the logarithm of both sides of Eq. (7.37), we have

$$(K_{10}-\gamma)t_{\max} = \ln\left(\frac{K_{10}}{\gamma}\right)$$

Solving for $t_{\max}$ yields

$$t_{\max} = \frac{\ln\left(\frac{K_{10}}{\gamma}\right)}{K_{10}-\gamma} \tag{7.38}$$

The maximum quantity of drug in the compartment is

$$q_1(t_{\max}) = \frac{q_2(0)}{K_{10}-\gamma}\left(e^{-\gamma t_{\max}}-e^{-K_{10}t_{\max}}\right)\Bigg|_{t_{\max}=\frac{\ln\left(\frac{K_{10}}{\gamma}\right)}{K_{10}-\gamma}}$$

$$= \frac{q_2(0)}{K_{10}-\gamma}\left(e^{-\frac{\gamma}{K_{10}-\gamma}\ln\left(\frac{K_{10}}{\gamma}\right)}-e^{-\frac{K_{10}}{K_{10}-\gamma}\ln\left(\frac{K_{10}}{\gamma}\right)}\right)$$

$$= \frac{q_2(0)}{K_{10}-\gamma}\left(e^{\ln\left(\left(\frac{K_{10}}{\gamma}\right)^{-\frac{\gamma}{K_{10}-\gamma}}\right)} - e^{\ln\left(\left(\frac{K_{10}}{\gamma}\right)^{-\frac{K_{10}}{K_{10}-\gamma}}\right)} \right) = \frac{q_2(0)}{K_{10}-\gamma}\left(\left(\frac{K_{10}}{\gamma}\right)^{-\frac{\gamma}{K_{10}-\gamma}} - \left(\frac{K_{10}}{\gamma}\right)^{-\frac{K_{10}}{K_{10}-\gamma}} \right)$$

$$= \frac{q_2(0)}{K_{10}-\gamma}\left(\left(\frac{\gamma}{K_{10}}\right)^{\frac{\gamma}{K_{10}-\gamma}} - \left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}} \right) = \frac{q_2(0)}{K_{10}-\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}}\left(\left(\frac{\gamma}{K_{10}}\right)^{\frac{\gamma-K_{10}}{K_{10}-\gamma}} - 1 \right)$$

$$= \frac{q_2(0)}{K_{10}-\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}}\left(\left(\frac{K_{10}}{\gamma}\right) - 1 \right) = \frac{q_2(0)}{K_{10}-\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}}\left(\frac{K_{10}-\gamma}{\gamma}\right)$$

$$= \frac{q_2(0)}{\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}} \tag{7.39}$$

The maximum concentration of the drug is

$$c_1(t_{\max}) = \frac{q_2(0)}{V_1\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}} \tag{7.40}$$

With an exponential input, the maximum drug concentration occurs at time $t_{\max} = \ln\frac{\left(\frac{K_{10}}{\gamma}\right)}{K_{10}-\gamma}$ with value $\frac{q_2(0)}{V_1\gamma}\left(\frac{\gamma}{K_{10}}\right)^{\frac{K_{10}}{K_{10}-\gamma}}$ according to Eqs. (7.38) and (7.40). When giving a bolus injection, the maximum drug concentration occurs at time 0 with value $\frac{q_2(0)}{V_1}$ according to Example Problem 7.5. Thus, a bolus injection achieves a higher concentration of the drug in the plasma and is faster than an exponential input.

When administering a drug in the body, the persistence of the drug is an important parameter to judge the clinical effectiveness and to determine a dose administration schedule. To produce a therapeutic effect, a minimum drug concentration in the plasma is required so the drug can diffuse to its target site. Maximum drug concentration is an important parameter, since a dosage at too high a level can be toxic. The next section describes how to maintain a minimum and maximum dosage.

7.5.4 Repeat Dosages

To maintain the concentration of a drug in the body, the drug must be administered on a regular basis to keep the concentration above a minimum value. For simplicity, assume the model in Figure 7.8 and Eq. (7.29), and that the input is a series of boluses, with magnitude

ζ, and the time between dosages is T. With the first dose administered at $t = 0$ and zero initial conditions, we have

$$q_1(t) = \zeta e^{-K_{10}t}$$

for $0 < t < T$. The maximum quantity in the compartment is at $t = 0$, and the minimum is at $t = T^-$. At $t = T$, the second dose is given, and according to superposition, the quantity is

$$q_1(t) = \zeta e^{-K_{10}t} + \zeta e^{-K_{10}(t-T)}$$

for $T \leq t < 2T$, with a maximum of $\zeta e^{-K_{10}T} + \zeta$ at $t = T$, and a minimum of $\zeta e^{-K_{10}2T^-} + \zeta e^{-K_{10}T^-} = \zeta e^{-K_{10}T^-}\left(1 + \zeta e^{-K_{10}T^-}\right)$ at $t = 2T^-$.

In general, for any interval $(n-1)T \leq t < nT$, we have

$$q_1(t) = \zeta e^{-K_{10}t} + \zeta e^{-K_{10}(t-T)} + \cdots + \zeta e^{-K_{10}(t-(n-2)T)} + \zeta e^{-K_{10}(t-(n-1)T)}$$

with the maximum in the interval at $t = (n-1)T$ of

$$q_{1_{\max}}((n-1)T) = \zeta e^{-K_{10}(n-1)T} + \zeta e^{-K_{10}(n-2)T} + \cdots + \zeta e^{-K_{10}T} + \zeta \tag{7.41}$$

Equation (7.41) is written in closed form as

$$q_{1_{\max}}((n-1)T) = \zeta\left(\frac{1 - e^{-K_{10}nT}}{1 - e^{-K_{10}T}}\right) \tag{7.42}$$

To find the minimum quantity at $t = nT^-$,

$$q_{1_{\min}}(nT^-) = \zeta e^{-K_{10}nT^-} + \zeta e^{-K_{10}(n-1)T^-} + \cdots + \zeta e^{-K_{10}2T^-} + \zeta e^{-K_{10}T^-} \tag{7.43}$$

Next, we write a closed form expression for Eq. (7.43), giving

$$q_{1_{\min}}(nT^-) = \zeta e^{-K_{10}T^-}\left(\frac{1 - e^{-K_{10}nT^-}}{1 - e^{-K_{10}T^-}}\right) \tag{7.44}$$

As $n \to \infty$, Eq. (7.42) approaches

$$q_{1_{\max}} = \frac{\zeta}{1 - e^{-K_{10}T}} \tag{7.45}$$

and Eq. (7.43) approaches

$$q_{1_{\min}} = \frac{\zeta e^{-K_{10}T}}{1 - e^{-K_{10}T}} \tag{7.46}$$

With the parameter values shown in Figure 7.12, the maximum approaches

$$q_{1_{\max}} = \frac{\zeta}{1 - e^{-K_{10}T}} = 3.86$$

and the minimum approaches

$$q_{1_{\min}} = \frac{\zeta e^{-K_{10}T}}{1 - e^{-K_{10}T}} = 2.86$$

To convert the quantities into concentrations, simply divide them by the volume of the compartment.

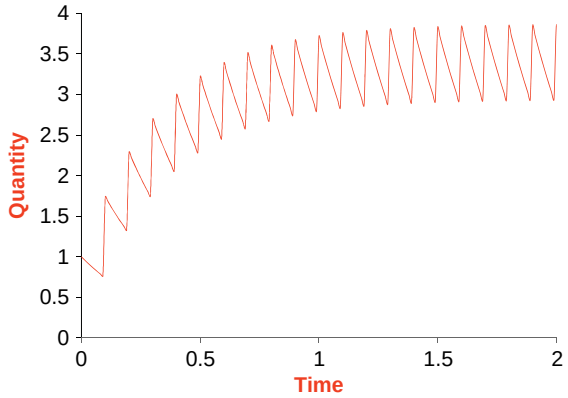


FIGURE 7.12 Illustration of change in quantity due to repeated bolus injections using a one-compartment model. Values used: $\zeta = 1$, $K_{10} = 3$, and $T = 0.1$.

EXAMPLE PROBLEM 7.6

A 3 g bolus of antibiotic is administered to a human with a plasma volume of 3 L. The average impulse response for this drug is shown in Figure 7.13. Assuming a one-compartment model, determine the transfer rate. If the concentration of the drug is not to fall below 30 percent of the initial dosage at steady state, how often does the drug need to be given to maintain this minimum level?

Solution

To determine the transfer rate, K_{10}, the curve in Figure 7.13 is used to determine the half-life, which is approximately $t_{\frac{1}{2}} = 1.4$. Using Eq. (7.4.2), $t_{\frac{1}{2}} = 1.4 = \dfrac{\ln(2)}{K_{10}}$, K_{10} is found as

$$K_{10} = \frac{\ln(2)}{1.4} \simeq 0.5$$

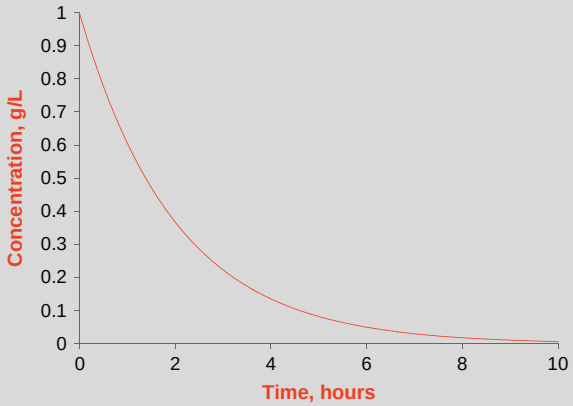


FIGURE 7.13 Illustration for Example Problem 7.6 with 3 g bolus in a 3 L plasma compartment.

In the first interval, $c_1 = \dfrac{q_1(0)e^{-\frac{t}{2}}}{V_1} = e^{-\frac{t}{2}}$, and $c_{1_{\min}}$ equals 30 percent of the initial value, or $c_{1_{\min}} = 0.3$. At steady state, the minimum concentration is determined from Eq. (7.46) as

$$c_{1_{\min}} = 0.3 = \frac{\zeta e^{-K_{10}T}}{V_1(1 - e^{-K_{10}T})} = \frac{e^{-\frac{T}{2}}}{\left(1 - e^{-\frac{T}{2}}\right)}$$

Solving the previous equation for T, we have $T \simeq 3$ hours. Thus, the antibiotic needs to be administered every 3 hours, as shown in Figure 7.14.

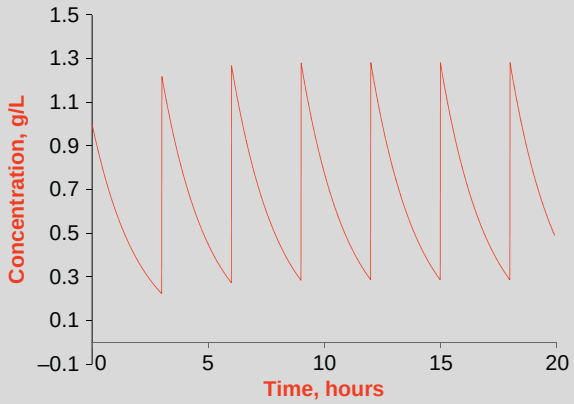


FIGURE 7.14 Change in concentration with a repeat dose of the antibiotic every 3 hours.

7.6 TWO-COMPARTMENT MODELING

The general form of the two-compartment model is shown in Figure 7.15. Here we assume that all transfer rates are constants and $K_{ij} \geq 0$.

We begin with the general form of the two-compartment model and then examine special cases with some of the transfer rates or inputs set equal to zero and other considerations. To analyze the system in Figure 7.15, conservation of mass is used to write a differential equation for each compartment describing the rate of change of the quantity of solute in the compartment, given as accumulation = input – output, where

Compartment 1	Compartment 2
Accumulation $= \dot{q}_1$	Accumulation $= \dot{q}_2$
Input $= f_1(t) + K_{21}q_2$	Input $= f_2(t) + K_{12}q_1$
Ouput $= (K_{10} + K_{12})q_1$	Ouput $= (K_{20} + K_{21})q_2$

Therefore,

$$\begin{aligned} \dot{q}_1 &= f_1(t) + K_{21}q_2 - (K_{10} + K_{12})q_1 \\ \dot{q}_2 &= f_2(t) + K_{12}q_1 - (K_{20} + K_{21})q_2 \end{aligned} \tag{7.47}$$

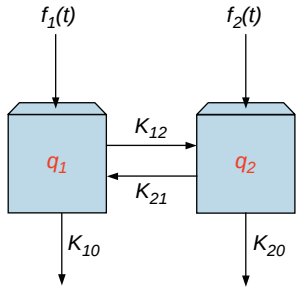


FIGURE 7.15 A general two-compartment model. Compartment 1 has volume V_1 and compartment 2 has volume V_2.

The easiest way to analyze the system in Eq. (7.47) is using the D-Operator or its equivalent, the Laplace variable s, where we represent the system in matrix form as

$$D\mathbf{IQ} = \mathbf{AQ} + \mathbf{F} \tag{7.48}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -(K_{10} + K_{12}) & K_{21} \\ K_{12} & -(K_{20} + K_{21}) \end{bmatrix}, \mathbf{F} = \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}$$

Equation (7.48) is solved as

$$\mathbf{Q} = (D\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} = \frac{1}{\det(D\mathbf{I} - \mathbf{A})}\text{adj}(D\mathbf{I} - \mathbf{A})\mathbf{F}$$

or

$$\det(D\mathbf{I} - \mathbf{A})\mathbf{Q} = \text{adj}(D\mathbf{I} - \mathbf{A})\mathbf{F} \tag{7.49}$$

and using MATLAB, we have

```
>> syms D q1 q2 K10 K20 K12 K21
>> A=[-(K10+K12) K21;K12 -(K20+K21)]
>> det(D*eye(2)-A)
ans =
  D^2+D*K20+D*K21+K10*D+K10*K20+K10*K21+K12*D+K12*K20
>> adj=det(D*eye(2)-A)*inv(D*eye(2)-A)

  adj =
  [ D+K20+K21,    K21]
  [    K12, D+K10+K12]
```

Substituting these values from MATLAB into Eq. (7.49) gives

$$\begin{aligned}\left(D^2 + D(K_{10} + K_{12} + K_{20} + K_{21}) + (K_{10}K_{20} + K_{10}K_{21} + K_{12}K_{20})\right)\mathbf{Q} \\ = \begin{bmatrix} D + (K_{20} + K_{21}) & K_{21} \\ K_{12} & D + (K_{10} + K_{12}) \end{bmatrix}\mathbf{F}\end{aligned} \tag{7.50}$$

Returning to the time domain, we have the following independent differential equations:

$$\begin{aligned}\ddot{q}_1 + (K_{10} + K_{12} + K_{20} + K_{21})\dot{q}_1 + (K_{10}K_{20} + K_{10}K_{21} + K_{12}K_{20})q_1 \\ = \frac{df_1(t)}{dt} + (K_{20} + K_{21})f_1(t) + K_{21}f_2(t) \\ \ddot{q}_2 + (K_{10} + K_{12} + K_{20} + K_{21})\dot{q}_2 + (K_{10}K_{20} + K_{10}K_{21} + K_{12}K_{20})q_2 \\ = K_{12}\, f_1(t) + \frac{df_2(t)}{dt} + (K_{10} + K_{12})f_2(t)\end{aligned} \tag{7.51}$$

Note that the characteristic equation, det(D**I**−**A**), is identical for both q_1 and q_2, and the form of the natural response is the same for either variable. Also note that the coefficients in the natural response are not identical for q_1 and q_2, and depend on the input to the compartment and the initial conditions.

The roots of the characteristic equation are determined using MATLAB as

```
>> eig(A)
ans =
-1/2*K10-1/2*K12-1/2*K20-1/2*K21+1/2*(K10^2+2*K10*K12-
2*K10*K20-2*K10*K21+K12^2-2*K12*K20+2*K21*K12+K20^2+
2*K20*K21+K21^2)^(1/2)

-1/2*K10-1/2*K12-1/2*K20-1/2*K21-1/2*(K10^2+2*K10*K12-
2*K10*K20-2*K10*K21+K12^2-2*K12*K20+2*K21*K12+K20^2+
2*K20*K21+K21^2)^(1/2)
```

This expression simplifies to

$$s_{1,2} = -\frac{(K_{10} + K_{12} + K_{20} + K_{21})}{2} \pm \frac{1}{2}\sqrt{(K_{20} + K_{21} - K_{10} - K_{12})^2 + 4K_{21}K_{12}} \tag{7.52}$$

From Eq. (7.52), we note that there can be no positive real roots and no imaginary roots if all $K_{ij} \geq 0$. If $(K_{20} + K_{21} - K_{10} - K_{12})^2 + 4K_{21}K_{12} = 0$, then the roots are repeated and equal to $s_{1,2} = -\dfrac{(K_{10} + K_{12} + K_{20} + K_{21})}{2}$. For repeated roots to happen, $(K_{20} + K_{21})$ must equal $(K_{10} + K_{12})$, and either K_{21} or K_{12} must be zero. If K_{21} and K_{12} are both equal to zero, then there is no movement of solute between the compartments.

In the following example, we revisit Fick's Law of diffusion using compartmental analysis and compute the concentration. The difference in analysis involves the transfer rates as

functions of the volume of each compartment—that is $K_{12} = \frac{K}{V_1}$ and $K_{21} = \frac{K}{V_2}$, where $K = \frac{DA}{\Delta x}$. The system in Example 7.7 is called a *closed compartment* because there is no output to the environment in a closed system.

EXAMPLE PROBLEM 7.7

Consider the two-compartment model in Figure 7.16, with $q_1(0) = \zeta$ and $q_2(0) = 0$. Solve for the concentration in each compartment.

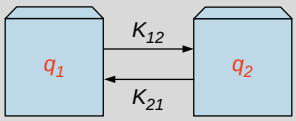


FIGURE 7.16 Illustration for Example Problem 7.7.

Solution

Conservation of mass for each compartment is

$$\dot{q}_1 = K_{21}q_2 - K_{12}q_1$$
$$\dot{q}_2 = K_{12}q_1 - K_{21}q_2$$

Using the D-Operator method gives

$$\ddot{q}_1 + (K_{12} + K_{21})\dot{q}_1 = 0$$
$$\ddot{q}_1 + (K_{12} + K_{21})\dot{q}_2 = 0$$

The roots are $s_{1,2} = 0, -(K_{12} + K_{21})$, which gives

$$q_1(t) = B_1 + B_2 e^{-(K_{12}+K_{21})t}$$
$$q_2(t) = B_3 + B_4 e^{-(K_{12}+K_{21})t}$$

We use the initial conditions to solve for B_i as follows

$$q_1(0) = \zeta = B_1 + B_2 e^{-(K_{12}+K_{21})t}|_{t=0} = B_1 + B_2$$

To find $\dot{q}_1(0)$, we use the conservation of mass equation for $\dot{q}_1$ at time zero

$$\dot{q}_1(0) = K_{21}q_2(0) - K_{12}q_1(0) = -K_{12}\zeta$$

and from the solution,

$$\dot{q}_1 = \frac{d\left(B_1 + B_2 e^{-(K_{12}+K_{21})t}\right)}{dt} = -(K_{12} + K_{21})B_2 e^{-(K_{12}+K_{21})t}$$

$$\dot{q}_1(0) = -K_{12}\zeta = -(K_{12} + K_{21})B_2$$

To solve for B_1 and B_2, we evaluate

$$\begin{bmatrix} 1 & 1 \\ 0 & -(K_{12}+K_{21}) \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \zeta \\ -K_{12}\zeta \end{bmatrix}$$

which gives

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\zeta K_{21}}{(K_{12}+K_{21})} \\ \dfrac{\zeta K_{12}}{(K_{12}+K_{21})} \end{bmatrix}$$

and

$$q_1(t) = \frac{\zeta}{(K_{12}+K_{21})}\left(K_{21} + K_{12}e^{-(K_{12}+K_{21})t}\right)u(t)$$

The concentration is

$$c_1(t) = \frac{\zeta}{V_1(K_{12}+K_{21})}\left(K_{21} + K_{12}e^{-(K_{12}+K_{21})t}\right)u(t)$$

Repeating the same steps for q_2 as before gives

$$q_2(t) = \frac{K_{12}\zeta}{(K_{12}+K_{21})}\left(1 - e^{-(K_{12}+K_{21})t}\right)u(t)$$

and

$$c_2(t) = \frac{K_{12}\zeta}{V_2(K_{12}+K_{21})}\left(1 - e^{-(K_{12}+K_{21})t}\right)u(t)$$

or using $q_2 = \zeta - q_1$ gives the same result.

A more straightforward solution involves substituting $q_2 = \zeta - q_1$ into $\dot{q}_1 = K_{21}q_2 - K_{12}q_1$, and solving $\dot{q}_1 = K_{21}(\zeta - q_1) - K_{12}q_1 = K_{21}\zeta - (K_{21}+K_{12})q_1$.

If $K_{12} = \dfrac{K}{V_1}$ and $K_{21} = \dfrac{K}{V_2}$, and $V_1 = V_2 = V$, then these results are the same as those computed using Fick's Law of diffusion in Section 7.2.2:

$$c_1 = \frac{\zeta}{2}\left(e^{\frac{-Kt}{V}} + 1\right)u(t)$$

and

$$c_2 = \frac{\zeta}{2}\left(1 - e^{\frac{-Kt}{V}}\right)u(t)$$

In a two-compartment model, the half-life is defined using two terms based on the roots of the characteristic equation. The half-life associated with the smaller root is called the elimination half-life, and the distribution half-life is used for the larger root.

7.6.1 Source Compartment

A compartment that only outputs to other compartments, without any inputs from other compartments, is called a source compartment. A source compartment has an input $f(t)$. This type of compartment is simply a one-compartment model that can be solved independently of the other compartments in the system. The output of the source compartment is an exponential decay, as described in Section 7.5. While in many situations, the source compartment does not send the solute to the environment, it is perfectly fine for a source compartment to do so.

Using the model shown in Figure 7.15, a source compartment exists if either K_{12} or K_{21} is zero. For repeated roots in a two-compartment model, a source compartment must be one of the compartments and $(K_{20} + K_{21})$ must equal $(K_{10} + K_{12})$.

In Example Problem 7.8, the digestive system is introduced as a source compartment. By including a digestive system component, the solute is not instantaneously delivered into the plasma but is slowly released from the digestive system into the plasma through a bolus input.

EXAMPLE PROBLEM 7.8

Consider the two-compartment model shown in Figure 7.17 with the ingestion of a bolus solute in the digestive system and removal of the solute via metabolism and excretion in urine. Solve for the plasma concentration.

Solution

This model has a source compartment. Rather than solving the problem with a bolus input, the initial condition is changed to $q_2(0)$ with no input. The conservation of mass for each compartment is

$$\dot{q}_1 = K_{21}q_2 - (K_{1M} + K_{1U})q_1 \tag{7.53}$$

$$\dot{q}_2 = -K_{21}q_2 \tag{7.54}$$

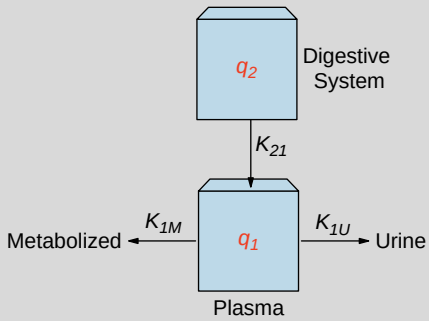


FIGURE 7.17 A two-compartment model with realistic ingestion of solute and removal from the plasma by metabolism and excretion in urine. It normally takes about 30 minutes to pass through the digestive system.

Since Eq. (7.54) involves only q_2, it is easily solved as $q_2 = q_2(0)e^{-K_{21}t}u(t)$. By substituting the solution for q_2 into Eq. (7.53), we now have one equation, giving

$$\dot{q}_1 = q_2(0)K_{21}e^{-K_{21}t} - (K_{1M} + K_{1U})q_1$$

and after rearranging

$$\dot{q}_1 + (K_{1M} + K_{1U})q_1 = q_2(0)K_{21}e^{-K_{21}t} \tag{7.55}$$

This is a first-order differential equation with a forcing function $q_2(0)K_{21}e^{-K_{21}t}$. The natural solution is $q_{1_n} = B_1e^{-(K_{1M}+K_{1U})t}$ and the forced response is $q_{1_f} = B_2e^{-K_{21}t}$. To determine B_2, $q_{1_f} = B_2e^{-K_{21}t}$ is substituted into Eq. (7.55), which gives

$$-K_{21}B_2e^{-K_{21}t} + (K_{1M} + K_{1U})B_2e^{-K_{21}t} = q_2(0)K_{21}e^{-K_{21}t}$$

Solving for B_2 gives

$$B_2 = \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}$$

The complete response is

$$\begin{aligned} q_1 &= q_{1_n} + q_{1_f} = B_1e^{-(K_{1M}+K_{1U})t} + B_2e^{-K_{21}t} \\ &= B_1e^{-(K_{1M}+K_{1U})t} + \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}e^{-K_{21}t} \end{aligned} \tag{7.56}$$

B_1 is found using the initial condition $q_1(0) = 0$

$$q_1(0) = 0 = \left[B_1e^{-(K_{1M}+K_{1U})t} + \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}e^{-K_{21}t}\right]_{t=0} = B_1 + \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}$$

giving

$$B_1 = -\frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}$$

and

$$q_1 = \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}\left(e^{-K_{21}t} - e^{-(K_{1M}+K_{1U})t}\right)u(t) \tag{7.57}$$

or in terms of concentration,

$$c_1 = \frac{1}{V_1}\frac{q_2(0)K_{21}}{(K_{1M} + K_{1U} - K_{21})}\left(e^{-K_{21}t} - e^{-(K_{1M}+K_{1U})t}\right)u(t) \tag{7.58}$$

To determine the time when the maximum solute is in compartment 1 in Example Problem 7.8, Eq. (7.57) is differentiated with respect to t, set equal to zero, and solved as follows:

$$\begin{aligned} \dot{q}_1 &= \frac{d}{dt}\left(\frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}\left(e^{-K_{21}t} - e^{-(K_{1M}+K_{1U})t}\right)\right) \\ &= \frac{q_2(0)K_{21}}{K_{1M} + K_{1U} - K_{21}}\left(-K_{21}e^{-K_{21}t} + (K_{1M} + K_{1U})e^{-(K_{1M}+K_{1U})t}\right) \end{aligned} \tag{7.59}$$

Setting Eq. (7.59) equal to zero and $t = t_{\max}$ gives

$$\frac{q_2(0)K_3}{K_{1M}+K_{1U}-K_{21}}\left(-K_{21}e^{-K_{21}t_{\max}}+(K_{1M}+K_{1U})e^{-(K_{1M}+K_{1U})t_{\max}}\right)=0$$

or

$$K_{21}e^{-K_{21}t_{\max}}=(K_{1M}+K_{1U})e^{-(K_{1M}+K_{1U})t_{\max}}$$

Multiplying both sides of the previous equation by $e^{(K_{1M}+K_{1U})t_{\max}}$ and dividing by K_{21} gives

$$e^{(K_{1M}+K_{1U})t}e^{-K_{21}t_{\max}}=e^{(K_{1M}+K_{1U}-K_{21})t}=\frac{K_{1M}+K_{1U}}{K_{21}}$$

Taking the logarithm of both sides gives

$$(K_{1M}+K_{1U}-K_{21})t_{\max}=\ \ln\left(\frac{K_{1M}+K_{1U}}{K_{21}}\right)$$

Solving for $t_{\max}$ yields

$$t_{\max}=\frac{\ln\left(\frac{K_{1M}+K_{1U}}{K_{21}}\right)}{(K_{1M}+K_{1U}-K_{21})} \tag{7.60}$$

It should be clear from Eq. (7.59) that the smaller the term $K_{1M}+K_{1U}$ compared to K_{21}, the more time it takes to reach the maximum concentration or quantity in the plasma.

EXAMPLE PROBLEM 7.9

Suppose 50 g of solute is ingested. Find the maximum amount of solute in the plasma if the compartmental model in Figure 7.17 is used with $K_{1M}+K_{1U}=0.005\ \text{min}^{-1}$ and $K_{21}=0.02\ \text{min}^{-1}$.

Solution

Using Eq. (7.60) gives

$$t_{\max}=\frac{\ln\left(\frac{K_{1M}+K_{1U}}{K_{21}}\right)}{(K_{1M}+K_{1U}-K_{21})}=\frac{\ln\left(\frac{0.005}{0.02}\right)}{0.005-0.02}=\frac{\ln(0.25)}{-0.015}=92.42\ \text{min}$$

The maximum amount of solute in compartment 1 at $t_{\max}$ is therefore

$$\begin{aligned}q_1(t_{\max})&=\frac{q_2(0)K_{21}}{K_{1M}+K_{1U}-K_{21}}\left(e^{-K_{21}t}-e^{-(K_{1M}+K_{1U})t}\right)\Big|_{t=92.42}\\&=\frac{50\times0.02}{0.005-0.02}\left(e^{-0.02\times92.42}-e^{-0.005\times92.42}\right)=31.5\ \text{g}\end{aligned}$$

The next example introduces an encapsulated pill input that releases a portion immediately and the remainder continuously until the pill completely dissolves. Mathematically this input is approximated as $\zeta\delta(t)+(1-\zeta)(u(t)-u(t-t_0))$. To estimate the gastric transfer

rate and the fraction released immediately experimentally, the pill is dissolved in a solution similar to the stomach, and the concentration is measured. From this data, parameter values can be determined.

EXAMPLE PROBLEM 7.10

Consider the two-compartment model shown in Figure 7.15 with $K_{12} = K_{20} = 0$, $K_{21} = K_{10} = 0.2$, $f_1(t) = 0$, and $f_2(t) = 20\delta(t) + 80(u(t) - u(t-30))$. Assume that the initial conditions are zero (not including that provided by $20\delta(t)$). Solve for the quantity in each compartment.

Solution

Since $K_{12} = 0$, this model has a source compartment. The solution is carried out using superposition, separating the input $f_2(t)$ into $20\delta(t)$, $80u(t)$, and $80u(t-30)$, and then summing the individual responses to get the complete response.

The conservation of mass for each compartment is

$$\dot{q}_1 = 0.2q_2 - 0.2q_1 \tag{7.61}$$

$$\dot{q}_2 = f_2 - 0.2q_2 \tag{7.62}$$

20δ(t) *Input*

First, consider the $20\delta(t)$ input. As in Example Problem 7.8, we treat the impulse input as a change in an initial condition, yielding $q_{2_\delta} = 20e^{-0.2t}u(t)$. Substituting this result into Eq. (7.61) gives

$$\dot{q}_{1_\delta} = 4e^{-0.2t} - 0.2q_{1_\delta} \tag{7.63}$$

The root for Eq. (7.63) is $s = -0.2$, and has a natural solution $q_{1_n} = B_1e^{-0.2t}$. The input in Eq. (7.63) has the same form as the natural solution (expected since $(K_{20} + K_{21})$ equals $(K_{10} + K_{12})$ and $K_{12} = 0$), and so the forced response is $q_{1_f} = B_2te^{-0.2t}$. Substituting q_{1_f} into Eq. (7.63) gives $B_2 = 4$. The complete response is

$$q_{1_\delta} = q_{1_n} + q_{1_f} = B_1e^{-0.2t} + 4te^{-0.2t} \tag{7.64}$$

B_1 is found using the initial condition $q_1(0) = 0$ and

$$q_{1_\delta}(0) = 0 = \left[B_1e^{-0.2t} + 4te^{-0.2t}\right]_{t=0} = B_1$$

Thus,

$$q_{1_\delta} = 4te^{-0.2t}u(t) \tag{7.65}$$

80u(t) *Input*

Next, consider the $80u(t)$ input. The conservation of mass equations are

$$\dot{q}_{1_u} = 0.2q_{2_u} - 0.2q_{1_u} \tag{7.66}$$

$$\dot{q}_{2_u} = 80 - 0.2q_{2_u} \tag{7.67}$$

Continued

Solving Eq. (7.67) gives $q_{2_u} = 400\left(1 - e^{-0.2t}\right)$. Substituting q_{2_u} into Eq. (7.66) gives

$$\dot{q}_{1_u} = 80\left(1 - e^{-0.2t}\right) - 0.2q_{1_u} \tag{7.68}$$

The root for Eq. (7.68) is $s = -0.2$, and has a natural solution $q_{1_n} = B_1 e^{-0.2t}$. The input in Eq. (7.66) has the same term as in the natural solution, and so the forced response is $q_{1_f} = B_3 + B_2 t e^{-0.2t}$. Substituting q_{1_f} into Eq. (7.68) gives $B_2 = -80$ and $B_3 = 400$. The complete response is

$$q_{1_u} = q_{1_n} + q_{1_f} = B_1 e^{-0.2t} + 400 - 80te^{-0.2t} \tag{7.69}$$

B_1 is found using the initial condition $q_1(0) = 0$ and

$$q_{1_u}(0) = 0 = \left[B_1 e^{-0.2t} + 400 - 80te^{-0.2t}\right]_{t=0} = B_1 + 400$$

and $B_1 = -400$. Thus,

$$q_{1_u} = \left(400 - 400e^{-0.2t} - 80te^{-0.2t}\right)u(t) \tag{7.70}$$

$-80u(t-30)$ *Input*

Next, consider the $-80u(t - 30)$ input. By the property of a linear system, then

$$q_{1_{u-30}} = -\left(400 - 400e^{-0.2(t-30)} - 80(t-30)e^{-0.2(t-30)}\right)u(t-30)$$

$$q_{2_{u-30}} = -400\left(1 - e^{-0.2(t-30)}\right)u(t-30)$$

Complete Solution

The complete response is

$$q_1 = q_{1_\delta} + q_{1_u} + q_{1_{u-30}} = 4te^{-0.2t}u(t) + \left(400 - 400e^{-0.2t} - 80te^{-0.2t}\right)u(t) - \left(400 - 400e^{-0.2(t-30)} - 80(t-30)e^{-0.2(t-30)}\right)u(t-30)$$

$$q_2 = q_{2_\delta} + q_{2_u} + q_{2_{u-30}} = 20e^{-0.2t}u(t) + 400\left(1 - e^{-0.2t}\right)u(t) - 400\left(1 - e^{-0.2(t-30)}\right)u(t-30)$$

which is plotted in Figure 7.18.

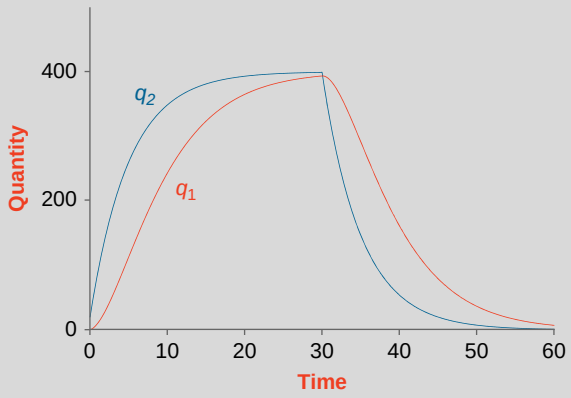


FIGURE 7.18 Plot of the solute quantities for Example Problem 7.10.

When delivering anesthesia, a similar input is used as in Example Problem 7.10—that is, a bolus plus constant infusion. The reason for this type of input is to quickly raise the anesthesia to a desired level (bolus) and then to maintain the level for the operation (step).

7.6.2 Sink Compartment

A sink compartment is one that has only inputs and no output. Similar to the source compartment, a sink acts like an integrator and has a zero root. Moreover, the solution of the nonsink compartment is independent of the sink compartment in the two-compartment case. Once solved, the quantity in the nonsink compartment is used to solve the quantity in the sink compartment. Using the model shown in Figure 7.15, a sink compartment exists if either K_{12} and K_{10}, or K_{21} and K_{20} are zero.

EXAMPLE PROBLEM 7.11

Consider the two-compartment model shown in Figure 7.15 with $K_{12} = K_{10} = 0$, $K_{21} = 0.2$, $K_{20} = 1$, $f_1(t) = 0$, and $f_2(t) = 10\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

Solution

Since $K_{21} = 0.2$ and $K_{12} = 0$, this compartment model has a sink for compartment 1. As before, rather than solving the problem with a bolus input, the initial condition is changed in compartment 2 to $q_2(0) = 10$, with zero input. The conservation of mass for each compartment is

$$\dot{q}_1 = K_{21}q_2 = 0.2q_2 \tag{7.71}$$

$$\dot{q}_2 = -(K_{21} + K_{20})q_2 = -1.2q_2 \tag{7.72}$$

Since Eq. (7.72) involves only q_2, we solve directly to get $q_2 = 10e^{-1.2t}u(t)$. Next, substitute q_2 into (Eq. 7.71), yielding

$$\dot{q}_1 = 0.2q_2 = 2e^{-1.2t} \tag{7.73}$$

Eq. (7.73) gives a single root at $s = 0$, and $q_{1_n} = B_1$. The forced response is $q_{1_f} = B_2e^{-1.2t}$, which when substituted into Eq. (7.71) gives $B_2 = -1.67$. The complete response is

$$q_1 = q_{1_n} + q_{1_f} = B_1 - 1.667e^{-1.2t} \tag{7.74}$$

With $q_1(0) = 0$, we have $B_1 = 1.667$ from Eq. (7.74), and the complete response is

$$q_1 = 1.667\left(1 - e^{-1.2t}\right)u(t) \tag{7.75}$$

This result indicates that more than 80 percent of the solute has moved from the system into the environment. If $K_{20} = 0$, then $q_2 = 10e^{-0.2t}u(t)$ and $q_1 = 10\left(1 - e^{-0.2t}\right)u(t)$. Here, all the solute exponentially moves from compartment 2 to 1 as expected.

EXAMPLE PROBLEM 7.12

Consider a two-compartment system for the distribution of creatinine in the body illustrated in Figure 7.19. Compartment 1 represents the plasma and compartment 2 the muscle. Creatinine is a waste product of metabolism in the muscle that's cleared from the body through the urine (transfer rate K_{10}). Assume creatinine production in the muscle is $f_2(t)$ and is given by a step input. Find the concentration of creatinine in the plasma compartment.

Solution

The differential equations describing the rate of change of creatinine in the compartments 1 and 2 are written by using the conservation of mass equation as

$$\dot{q}_1 = K_{21}q_2 - (K_{10} + K_{12})q_1 \tag{7.75}$$

$$\dot{q}_2 = K_{12}q_1 - K_{21}q_2 + f_2 = K_{12}q_1 - K_{21}q_2 + 1 \tag{7.76}$$

The D-Operator is used to remove q_2, giving

$$\ddot{q}_1 + (K_{10} + K_{12} + K_{21})\dot{q}_1 + K_{10}K_{21}q_1 = K_{21} \tag{7.77}$$

The roots of the characteristic equation are

$$s_{1,2} = -\frac{(K_{10} + K_{12} + K_{21})}{2} \pm \frac{1}{2}\sqrt{(K_{21} - K_{10} - K_{12})^2 - 4K_{21}K_{10}}$$

The natural response is an overdamped response:

$$q_{1_n} = B_1 e^{s_1 t} + B_2 e^{s_2 t}$$

The forced response is a constant (B_3) and when substituted into the differential equation yields $B_3 = \dfrac{1}{K_{10}}$. The complete response is

$$q_1 = B_1 e^{s_1 t} + B_2 e^{s_2 t} + \frac{1}{K_{10}}$$

and

$$c_1 = \frac{1}{V_1}\left(B_1 e^{s_1 t} + B_2 e^{s_2 t} + \frac{1}{K_{10}}\right)$$

for $t \geq 0$. The constants B_1 and B_2 are determined using the initial conditions.

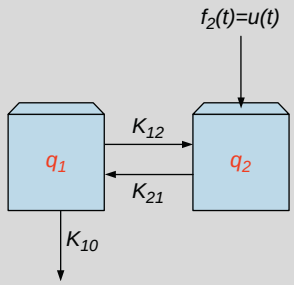


FIGURE 7.19 Illustration for Example Problem 7.12.

7.7 THREE-COMPARTMENT MODELING

The general form of the three-compartment model is shown in Figure 7.20. As before, we begin with the general form of the three-compartment model and then examine special cases. To analyze the system in Figure 7.20, conservation of mass is used to write a differential equation for each compartment describing the rate of change of the quantity of solute in the compartment, given as accumulation = input – output, where

Compartment 1	Compartment 2	Compartment 3
Accumulation $= \dot{q}_1$	Accumulation $= \dot{q}_2$	Accumulation $= \dot{q}_3$
Input $= f_1(t) + K_{21}q_2 + K_{31}q_3$	Input $= f_2(t) + K_{12}q_1 + K_{32}q_3$	Input $= f_3(t) + K_{13}q_1 + K_{23}q_2$
Ouput $= (K_{10} + K_{12} + K_{13})q_1$	Ouput $= (K_{20} + K_{21} + K_{23})q_2$	Ouput $= (K_{30} + K_{31} + K_{32})q_3$

Therefore,

$$\begin{aligned}
\dot{q}_1 &= f_1(t) + K_{21}q_2 + K_{31}q_3 - (K_{10} + K_{12} + K_{13})q_1 \\
\dot{q}_2 &= f_2(t) + K_{12}q_1 + K_{32}q_3 - (K_{20} + K_{21} + K_{23})q_2 \\
\dot{q}_3 &= f_3(t) + K_{13}q_1 + K_{23}q_2 - (K_{30} + K_{31} + K_{32})q_3
\end{aligned} \tag{7.78}$$

The D-Operator is used to simplify the system, where Eq. (7.78) is written in matrix form as

$$D\mathbf{IQ} = \mathbf{AQ} + \mathbf{F} \tag{7.79}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ q_2 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -(K_{10} + K_{12} + K_{13}) & K_{21} & K_{31} \\ K_{12} & -(K_{20} + K_{21} + K_{23}) & K_{32} \\ K_{13} & K_{23} & -(K_{30} + K_{31} + K_{32}) \end{bmatrix}, \mathbf{F} = \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_2(t) \end{bmatrix}$$

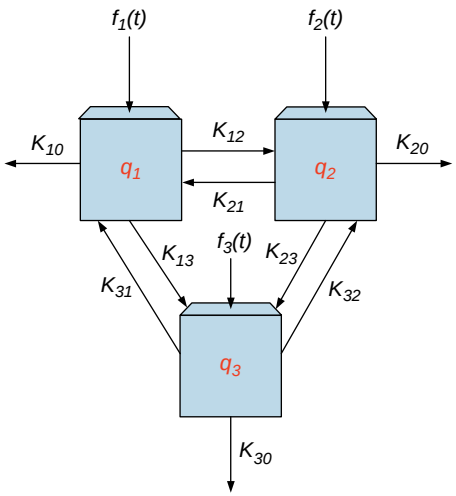


FIGURE 7.20 A general three-compartment model. Compartment 1 has volume V_1, compartment 2 has volume V_2, and compartment 3 has volume V_3.

To make the solution more readable, matrix $\mathbf{A}$ is written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Solving Eq. (7.79) gives

$$\mathbf{Q} = (D\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} = \frac{1}{\det(D\mathbf{I} - \mathbf{A})}\text{adj}(D\mathbf{I} - \mathbf{A})\mathbf{F} \tag{7.80}$$

or

$$\det(D\mathbf{I} - \mathbf{A})\mathbf{Q} = \text{adj}(D\mathbf{I} - \mathbf{A})\mathbf{F}$$

and using MATLAB, we have

```
>> syms D q1 q2 q3 a11 a12 a13 a21 a22 a23 a31 a32 a33
>> A=[a11 a12 a13; a21 a22 a23; a31 a32 a33];
>> det(D*eye(3)-A)
ans =
D^3-D^2*a33-a22*D^2+D*a22*a33-D*a23*a32-
a11*D^2+a11*D*a33+a11*a22*D-a11*a22*a33+a11*a23*a32-
a21*a12*D+a21*a12*a33-a21*a13*a32-a31*a12*a23-
a31*a13*D+a31*a13*a22
>> adj=det(D*eye(3)-A)*inv(D*eye(3)-A)
  adj =
[ D^2-D*a33-a22*D+a22*a33-a23*a32, a12*D-a12*a33+a13*a32,
a12*a23+a13*D-a13*a22]
[ a21*D-a21*a33+a23*a31, D^2-D*a33-a11*D+a11*a33-a13*a31,
a23*D-a23*a11+a13*a21]
[ a21*a32+a31*D-a31*a22, a32*D-a32*a11+a12*a31, D^2-a22*D-
a11*D+a11*a22-a12*a21]
```

Substituting the values from MATLAB into Eq. (7.80) gives

$$\begin{pmatrix} D^3 - (a_{11}+a_{22}+a_{33})D^2 + (a_{11}a_{22}+a_{22}a_{33}+a_{33}a_{11}-a_{12}a_{21}-a_{13}a_{31}-a_{23}a_{32})D \\ -a_{11}a_{22}a_{33}+a_{11}a_{23}a_{32}+a_{22}a_{13}a_{31}+a_{33}a_{12}a_{21}-a_{12}a_{23}a_{31}-a_{13}a_{32}a_{21} \end{pmatrix}\mathbf{Q} =$$

$$\begin{bmatrix} D^2-(a_{33}+a_{22})D+a_{22}a_{33}-a_{23}a_{32} & a_{12}D-a_{12}a_{33}+a_{13}a_{32} & a_{13}D+a_{12}a_{23}-a_{13}a_{22} \\ a_{21}D-a_{21}a_{33}+a_{23}a_{31} & D^2-(a_{33}+a_{11})D+a_{11}a_{33}-a_{13}a_{31} & a_{23}D-a_{23}a_{11}+a_{13}a_{21} \\ a_{31}D+a_{21}a_{32}-a_{31}a_{22} & a_{32}D-a_{32}a_{11}+a_{12}a_{31} & D^2-(a_{22}+a_{11})D+a_{11}a_{22}-a_{12}a_{21} \end{bmatrix}\mathbf{F} \tag{7.81}$$

Returning to the time domain gives the following independent differential equations:

$$\begin{aligned} &\dddot{q}_1 - (a_{11}+a_{22}+a_{33})\ddot{q}_1 + (a_{11}a_{22}+a_{22}a_{33}+a_{33}a_{11}-a_{12}a_{21}-a_{13}a_{31}-a_{23}a_{32})\dot{q}_1 \\ &\quad + (a_{11}a_{23}a_{32}-a_{11}a_{22}a_{33}+a_{22}a_{13}a_{31}+a_{33}a_{12}a_{21}-a_{12}a_{23}a_{31}-a_{13}a_{32}a_{21})q_1 \\ &\quad = \ddot{f}_1 - (a_{33}+a_{22})\dot{f}_1 + (a_{22}a_{33}-a_{23}a_{32})f_1 + a_{12}\dot{f}_2 - (a_{12}a_{33}-a_{13}a_{32})f_2 \\ &\quad + a_{13}\dot{f}_3 + (a_{12}a_{23}-a_{13}a_{22})f_3 \end{aligned} \tag{7.82}$$

$$\begin{aligned}&\dddot{q}_2-(a_{11}+a_{22}+a_{33})\ddot{q}_2+(a_{11}a_{22}+a_{22}a_{33}+a_{33}a_{11}-a_{12}a_{21}-a_{13}a_{31}-a_{23}a_{32})\dot{q}_2\\&\quad+(a_{11}a_{23}a_{32}-a_{11}a_{22}a_{33}+a_{22}a_{13}a_{31}+a_{33}a_{12}a_{21}-a_{12}a_{23}a_{31}-a_{13}a_{32}a_{21})q_2\\&\quad=a_{21}\dot{f}_1-(a_{21}a_{33}-a_{23}a_{31})f_1+\ddot{f}_2-(a_{33}+a_{11})\dot{f}_2+(a_{11}a_{33}-a_{13}a_{31})f_2\\&\quad+a_{23}\dot{f}_3-(a_{23}a_{11}-a_{13}a_{21})f_3\end{aligned}\tag{7.83}$$

$$\begin{aligned}&\dddot{q}_3-(a_{11}+a_{22}+a_{33})\ddot{q}_3+(a_{11}a_{22}+a_{22}a_{33}+a_{33}a_{11}-a_{12}a_{21}-a_{13}a_{31}-a_{23}a_{32})\dot{q}_3\\&\quad+(a_{11}a_{23}a_{32}-a_{11}a_{22}a_{33}+a_{22}a_{13}a_{31}+a_{33}a_{12}a_{21}-a_{12}a_{23}a_{31}-a_{13}a_{32}a_{21})q_3\\&\quad=a_{31}\dot{f}_1+(a_{21}a_{32}-a_{31}a_{22})f_1+a_{32}\dot{f}_2-(a_{32}a_{11}-a_{12}a_{31})f_2\\&\quad+\ddot{f}_3-(a_{22}+a_{11})\dot{f}_3+(a_{11}a_{22}-a_{12}a_{21})f_3\end{aligned}\tag{7.84}$$

The characteristic equation, $\det(D\mathbf{I}-\mathbf{A})$, is identical for q_1, q_2, and q_3, as well as the form of the natural response. Note that the coefficients in the natural response are not identical and depend on the input to the compartments and the initial conditions. The characteristic equation is

$$\begin{aligned}&s^3-(a_{11}+a_{22}+a_{33})s^2+(a_{11}a_{22}+a_{22}a_{33}+a_{33}a_{11}-a_{12}a_{21}-a_{13}a_{31}-a_{23}a_{32})s\\&\quad-a_{11}a_{22}a_{33}+a_{11}a_{23}a_{32}+a_{22}a_{13}a_{31}+a_{33}a_{12}a_{21}-a_{12}a_{23}a_{31}-a_{13}a_{32}a_{21}=0\end{aligned}\tag{7.85}$$

The roots of the characteristic equation are determined using the MATLAB command "`eig`" and may be underdamped, overdamped, or critically damped, depending on the transfer rates. The expressions for the roots are far too complex to be usable and will not be written here. With a two-compartment model, there was only one way for repeat roots to occur. With a three-compartment model, there are many more configurations for repeat roots to occur. Complex roots can occur under certain conditions, which are discussed in Section 7.7.2.

In the remainder of this section and the next, we consider special cases of the three-compartment model: mammillary, catenary, and unilateral. Each model may be closed and may have sink and source compartments.

7.7.1 Mammillary Three-Compartment Model

A mammillary three-compartment model is shown in Figure 7.21, which is characterized by a central compartment connected to two peripheral compartments. All exchanges of the solute are through the central compartment, and there is no direct exchange of solute

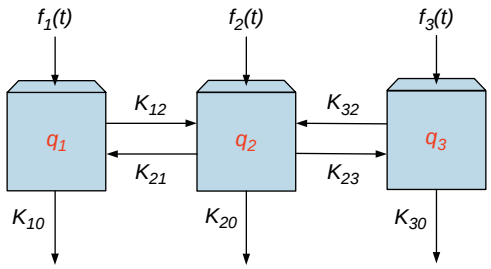


FIGURE 7.21 A mammillary three-compartment model.

between compartments 1 and 3. Each compartment can have an input and an output to the environment.

The mammillary three-compartment model is given by the following set of equations:

$$\begin{aligned}\dot{q}_1 &= f_1(t) + K_{21}q_2 - (K_{10} + K_{12})q_1 \\ \dot{q}_2 &= f_2(t) + K_{12}q_1 - (K_{20} + K_{21} + K_{23})q_2 + K_{32}q_3 \\ \dot{q}_3 &= f_3(t) + K_{23}q_2 - (K_{30} + K_{32})q_3\end{aligned} \tag{7.86}$$

With

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} -(K_{10} + K_{12}) & K_{21} & 0 \\ K_{12} & -(K_{20} + K_{21} + K_{23}) & K_{32} \\ 0 & K_{23} & -(K_{30} + K_{32}) \end{bmatrix}$$

and Eq. (7.81), we have

$$\begin{aligned}&\begin{pmatrix} D^3 - (a_{11} + a_{22} + a_{33})D^2 + (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32})D \\ -a_{11}a_{22}a_{33} + a_{11}a_{23}a_{32} + a_{33}a_{12}a_{21} - a_{12}a_{23}a_{31} \end{pmatrix}\mathbf{Q} = \\ &\begin{bmatrix} D^2 - (a_{33} + a_{22})D + a_{22}a_{33} - a_{23}a_{32} & a_{12}D - a_{12}a_{33} & a_{12}a_{23} \\ a_{21}D - a_{21}a_{33} & D^2 - (a_{33} + a_{11})D + a_{11}a_{33} & a_{23}D - a_{23}a_{11} \\ a_{21}a_{32} & a_{32}D - a_{32}a_{11} & D^2 - (a_{22} + a_{11})D + a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}\mathbf{F}\end{aligned} \tag{7.87}$$

Returning to the time domain gives the following independent differential equations:

$$\begin{aligned}&\dddot{q}_1 - (a_{11} + a_{22} + a_{33})\ddot{q}_1 + (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32})\dot{q}_1 \\ &\quad + (a_{11}a_{23}a_{32} + a_{33}a_{12}a_{21} - a_{11}a_{22}a_{33})q_1 \\ &\quad = \ddot{f}_1 - (a_{33} + a_{22})\dot{f}_1 + (a_{22}a_{33} - a_{23}a_{32})f_1 + a_{12}\dot{f}_2 - a_{12}a_{33}f_2\end{aligned} \tag{7.88}$$

$$\begin{aligned}&\dddot{q}_2 - (a_{11} + a_{22} + a_{33})\ddot{q}_2 + (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32})\dot{q}_2 \\ &\quad + (a_{11}a_{23}a_{32} + a_{33}a_{12}a_{21} - a_{11}a_{22}a_{33})q_2 \\ &\quad = a_{21}\dot{f}_1 - (a_{21}a_{33} - a_{23}a_{31})f_1 + \ddot{f}_2 - (a_{33} + a_{11})\dot{f}_2 + a_{11}a_{33}f_2 + a_{23}\dot{f}_3 - a_{23}a_{11}f_3\end{aligned} \tag{7.89}$$

$$\begin{aligned}&\dddot{q}_3 - (a_{11} + a_{22} + a_{33})\ddot{q}_3 + (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32})\dot{q}_3 \\ &\quad + (a_{11}a_{23}a_{32} + a_{33}a_{12}a_{21} - a_{11}a_{22}a_{33})q_3 \\ &\quad = a_{21}a_{32}f_1 + a_{32}\dot{f}_2 - a_{32}a_{11}f_2 + \ddot{f}_3 - (a_{22} + a_{11})\dot{f}_3 + (a_{11}a_{22} - a_{12}a_{21})f_3\end{aligned} \tag{7.90}$$

The roots of a mammillary three-compartment model are all real and determined from the characteristic equation

$$\begin{aligned}&s^3 - (a_{11} + a_{22} + a_{33})s^2 + (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32})s \\ &\quad + a_{11}a_{23}a_{32} + a_{33}a_{12}a_{21} - a_{11}a_{22}a_{33} = 0\end{aligned}$$

EXAMPLE PROBLEM 7.13

Consider the mammillary three-compartment model shown in Figure 7.21, with a loss of solute to the environment only from compartment 1 and input only from compartment 2. Additionally, K_{12}=2, K_{21}=1.5, K_{10}=0.5, K_{23}=1.3, K_{32}=0.4, and $f_2(t)=10\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

Solution

With the input $f_2(t)=10\delta(t)$ transformed into a change in initial condition for compartment 2 to $q_2(0)$=10 and no input, conservation of mass for each compartment yields

$$\begin{aligned}\dot{q}_1 &= K_{21}q_2 - (K_{10}+K_{12})q_1 = -2.5q_1 + 1.5q_2\\ \dot{q}_2 &= K_{12}q_1 - (K_{21}+K_{23})q_2 + K_{32}q_3 = 2q_1 - 2.8q_2 + 0.4q_3\\ \dot{q}_3 &= K_{23}q_2 - (K_{30}+K_{32})q_3 = 1.3q_2 - 0.4q_3\end{aligned} \tag{7.91}$$

Using the D-Operator method with MATLAB, we get

```
>> syms D
>> A=[-2.5 1.5 0;2 -2.8 0.4;0 1.3 -0.4];
>> det(D*eye(3)-A)
ans =

   D^3+57/10*D^2+28/5*D+3/10
```

and

$$\dddot{q}_1 + 5.7\ddot{q}_1 + \frac{28}{5}\dot{q}_1 + 0.3q_1 = 0$$

$$\dddot{q}_2 + 5.7\ddot{q}_2 + \frac{28}{5}\dot{q}_2 + 0.3q_2 = 0$$

$$\dddot{q}_3 + 5.7\ddot{q}_3 + \frac{28}{5}\dot{q}_3 + 0.3q_3 = 0$$

Using the "`eig(A)`" command gives the roots as −4.46, −1.18, and −0.06. Thus, we have

$$\begin{aligned}q_1 &= B_1e^{-4.46t} + B_2e^{-1.18t} + B_3e^{-0.06t}\\ q_2 &= B_4e^{-4.46t} + B_5e^{-1.18t} + B_6e^{-0.06t}\\ q_3 &= B_7e^{-4.46t} + B_8e^{-1.18t} + B_9e^{-0.06t}\end{aligned}$$

(since the forced response is zero). Note that since there is no input, all we needed to do was define the matrix **A** and then use the "`eig(A)`" command (i.e., no need to use the "`det`" command). However, we shall use the "`det`" command because it gives the intermediate result.

The initial conditions are $q_1(0)$=0, $q_2(0)$=10, and $q_3(0)$=0. To determine the initial conditions for the derivative terms, we use Eq. (7.91) and get

$$\begin{aligned}\dot{q}_1(0) &= -2.5q_1(0) + 1.5q_2(0) = 15\\ \dot{q}_2(0) &= 2q_1(0) - 2.8q_2(0) + 0.4q_3(0) = -28\\ \dot{q}_3(0) &= 1.3q_2(0) - 0.4q_3(0) = 13\end{aligned}$$

Continued

To determine the initial conditions for the second derivative, we take the derivative of Eq. (7.91) and set $t = 0$, giving

$$\ddot{q}_1(0) = -2.5\dot{q}_1(0) + 1.5\dot{q}_2(0) = -79.5$$
$$\ddot{q}_2(0) = 2\dot{q}_1(0) - 2.8\dot{q}_2(0) + 0.4\dot{q}_3(0) = 113.6$$
$$\ddot{q}_3(0) = 1.3\dot{q}_2(0) - 0.4\dot{q}_3(0) = -41.6$$

Solution details are provided for q_1 here, and a final solution for q_2 and q_3. Using the initial conditions, we solve for B_1, B_2 and B_3 from

$$q_1(0) = 0 = B_1 + B_2 + B_3$$
$$\dot{q}_1(0) = 15 = -4.46B_1 - 1.18B_2 - 0.06B_3$$
$$\ddot{q}_1(0) = -79.5 = 19.9B_1 + 1.4B_2 + 0.0036B_3$$

giving

$$B_1 = -4.219,\ B_2 = 3.1818,\ \text{and}\ B_3 = 1.0372.$$

Therefore,

$$q_1 = \left(-4.219e^{-4.46t} + 3.1818e^{-1.18t} + 1.0372e^{-0.06t}\right)u(t)$$

We repeat this process for q_2 and q_3, yielding

$$q_2 = \left(5.51e^{-4.46t} + 2.8179e^{-1.18t} + 1.6721e^{-0.06t}\right)u(t)$$
$$q_3 = \left(-1.762e^{-4.46t} - 4.6849e^{-1.18t} + 6.4469e^{-0.06t}\right)u(t)$$

7.7.2 The Unilateral Three-Compartment Model

A unilateral three-compartment model is shown in Figure 7.22, which is characterized by a closed loop of connected compartments, whereby the solute circulates around the loop in one direction only. Each compartment can have an input and an output to the environment.

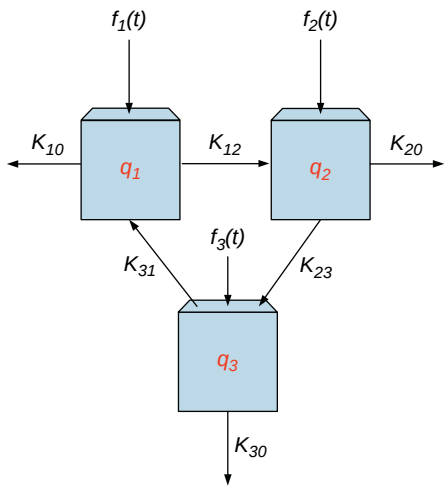


FIGURE 7.22 A unilateral three-compartment model.

In general, the unilateral three-compartment model is given by the following set of equations:

$$\begin{aligned}\dot{q}_1 &= f_1(t) + K_{31}q_3 - (K_{10} + K_{12})q_1 \\ \dot{q}_2 &= f_2(t) + K_{12}q_1 - (K_{20} + K_{23})q_2 \\ \dot{q}_3 &= f_3(t) + K_{23}q_2 - (K_{30} + K_{31})q_3\end{aligned} \tag{7.92}$$

To examine a simple unilateral three-compartment model with complex roots, consider a closed system (i.e., $K_{10} = K_{20} = K_{30} = 0$). From Eq. (7.85), the roots are found from the characteristic equation, given as

$$s^3 + (K_{12} + K_{23} + K_{31})s^2 + (K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})s = 0 \tag{7.93}$$

which are $s_1 = 0$, and

$$s_{2,3} = -\frac{(K_{12} + K_{23} + K_{31})}{2} \pm \frac{1}{2}\sqrt{(K_{12} + K_{23} + K_{31})^2 - 4(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})}$$

Complex roots occur when $4(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12}) > (K_{12} + K_{23} + K_{31})^2$. Repeated roots occur when $4(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12}) = (K_{12} + K_{23} + K_{31})^2$.

Consider the case of complex roots and a zero root, which gives rise to a natural solution of the form

$$q_i = B_1 + e^{-\alpha t}(B_2 \cos\omega_d t + B_3 \sin\omega_d t) = B_1 + B_4 e^{-\alpha t} \cos(\omega_d t + \phi)$$

where α and ω_d are the real and imaginary part of the complex root, and the B_i terms are determined from initial conditions after the forced response is determined. We write the complex roots in standardized format as $s_{2,3} = -\zeta\omega_0 \pm \omega_0\sqrt{\zeta^2 - 1}$, which has a characteristic equation of

$$s^2 + 2\,\zeta\omega_0 s + \omega_0^2 = 0 \tag{7.94}$$

The system is at its most oscillatory when $\zeta = 0$, a pure sinusoid.

To get a better understanding of the system, we determine the extent of its oscillatory behavior by finding the optimal values of the transfer rates to achieve maximum oscillatory behavior (i.e., minimum ζ). To write an expression for ζ, we use the coefficients of the characteristic equation (Eq. (7.93)) and set them equal to the terms in Eq. (7.94):

$$\begin{aligned}2\zeta\omega_0 &= (K_{12} + K_{23} + K_{31}) \\ \omega_0^2 &= (K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})\end{aligned}$$

which gives

$$\omega_0 = \sqrt{(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})}$$

and

$$\zeta = \frac{1}{2\omega_0} = \frac{(K_{12} + K_{23} + K_{31})}{2\sqrt{(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})}} \tag{7.95}$$

To find the minimum ζ, we find $\dfrac{\partial\zeta}{\partial K_{12}} = 0$, $\dfrac{\partial\zeta}{\partial K_{23}} = 0$, and $\dfrac{\partial\zeta}{\partial K_{31}} = 0$, which allows us to determine the conditions that allow minimum ζ. First, we use the chain rule to find

$$\frac{\partial \zeta}{\partial K_{12}} = \frac{2(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})^{\frac{1}{2}} - (K_{23} + K_{31})(K_{12} + K_{23} + K_{31})(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})^{-\frac{1}{2}}}{4(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})}$$

$$= \frac{2(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12}) - (K_{23} + K_{31})(K_{12} + K_{23} + K_{31})}{4(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12})^{\frac{3}{2}}} = 0$$

The minimum occurs when the numerator is zero—that is,

$$\begin{aligned} 2(K_{12}K_{23} + K_{23}K_{31} + K_{31}K_{12}) &= (K_{23} + K_{31})(K_{12} + K_{23} + K_{31}) \\ &= K_{23}K_{12} + K_{23}^2 + K_{23}K_{31} + K_{31}K_{12} + K_{31}K_{23} + K_{31}^2 \end{aligned}$$

Simplifying, we have

$$K_{12}(K_{23} + K_{31}) = K_{23}^2 + K_{31}^2$$

or

$$K_{12} = \frac{K_{23}^2 + K_{31}^2}{(K_{23} + K_{31})}$$

Repeating for $\frac{\partial \zeta}{\partial K_{23}} = 0$, we get $K_{23} = \frac{K_{12}^2 + K_{31}^2}{(K_{12} + K_{31})}$, and for $\frac{\partial \zeta}{\partial K_{31}} = 0$, we get $K_{31} = \frac{K_{12}^2 + K_{23}^2}{(K_{12} + K_{23})}$. The only way these relationships are valid is if $K_{12} = K_{23} = K_{31} = K$, and from Eq. (7.95), we find $\zeta = \frac{3}{2\sqrt{3}} = 0.866$, which does not have a very noticeable oscillatory behavior. We will see in the next section that a more noticeable oscillatory response is possible with more than three compartments.

EXAMPLE PROBLEM 7.14

Consider the unilateral three-compartment model shown in Figure 7.22 with no loss of solute to the environment from any compartments and an input for compartment 3 only. Additionally, $K_{12} = K_{23} = K_{31} = 2$, and $f_3(t) = 5\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

Solution

As before, we transform the input, $f_3(t) = 5\delta(t)$, into a change in initial condition for compartment 3 to $q_3(0)$=5 and no input. The conservation of mass for each compartment yields

$$\begin{aligned} \dot{q}_1 &= K_{31}q_3 - (K_{10} + K_{12})q_1 = -2q_1 + 2q_3 \\ \dot{q}_2 &= K_{12}q_1 - (K_{20} + K_{23})q_2 = 2q_1 - 2q_2 \\ \dot{q}_3 &= K_{23}q_2 - (K_{30} + K_{31})q_3 = 2q_2 - 2q_3 \end{aligned} \tag{7.96}$$

Using the D-Operator method with MATLAB, we get

```
>> syms D
>> A=[-2 0 2;2 -2 0;0 2 -2];
>> det(D*eye(3)-A)
ans =

   D^3+6*D^2+12*D
```

and

$$\dddot{q}_1 + 6\ddot{q}_1 + 12\dot{q}_1 = 0$$
$$\dddot{q}_2 + 6\ddot{q}_2 + 12\dot{q}_2 = 0$$
$$\dddot{q}_3 + 6\ddot{q}_3 + 12\dot{q}_3 = 0$$

The roots from the characteristic equation are $0, -3 \pm j1.7321$. The complete solution is the natural solution, since the forced response is zero, and is given by

$$q_1 = B_1 + e^{-3t}(B_2 \cos 1.7321t + B_3 \sin 1.7321t)$$
$$q_2 = B_4 + e^{-3t}(B_5 \cos 1.7321t + B_6 \sin 1.7321t)$$
$$q_3 = B_7 + e^{-3t}(B_8 \cos 1.7321t + B_9 \sin 1.7321t)$$

The initial conditions are $q_1(0)=0$, $q_2(0)=0$, and $q_3(0)=5$. To determine the initial conditions for the derivative terms, we use Eq. (7.96) and get

$$\dot{q}_1(0) = -2q_1(0) + 2q_3(0) = 10$$
$$\dot{q}_2(0) = 2q_1(0) - 2q_2(0) = 0$$
$$\dot{q}_3(0) = 2q_2(0) - 2q_3(0) = -10$$

To determine the initial conditions for the second derivative, we take the derivative of Eq. (7.96) with $t = 0$, giving

$$\ddot{q}_1(0) = -2\dot{q}_1(0) + 2\dot{q}_3(0) = -40$$
$$\ddot{q}_2(0) = 2\dot{q}_1(0) - 2\dot{q}_2(0) = 20$$
$$\ddot{q}_3(0) = 2\dot{q}_2 - 2\dot{q}_3 = 20$$

For q_1, we have

$$q_1(0) = 0 = B_1 + B_2$$
$$\dot{q}_1(0) = 10 = -3B_2 + 1.7321B_3$$
$$\ddot{q}_1(0) = -40 = 6B_2 - 10.4B_3$$

which gives $B_1 = \frac{5}{3}, B_2 = -\frac{5}{3}$, and $B_3 = 2.9$, and

$$q_1 = \left(\frac{5}{3} - e^{-3t}\left(\frac{5}{3} \cos 1.7321t - 2.9 \sin 1.7321t\right)\right)u(t)$$

Repeating for q_2 and q_3, we have

$$q_2 = \left(\frac{5}{3} - e^{-3t}\left(\frac{5}{3} \cos 1.7321t + 2.9 \sin 1.7321t\right)\right)u(t)$$

$$q_3 = \frac{5}{3}\left(1 + 2e^{-3t} \cos 1.7321t\right)u(t)$$

Continued

Illustrated in Figure 7.23 is a plot of the quantity in each compartment. While it is difficult to see the oscillations, the first peak is evident by the overshoot or undershoot. To determine the time at peak undershoot for q_3, we use the technique of Section 2.9.2 by finding the time that satisfies $\frac{\partial q_3}{\partial t} = 0$, which gives $T_{p_3} = 1.21$. Similarly, $T_p = 1.81$ for both q_1 and q_2.

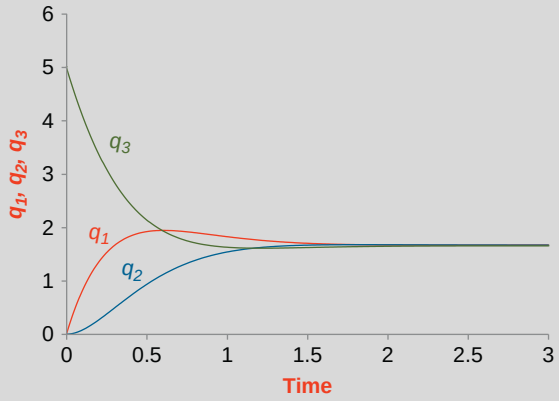


FIGURE 7.23 Illustration of the quantity in each compartment in Example Problem 7.14.

7.7.3 Source Compartment

In Section 7.5.1, a source compartment in a two-compartment model was described as one that only has output to other compartments, without any inputs from other compartments. A source compartment also appears in three-compartment models, whose output is solved independent of the other compartments as before.

The following example involves a three-compartment mammillary model with a source compartment, as illustrated in Figure 7.24. The body is now divided into the digestive system, plasma, and the tissues to more accurately depict their behavior.

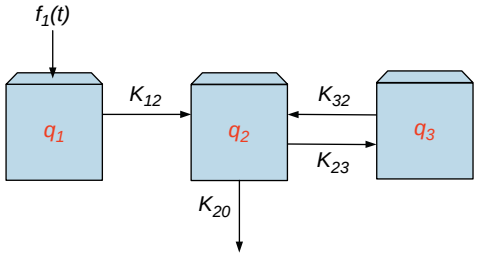


FIGURE 7.24 Illustration for Example Problem 7.15. Compartment 1 is the digestive system, compartment 2 is the plasma, and compartment 3 is the tissues.

EXAMPLE PROBLEM 7.15

Consider the mammillary three-compartment model with the source compartment shown in Figure 7.24. The input is $f_1(t)=\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

Solution

Once again, the input is transformed into a change in initial condition for compartment 1, $q_1(0)=1$. The equations describing this model are

$$\dot{q}_1 = -K_{12}q_1 \tag{7.97}$$

$$\dot{q}_2 = K_{12}q_1 - (K_{20} + K_{23})q_2 + K_{32}q_3 \tag{7.98}$$

$$\dot{q}_3 = K_{23}q_2 - K_{32}q_3 \tag{7.99}$$

Since Eq. (7.97) involves only q_1, it is easily solved as $q_1 = q_1(0)e^{-K_{12}t}u(t) = e^{-K_{12}t}u(t)$. Substituting the solution for q_1 into Eqs. (7.98) and (7.99), we now have two equations as

$$\dot{q}_2 = K_{12}e^{-K_{12}t}u(t) - (K_{20} + K_{23})q_2 + K_{32}q_3 \tag{7.100}$$

$$\dot{q}_3 = K_{23}q_2 - K_{32}q_3 \tag{7.101}$$

The D-Operator gives the reconstructed differential equations for q_2 and q_3 as

```
>> syms D K20 K23 K32
>> A=[-(K20+K23) K32;K23 -K32];
>> det(D*eye(2)-A)
ans =
  D^2+D*K32+K20*D+K20*K32+K23*D
>> adj=det(D*eye(2)-A)*inv(D*eye(2)-A)
adj =

   [  D+K32, K32       ]
   [  K23,   D+K20+K23]
```

and

$$\ddot{q}_2 + (K_{32} + K_{20} + K_{23})\dot{q}_2 + K_{20}K_{32}q_2 = K_{12}(K_{32} - K_{12})e^{-K_{12}t} \tag{7.102}$$

$$\ddot{q}_3 + (K_{32} + K_{20} + K_{23})\dot{q}_3 + K_{20}K_{32}q_3 = K_{23}K_{12}e^{-K_{12}t} \tag{7.103}$$

The roots are

```
>> eig(A)
ans =

  -1/2*K20-1/2*K23-1/2*K32+1/2*(K20^2+2*K20*K23-
  2*K20*K32+K23^2+2*K32*K23+K32^2)^(1/2)

  -1/2*K20-1/2*K23-1/2*K32-1/2*(K20^2+2*K20*K23-
  2*K20*K32+K23^2+2*K32*K23+K32^2)^(1/2)
```

Continued

which simplifies to

$$s_{1,2} = -\frac{(K_{20} + K_{23} + K_{32})}{2} \pm \frac{1}{2}\sqrt{(K_{32} + K_{20} + K_{23})^2 - 4K_{32}K_{23}} \tag{7.104}$$

Only real roots are possible with $K_{ij} \geq 0$. The natural response for q_2 is

$$q_{2_n} = B_1 e^{s_1 t} + B_2 e^{s_2 t}$$

The forced response is $q_{2_f} = B_3 e^{-K_{12}t}$, which after substituting into Eq. (7.103) gives

$$B_3 = \frac{K_{23}K_{12}}{K_{12}^2 - K_{12}(K_{32} + K_{20} + K_{23}) + K_{20}K_{32}}$$

The complete response is then

$$q_2 = B_1 e^{s_1 t} + B_2 e^{s_2 t} + B_3 e^{-K_{12}t}$$

We use the initial conditions, $q_2(0) = 0$ and $\dot{q}_2(0) = K_{12}q_1(0)$, to solve for B_1 and B_2 as follows:

$$q_2(0) = B_1 + B_2 + B_3 = 0$$

and with

$$\dot{q}_2 = s_1 B_1 e^{s_1 t} + s_2 B_2 e^{s_2 t} - K_{12} B_3 e^{-K_{12}t},$$

we have

$$\dot{q}_2(0) = K_{12} = s_1 B_1 + s_2 B_2 - K_{12} B_3$$

To solve for B_1 and B_2, we evaluate

$$\begin{bmatrix} 1 & 1 \\ s_1 & s_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} -B_3 \\ K_{12}(1 + B_3) \end{bmatrix}$$

which gives

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \dfrac{K_{12}B_3 + K_{12} + B_3 s_2}{(s_1 - s_2)} \\ -\dfrac{K_{12}B_3 + K_{12} + B_3 s_1}{(s_1 - s_2)} \end{bmatrix}$$

The final solution is

$$q_2(t) = \left(\begin{array}{l} \dfrac{K_{12}B_3 + K_{12} + B_3 s_2}{(s_1 - s_2)} e^{s_1 t} - \dfrac{K_{12}B_3 + K_{12} + B_3 s_1}{(s_1 - s_2)} e^{s_2 t} \\ + \dfrac{K_{23}K_{12}}{K_{12}^2 - K_{12}(K_{32} + K_{20} + K_{23}) + K_{20}K_{32}} e^{-K_{12}t} \end{array} \right) u(t)$$

Repeating for q_3, we have

$$q_3(t) = \left(\frac{B_3(K_{12} + s_2)}{(s_1 - s_2)} e^{s_1 t} - \frac{B_3(K_{12} + s_1)}{(s_1 - s_2)} e^{s_2 t} + \frac{K_{23}K_{12}}{K_{12}^2 - K_{12}(K_{32} + K_{20} + K_{23}) + K_{20}K_{32}} e^{-K_{12}t} \right) u(t)$$

7.7.4 Sink Compartment

A sink compartment in a three-compartment model gives rise to a zero root and is described as a compartment with only inputs and no output to other compartments except to the environment. The solution for the sink compartment is found as usual using the D-Operator approach from the resulting quantities in the other two compartments.

To illustrate a sink compartment, Example Problem 7.16 involves a three-compartment model describing the transport of a thyroid hormone to the hepatic duct (sink compartment).[12] The thyroid system was first described in Example Problem 7.5 and is extended to this example. We will then extend the model in Section 7.8.4. Before describing the current model, more background material on the thyroid system is presented.

The thyroid hormones thyroxine (T4) and triiodothyronine (T3), which are produced by the thyroid gland, maintain the body temperature, regulate energy metabolism, and are important for growth and development. The thyroid hormones themselves do not exist inside the thyroid cell but are part of a large thyroglobulin molecule that consists of approximately 70 tyrosine amino acids. The metabolic rate falls to approximately 50 percent of normal without these hormones, and too much thyroid hormone can increase the metabolic rate by 100 percent above the normal rate. The pituitary gland controls the discharge of T4 and T3 through its release of the thyroid-stimulating hormone, TSH. As previously described, the pituitary gland is under the control of the hypothalamus through its release of the thyrotropin-releasing-hormone, TRH.

Ingested iodine, in the form of iodide, is an essential element in the formation of thyroid hormones. Blood flow through the thyroid gland is among the highest of any organ in the body, which allows the quick uptake of iodide. Typically after ingestion, 80 percent of the iodide is rapidly excreted by the kidneys, and the other 20 percent is taken up by the thyroid gland.

Once iodide is taken up by the thyroid gland, it is used in a series of enzyme reactions to create the thyroid hormones. Iodide is first transported across the membrane of the thyroid cell by a pump mechanism called the sodium-iodide symporter. The pump allows iodide concentrations in the thyroid cell to be much greater than in the plasma. Once inside the cell, iodide is oxidized by thyroidal peroxidase to iodine, which iodinates the tyrosine component of the thyroglobulin molecule to first form monoiodotyrosine (MIT) and then diiondotryosine (DIT). Thyroperoixidase then catalyzes the joining of two molecules of DIT to form T4 (a two-benzene ringed structure consisting of an inner tyrosyl ring and an outer phenolic ring, within the thyroglobulin-iodine molecule) or to a lesser extent, the joining of one molecule of MIT to DIT to form T3. Reverse T3 is also formed but is excluded from this discussion.

While the process of creating the thyroglobulin-iodine molecule is quick, the thyroid gland keeps approximately a 60-day supply in reserve. Thyroglobulin itself is not released into the plasma when the thyroid is stimulated by TSH, but T4 and T3 are released through a lysosomal protease enzyme action on the thyroglobulin-iodine molecule. Almost all of the output from the thyroid gland is T4 (greater than 90 percent). In addition, MIT and DIT are released from the thyroglobulin-iodine molecule when T4 and T3 are released; however, MIT and DIT do not leave the cell but are deiodinated, allowing the release of iodine.

[12]See Haddad et al., 2003, in references for original problem developement.

The iodine is then reused in the cell, repeating the enzyme reactions to form the thyroid hormones.

Once in the plasma, the thyroid hormones reversibly combine with proteins. The binding to proteins protects T4 and T3 from immediate metabolism and excretion as they are inactive in this mode. T4 binds primarily with thyroid binding globulin (TBG) and, to a lesser extent, to thyroxine-binding prealbumin (TBPA) and albumin. TBG in the plasma is present in low concentrations and has a high affinity for T4; it usually binds about 70 percent of the available T4. TBPA in the plasma is present in high concentrations but has a low affinity for T4. Ten percent of the plasma-bound hormone T4 is used each day, giving it a 7-day half-life. The 7-day half-life for T4 creates a stable pool of thyroid hormone in the plasma. Approximately 0.04 percent of the T4 is not bound to proteins in the plasma during normal conditions, and we call the unbound T4 *free*-T4.

T3 has a lower binding affinity for plasma proteins. Those T3 that are bound are primarily with TBPA and albumin. T3 is rapidly cleared from the plasma with a half-life of 1 day. Approximately 10 percent of T3 is not bound to proteins in the plasma during normal conditions.

While a small amount of T3 is released by the thyroid into the plasma, almost all of the T3 in the plasma is produced by deiodination enzymes in the liver and to a much lesser extent in the kidneys, where an iodine atom is removed from T4 that converts it into T3. T4 is also eliminated within the liver and to a lesser extent in the kidneys, by conjugation of sulfate or glucuronic acid with the phenolic hydroxyl group of the outer phenolic ring, turning T4 into T3. Also within the liver, T4 undergoes deamination and decarboxylase reactions that convert it into T3.

Once in the plasma, free-T3 moves into the interstitial space and easily moves across the cell membranes in the tissues. When transported into the cells, free-T3 moves into the cell nucleus and binds with thyroid hormone receptors, which then synthesize new proteins through gene transcription. These new proteins are connected to energy metabolism, body temperature, body weight, and the control of growth, reproduction, and differentiation. While T4 also moves into the cell and binds with thyroid hormone receptors, it takes about 10 times more T4 than T3 to equal the effect of T3 in gene transcription.

Example Problem 7.16 involves a three-compartment model describing the transport of T4 throughout the body. To model the transport of T4 in the body, a bolus of radioactive iodine-T4 is injected into the plasma. The use of radioactive iodine-T4 allows us to track the transport of T4 in the body apart from the natural T4. The pathways of interest in the following example, as described in Figure 7.25, involve the plasma, liver, and hepatic duct.

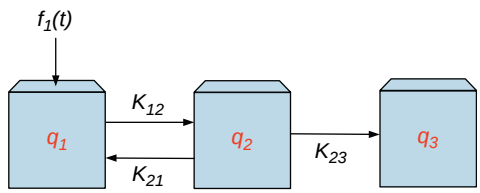


FIGURE 7.25 Illustration for Example Problem 7.16. Compartment 1 is the plasma, compartment 2 is the liver, and compartment 3 is the hepatic duct, simplified to a single compartment. The input $f_1(t)$ is a bolus of radioactive iodine-T4.

Within the liver, some of the T4 is converted into T3 and I^{131}. The I^{131} moves into the hepatic duct, where it is absorbed within the bile. The feedback control by the pituitary gland and the involvement of the kidneys are ignored in this example.

EXAMPLE PROBLEM 7.16

Consider a three-compartment thyroxine transport model as shown in Figure 7.25. The input is $f_1(t) = 0.1 \times 10^{-3}\delta(t)$g of radioactive iodine-T4. Additionally, K_{12}=0.6, K_{21}=0.5, and K_{23}=0.3. Assume that the initial conditions are zero. Solve for the quantity of radioactive iodine-T4 in the plasma compartment.

Solution

Since there is no output to another compartment for compartment 3, this compartment model has a sink for compartment 3. As before, rather than solving the problem with a bolus input, the initial condition is changed for compartment 1 to $q_1(0) = 0.1 \times 10^{-3}$, with zero input. The conservation of mass for each compartment yields

$$\begin{aligned}\dot{q}_1 &= K_{21}q_2 - K_{12}q_1 = -0.6q_1 + 0.5q_2\\ \dot{q}_2 &= K_{12}q_1 - (K_{21} + K_{23})q_2 = 0.6q_1 - 0.8q_2\\ \dot{q}_3 &= K_{23}q_2 = 0.3q_2\end{aligned} \tag{7.105}$$

Using the D-Operator method with MATLAB, we get

```
>> syms D
>> A=[-0.6 0.5 0;0.6 -0.8 0;0 0.3 0];
>> det(D*eye(3)-A)
ans =

   D^3+7/5*D^2+9/50*D
```

and

$$\dddot{q}_1 + \frac{7}{5}\ddot{q}_1 + \frac{9}{50}\dot{q}_1 = 0$$

The "`eig(A)`" command gives the roots as 0, −1.26, and −0.14. Thus, we have

$$q_1 = B_1 + B_2e^{-1.26t} + B_3e^{-0.14t}$$

(since the forced response is zero). The initial conditions are $q_1(0) = 0.1 \times 10^{-3}, q_2(0) = 0$, and $q_3(0)$=0. To determine the initial conditions for the derivative terms, we use Eq. (7.105), which gives

$$\begin{aligned}\dot{q}_1(0) &= -0.6q_1(0) + 0.5q_2(0) = -0.06 \times 10^{-3}\\ \dot{q}_2(0) &= 0.6q_1(0) - 0.8q_2(0) = 0.06 \times 10^{-3}\end{aligned}$$

To determine the initial conditions for the second derivative, we take the derivative of Eq. (7.105) and set t = 0, giving

$$\begin{aligned}\ddot{q}_1(0) &= -0.6\dot{q}_1(0) + 0.5\dot{q}_2(0) = 6.6 \times 10^{-5}\\ \ddot{q}_2(0) &= 0.6\dot{q}_1(0) - 0.8\dot{q}_2(0) = -8.4 \times 10^{-5}\end{aligned}$$

Continued

Using the initial conditions, we solve for B_1, B_2, and B_3 from

$$q_1(0) = 0.1 \times 10^{-3} = B_1 + B_2 + B_3$$
$$\dot{q}_1(0) = -0.06 \times 10^{-3} = -1.26B_2 - 0.14B_3$$
$$\ddot{q}_1(0) = 6.6 \times 10^{-5} = 1.4B_2 + 0.02B_3$$

giving

$$B_1 = 0.48 \times 10^{-4}, B_2 = 0.47 \times 10^{-4}, \text{ and } B_3 = 0.05 \times 10^{-4}, \text{ and}$$
$$q_1 = \left(0.48 + 0.47e^{-1.26t} + 0.05e^{-0.14t}\right) \times 10^{-4} u(t)$$

The model used in Example Problem 7.16 is too simple to capture the real transport dynamics of thyroid hormone. Some investigators have included multiple compartments for the hepatic duct and many other compartments. Some have included chemical reactions in the model. We will investigate these models in a later chapter.

7.8 MULTICOMPARTMENT MODELING

Realistic models of the body typically involve more than three compartments. The concepts described in the previous sections can be applied to a compartment model of any size. Each compartment is characterized by a conservation of mass differential equation describing the rate of change of the solute. Thus, for the case of n compartments, there are n equations of the form

$$\frac{dq_i}{dt} = input - output$$

where q_i is the quantity of solute in compartment i, which can be generalized for the system to

$$D\mathbf{IQ} = \mathbf{AQ} + \mathbf{F} \tag{7.106}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \ \mathbf{F} = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}$$

and for the first row in $\mathbf{A}$, we have

$$a_{11} = -(K_{10} + K_{12} + K_{13} + \cdots + K_{1n})$$
$$a_{12} - K_{12}$$
$$\vdots$$
$$a_{1n} = K_{1n}$$

and so on for the other rows in $\mathbf{A}$. Equation (7.106) is solved as before from

$$\mathrm{Q} = (D\mathrm{I} - \mathrm{A})^{-1}\mathrm{F} = \frac{1}{\det(D\mathrm{I} - \mathrm{A})}\mathrm{adj}(D\mathrm{I} - \mathrm{A})\mathrm{F} \tag{7.107}$$

or

$$\det(D\mathrm{I} - \mathrm{A})\mathrm{Q} = \mathrm{adj}(D\mathrm{I} - \mathrm{A})\mathrm{F}$$

MATLAB is used to reconstruct the differential equations, as before, in terms of a single variable and the inputs. The characteristic equation, $\det(D\mathbf{I}-\mathbf{A})$, is identical for $q_1, q_2, \ldots q_n$, as well as the form of the natural response. The roots of the characteristic equation are determined using the MATLAB command "`eig(A)`" and may be underdamped, overdamped, or critically damped, depending on the transfer rates. The expression for the roots is far too complex to be usable and will not be written here. Most models will have many elements in $\mathbf{A}$ as zero, which makes the solution much more tractable.

In the remainder of this section, we consider special cases of the multicompartment model: mammillary, catenary, and unilateral. Each model may be closed and may have sink and source compartments.

7.8.1 Mammillary Multicompartment Model

A mammillary n-compartment model is shown in Figure 7.26, which is characterized by a central compartment connected to $n-1$ peripheral compartments. All exchange of solute is through the central compartment, and there is no direct exchange of solute among the other compartments. Each compartment can have an input and an output to the environment.

The matrix $\mathbf{A}$, given in Eq. (7.106), has nonzero elements defined as

$$\begin{aligned} a_{11} &= -(K_{10}+K_{12}+K_{13}+\cdots+K_{1n}) \\ a_{ii} &= -(K_{i1}+K_{i0}), && 2 \le i \le n \\ a_{1i} &= K_{i1}, && 2 \le i \le n \\ a_{i1} &= K_{1i}, && 2 \le i \le n \end{aligned} \tag{7.108}$$

This system only has real roots.

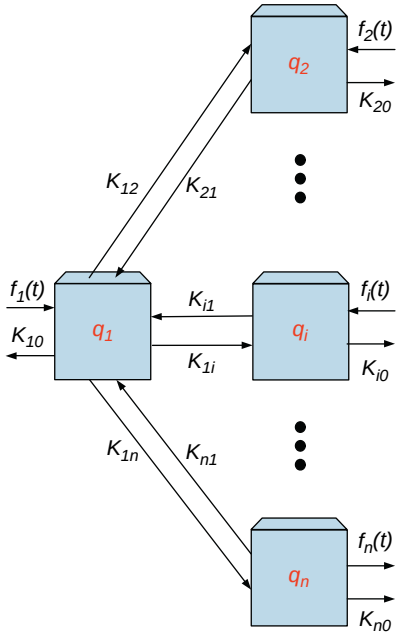


FIGURE 7.26 A mammillary n-compartment model.

7.8.2 Catenary Multicompartment Model

A catenary n-compartment model is shown in Figure 7.27, which is characterized by a chain of compartments, with each compartment exchanging solute with the two adjacent compartments, except for the first and last in the chain. Each compartment can have an input and an output to the environment.

The matrix **A**, given in Eq. (7.106), has nonzero elements defined as

$$\begin{aligned} a_{11} &= -(K_{10} + K_{12}) \\ a_{i,i-1} &= K_{i-1,i}, && 2 \le i \le n \\ a_{ii} &= -\left(K_{i0} + K_{i,(i-1)} + K_{i,(i+1)}\right), && 2 \le i \le n-1 \\ a_{i,i+1} &= K_{i+1,i}, && 2 \le i \le n-1 \\ a_{nn} &= -\left(K_{n0} + K_{n,(n-1)}\right) \end{aligned} \tag{7.109}$$

This system only has real roots.

7.8.3 Unilateral Multicompartment Model

A unilateral n-compartment model is shown in Figure 7.28, which is characterized by a closed loop of connected compartments, whereby solute circulates around the loop in one direction only. Each compartment can have an input and an output to the environment.

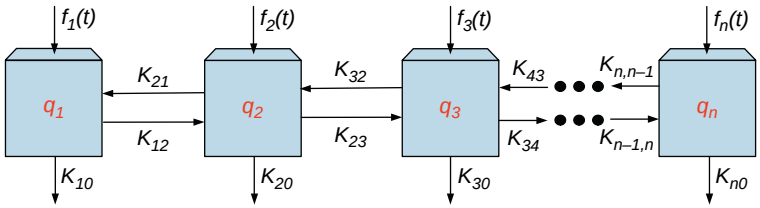


FIGURE 7.27 A catenary n-compartment model.

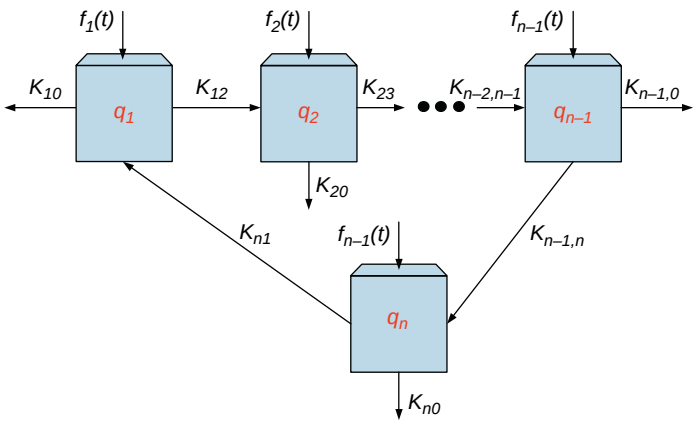


FIGURE 7.28 A unilateral n-compartment model.

The matrix **A**, given in Eq. (7.106), has nonzero elements defined as

$$\begin{aligned} a_{ii} &= -(K_{i0} + K_{i,i+1}), \qquad 1 \le i \le n-1 \\ a_{nn} &= -(K_{n0} + K_{n1}) \\ a_{1n} &= K_{n1} \\ a_{i,i-1} &= K_{i-1,i}, \qquad 2 \le i \le n \end{aligned} \tag{7.110}$$

In Section 7.7.2, we investigated the three-compartment unilateral complex roots case and determined that in the roots with the most oscillatory behavior, all the transfer rates were equal to the same value. We will continue this investigation to explore the oscillatory behavior for a closed system unilateral n-compartment model with equal transfer rates, *K*, and bolus input. The system matrix **A** is

$$A = \begin{bmatrix} -K & 0 & 0 & \cdots & 0 & K \\ K & -K & 0 & \cdots & 0 & 0 \\ 0 & K & -K & \cdots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \cdots & K & -K \end{bmatrix}$$

and the determinant of (D**I**-**A**) is

$$\begin{vmatrix} D-K & 0 & 0 & \cdots & 0 & K \\ K & D-K & 0 & \cdots & 0 & 0 \\ 0 & K & D-K & \cdots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \cdots & K & D-K \end{vmatrix} = (D+K)^n - K^n$$

As Godfrey illustrates, the roots of this system are

$$-K + K\left(\cos\frac{2\pi m}{n} + j \sin\frac{2\pi m}{n} \right), \qquad m = 1, 2, \cdots n \tag{7.111}$$

and lie evenly on the unit circle of radius K, centered at (K,0) in the complex plan. For a closed system, one of the roots is 0, and for an even *n*, there is another root at –2*K*. The remaining roots are complex and given by Eq. (7.111). For $m = 1$ and $m = n-1$, Godfrey shows that the damping ratio is equal to

$$\zeta = \sin\frac{\pi}{n} \tag{7.112}$$

and as *n* increases to infinity, the damping ratio approaches zero. Since the quantity within a compartment can never be less than zero, as *n* approaches infinity, the amplitude of the sinusoid approaches 0.

Our approach to solving a unilateral n-compartment model is the same as before, letting MATLAB do the work for us, as shown in the following example.

EXAMPLE PROBLEM 7.17

Consider the unilateral five-compartment model with no loss of solute to the environment from any compartments and an input for compartment 3 only. Additionally, all transfer rates equal 2 and $f_3(t) = 5\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in compartment 3.

Solution

As before, we transform the input, $f_3(t) = 5\delta(t)$, into a change in initial condition for compartment 3 to $q_3(0) = 5$ and no input. The conservation of mass for each compartment yields

$$\begin{aligned}
\dot{q}_1 &= -2q_1 + 2q_5 \\
\dot{q}_2 &= 2q_1 - 2q_2 \\
\dot{q}_3 &= 2q_2 - 2q_3 \\
\dot{q}_4 &= 2q_3 - 2q_4 \\
\dot{q}_5 &= 2q_4 - 2q_5
\end{aligned}$$

Using MATLAB, we have

```
>> syms D
>> A=[-2 0 0 0 2;2 -2 0 0 0;0 2 -2 0 0;0 0 2 -2 0;0 0 0 2 -2];
>> det(D*eye(5)-A)
ans =

   D^5+10*D^4+40*D^3+80*D^2+80*D
```

and

$$\frac{d^5q_3}{dt^5} + 10\frac{d^4q_3}{dt^4} + 40\frac{d^3q_3}{dt^3} + 80\frac{d^2q_3}{dt^2} + 80\frac{dq_3}{dt} = 0$$

The roots from the characteristic equation (i.e., eig(**A**)) are 0, $-3.6180 \pm j1.1756$, and $-1.3820 \pm j1.9021$. The complete solution for q_3 is the natural solution, since the forced response is zero, and is given by

$$q_3 = B_1 + e^{-3.618t}(B_2\ \cos 1.1756t + B_3\ \sin 1.1756t) + e^{-1.382t}(B_4\ \cos 1.9021t + B_5\ \sin 1.9021t)$$

The initial conditions for q_3 are found using the conservation of mass equations and successive derivatives, giving $q_3(0) = 5$, $\dot{q}_3(0) = -10$, $\ddot{q}_3(0) = 20$, $\dddot{q}_3(0) = -40$, and $\ddddot{q}_3(0) = 80$. Solving for the unknown coefficients using the initial conditions, we have

$$\begin{aligned}
q_3(0) &= 5 = B_1 + B_2 + B_4 \\
\dot{q}_3(0) &= -10 = -3.61B_2 + 1.1756B_3 - 1.382B_4 + 1.9021B_5 \\
\ddot{q}_3(0) &= 20 = 11.72B_2 - 8.51B_3 - 1.71B_4 - 5.26B_5 \\
\dddot{q}_3(0) &= -40 = -32.4B_2 + 44.58B_3 + 7.64B_4 + 4.02B_5 \\
\ddddot{q}_3(0) &= 80 = 64.79B_2 - 199.4B_3 - 18.18B_4 + 9.02B_5
\end{aligned}$$

and using MATLAB yields $B_1 = 1.0$, $B_2 = 2.0$, $B_3 = 0.0$, $B_4 = 2.00$, and $B_5 = 0.0$. Thus, for t$\geq$0, we have

$$q_3 = 1 + 2e^{-3.618t}\ \cos 1.1756t + 2.00e^{-1.382t}\ \cos 1.9021t$$

The same approach can be used to find the quantities in the other compartments, which are plotted in Figure 7.29. Note that the oscillation about the steady state of 1 is more pronounced in q_4 than q_3. In fact, the prominence of oscillation about steady state decreases as we move from q_4 to q_5 to q_1 to q_2 to q_3. In general, we see the most prominent oscillation in the compartment that receives the solute from the compartment stimulated by the bolus. Also note that the steady-state quantity in each compartment equals one.

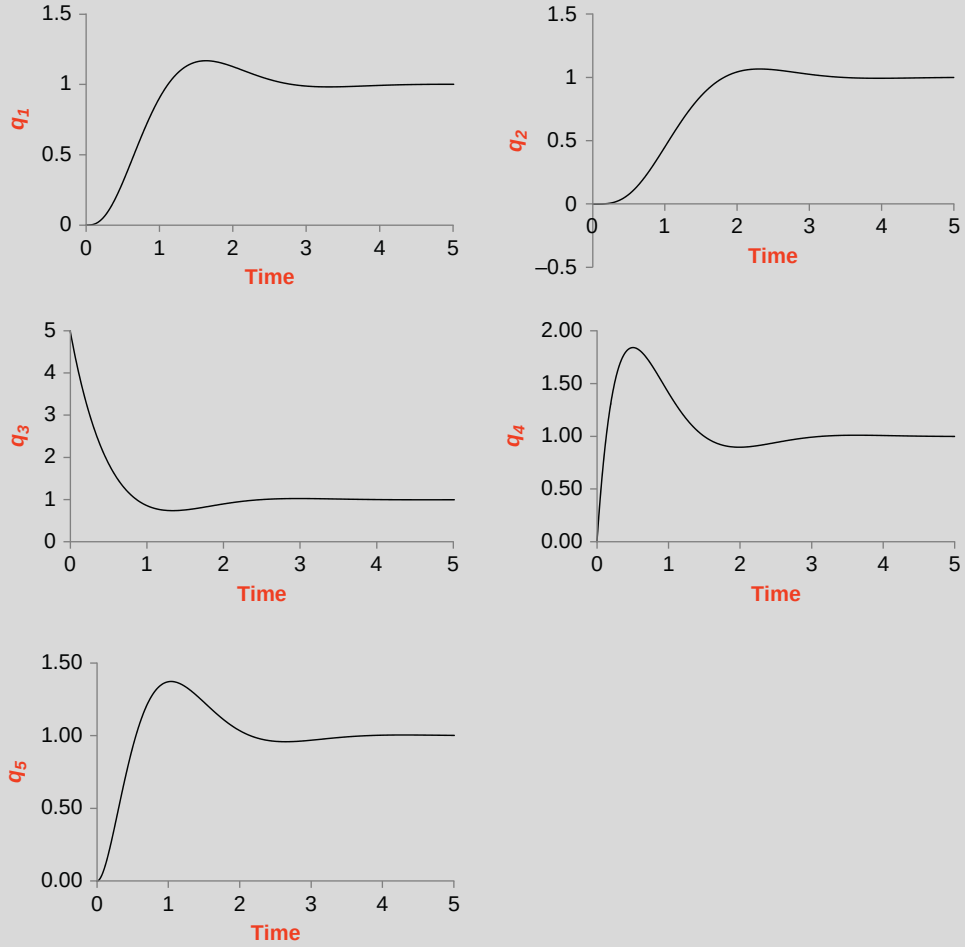


FIGURE 7.29 Plots of the responses for Example Problem 7.17.

As we increase the number of compartments in a closed unilateral system, in general, the more oscillatory the response becomes. In some cases, the oscillatory behavior is prominent but lasts only a very short time. In other cases, the oscillations may be less prominent but last for a longer period of time. Consider the closed system unilateral model shown in

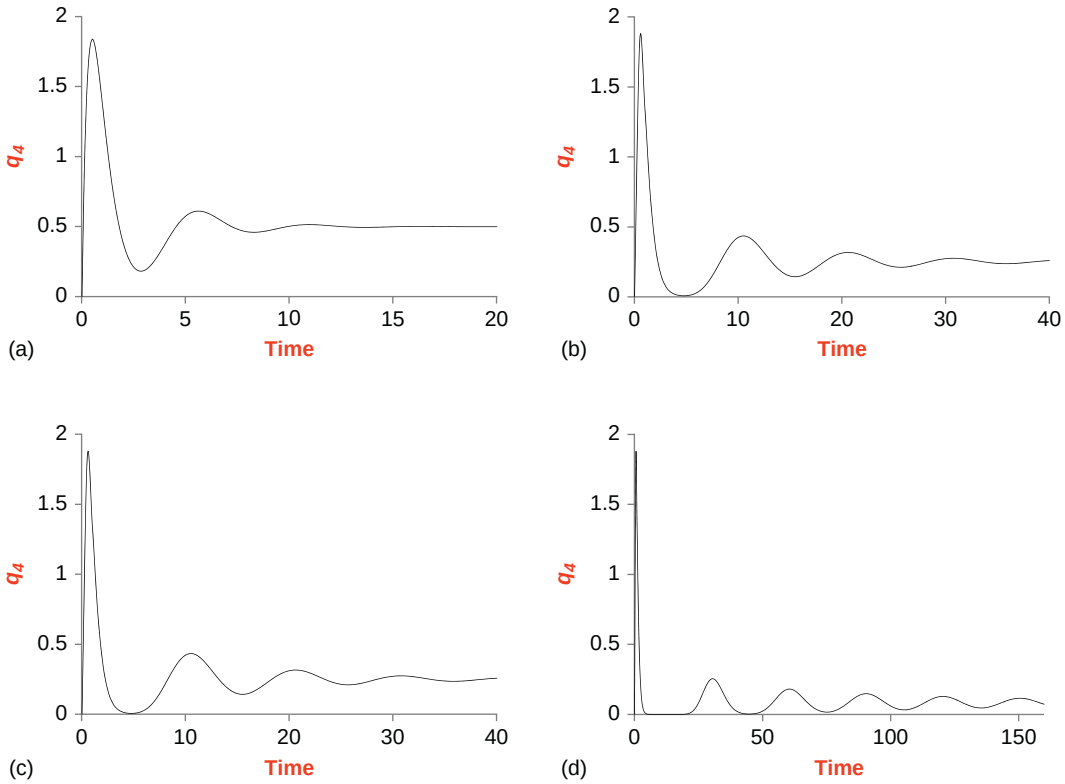


FIGURE 7.30 Plots of the response for q_4 for the closed unilateral system shown in Figure 7.28 for (a) 10, (b) 20, (c) 40, and (d) 80 compartments. A bolus input of 5 to compartment 3 is the only input.

Figure 7.28 (with zero initial conditions and equal transfer rates, $K = 2$) with input $f_3(t) = 5\delta(t)$. The response for a model of 10, 20, 40, and 80 compartments is shown in Figure 7.30 (note the time scale changes for each model). As shown, the number of oscillations increases as the number of compartments increases, and the steady-state value decreases. The time it takes for the solute to move through the system also increases as the number of compartments increases. For 20 compartments and higher, there is essentially no solute left in the compartment after the initial oscillation until the solute flows through the system.

If the system is open and solute is allowed to move into the environment, the oscillatory behavior is reduced. Consider the model used in Figure 7.28, with the exception that solute output to the environment is allowed in compartment 4, with a transfer rate of $K_{40} = 0.2$ (10 percent of the transfer rate among the compartments). Shown in Figure 7.31 is the response for q_4 with 40 and 80 compartments. An oscillatory response is still noted with fewer prominent oscillations as compared with no output to the environment. Also note that the peak oscillation for the first is much smaller than before. Finally, observe that the steady-state value is now zero.

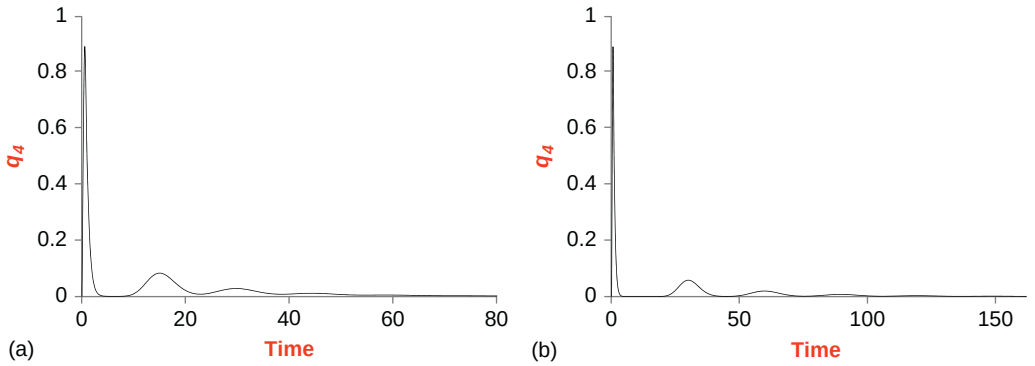


FIGURE 7.31 Plots of the response for q_4 for the open unilateral system shown in Figure 7.27 for (a) 40 and (b) 80 compartments. The model is identical to the one shown in Figure 7.29, except $K_{40} = 0.2$.

7.8.4 General Multicompartment Model

Although the previous models presented in this section are important, many systems are more complex and follow the form of a general multicompartment model. Some systems are composed of subsystems, described in Sections 7.8.1–7.8.3, which are linked together with transference among subunits. For example, a model[13] that describes thyroid hormone distribution and metabolism using two mammillary three-compartment models linked together is shown in Figure 7.32. Mammillary compartments 1–3 describe T3, and mammillary compartments 4–6 describe T4; compartments 1 and 4 are the same space, as are 2 and 5, and 3 and 6. The plasma is represented by compartments 2 and 5, compartments 1 and 4 represent the *fast* tissue (liver, kidneys, lung, heart, and gut), and compartments 3 and 6 represent the *slow* tissue (muscle, skin, and brain). *Fast* and *slow* indicate how quickly the hormones are synthesized via transfer rates K_{i0}. Transfer rates K_{41} and K_{63} are used to

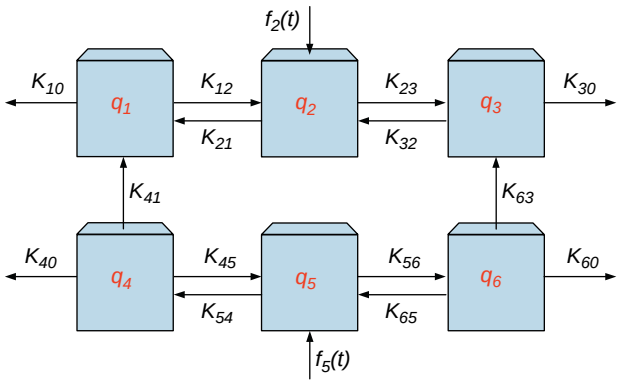


FIGURE 7.32 Six-compartment model that describes thyroid hormone distribution and metabolism.

[13]See DiStefano and Mori, 1977, in references for original problem development.

model the transformation of T4 into T3. Input to the system is through the plasma in either compartment q_2 or q_5.

Consider the movement of a drug in the body with the pharmacokinetic model in Figure 7.10. After ingestion, the drug moves into the blood, where it is distributed in the plasma. Drug distribution in the plasma is among water and proteins. Since drugs are relatively small molecules, they easily move through the capillaries and into most fluids and organs of the body. In addition, drugs move easily into the intracellular fluids of body tissues. Each arrow in Figure 7.10 needs to be defined with a transfer rate. Obtaining values for the transfer rates is usually very difficult or even impossible.

EXAMPLE PROBLEM 7.18

Consider the model[14] illustrated in Figure 7.33 for the oral input of the thyroid hormone replacement therapy in which the body does not produce any thyroid hormone. While the model is appropriate for either T3 or T4, here we track T3 and ignore T4 for simplicity. Assume T3 exists in compartment 1 (gut) in solid form and in compartment 2 (still the gut) in liquid form. Compartment 3 represents the plasma, and compartments 4 and 5 are the slow and fast tissues, respectively. Assume that the input is bolus, $f_1(t) = 25\delta(t)$, and that the initial conditions are zero. Further, $K_{12} = 1.1$, $K_{20} = 0.01$, $K_{23} = 0.9$, $K_{34} = 15$, $K_{43} = 30$, $K_{40} = 1.0$, $K_{35} = 0.5$, $K_{53} = 0.4$, and $K_{50} = 0.05$. Note that K_{40} and K_{50} represent T3 metabolism. Solve for the quantity in compartment 3.

Solution

The input is transformed into a new initial condition, $q_1(0)$. The conservation of mass for each compartment yields

$$\dot{q}_1 = -K_{12}q_1 = -1.1q_1 \tag{7.113}$$

$$\dot{q}_2 = K_{12}q_1 - (K_{20} + K_{23})q_2 = 1.1q_1 - 0.91q_2 \tag{7.114}$$

$$\dot{q}_3 = K_{23}q_2 - (K_{34} + K_{35})q_3 + K_{43}q_4 + K_{53}q_5 = 0.9q_2 - 15.5q_3 + 30q_4 + 0.4q_5 \tag{7.115}$$

$$\dot{q}_4 = K_{34}q_3 - (K_{40} + K_{43})q_4 = 15q_3 - 31q_4 \tag{7.116}$$

$$\dot{q}_5 = K_{35}q_3 - (K_{50} + K_{53})q_5 = 0.5q_3 - 0.45q_5 \tag{7.117}$$

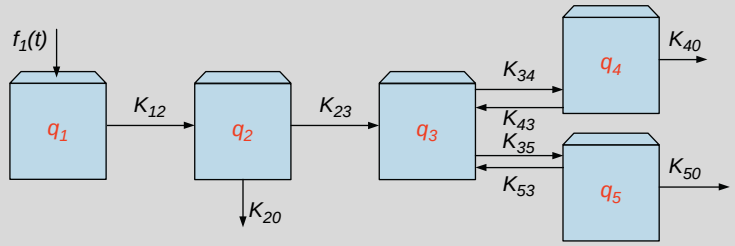


FIGURE 7.33 Illustration for Example Problem 7.18. Compartments 1 and 2 are the digestive system, compartment 3 is the plasma, and compartments 4 and 5 are the fast and slow tissues, respectively.

[14]See DiStefano and Mak, 1979, in references for original problem development.

Since the conservation of mass equation for q_1 involves only q_1, it is easily solved as $q_1 = 25e^{-1.1t}u(t)$. Substituting the solution for q_1 into Eq. (7.114), gives

$$\dot{q}_2 = 27.5e^{-1.1t} - 0.91q_2 \tag{7.118}$$

Equation (7.118) involves an input and only q_2, which is solved independently as

$$q_2 = \frac{K_{12}q_1(0)}{K_{12} - (K_{20} + K_{23})}\left(e^{-(K_{20}+K_{23})t} - e^{-K_{12}t}\right) = 144.74\left(e^{-0.91t} - e^{-1.1t}\right)u(t) \tag{7.119}$$

The solution for q_2 is substituted into Eq. (7.115), yielding

$$\begin{aligned}\dot{q}_3 &= K_{23}\frac{K_{12}q_1(0)}{K_{12} - (K_{20} + K_{23})}\left(e^{-(K_{20}+K_{23})t} - e^{-K_{12}t}\right) - 15.5q_3 + 30q_4 + 0.4q_5 \\ &= 130.27\left(e^{-0.91t} - e^{-1.1t}\right) - 15.5q_3 + 30q_4 + 0.4q_5\end{aligned} \tag{7.120}$$

Equations (7.120), (7.116), and (7.117) are solved using MATLAB and the D-Operator, as follows:

```
>> syms D
>> A=[-15.5 30 0.4;15 -31 0;0.5 0 -0.45];
>> det(D*eye(3)-A)
>> adj=det(D*eye(3)-A)*inv(D*eye(3)-A)
```

which gives the reconstructed differential equations for q_3 as

$$\dddot{q}_3 + 46.95\ddot{q}_3 + 51.225\dot{q}_3 + 7.525q_3 = 2531.8e^{-1.1t} - 1803.12e^{-0.91t} \tag{7.121}$$

with roots -45.84, -0.94, and -0.18. The natural response is

$$q_{3_n} = B_1e^{-45.84t} + B_2e^{-0.94t} + B_3e^{-0.18t}$$

The forced response is $q_{3_f} = B_4e^{-1.1t} + B_5e^{-0.91t}$, which after substituting into Eq. (7.113) gives

$$B_4 = 380.39 \text{ and } B_5 = 1870.41$$

The complete response is then

$$q_3 = B_1e^{-45.84t} + B_2e^{-0.94t} + B_3e^{-0.18t} + 380.39e^{-1.1t} + 1870.41e^{-0.91t}$$

We use the initial conditions, $q_3(0) = 0$, $\dot{q}_3(0) = 0$, and $\ddot{q}_3(0) = 24.75$, to solve for B_1, B_2, and B_3 as follows:

$$q_3(0) = 0 = B_1 + B_2 + B_3 + 380.39 + 1870.41$$

and with $\dot{q}_3 = -45.84B_1e^{-45.84t} - 0.94B_2e^{-0.94t} - 0.18B_3e^{-0.18t} - 418.39e^{-1.1t} - 1701.97e^{-0.91t}$ we have

$$\dot{q}_3(0) = 0 = -45.84B_1 - 0.94B_2 - 0.18B_3 - 418.39 - 1701.97$$

Continued

which with $\ddot{q}_3 = 2101B_1e^{-45.84t} + 0.88B_2e^{-0.94t} + 0.031B_3e^{-0.18t} + 460.23e^{-1.1t} + 1548.80e^{-0.91t}$ gives

$$\ddot{q}_3(0) = 24.75 = 2101B_1 + 0.88B_2 + 0.031B_3 + 460.23 + 1548.80$$

To solve for the unknown constants, we evaluate

$$\begin{bmatrix} 1 & 1 & 1 \\ -45.84 & -0.94 & -0.18 \\ 2101 & 0.88 & 0.031 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = \begin{bmatrix} -2250.7 \\ 2120.3 \\ -1984.3 \end{bmatrix}$$

which gives

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = \begin{bmatrix} .004 \\ -2259.51 \\ 8.844 \end{bmatrix}$$

Thus,

$$q_3 = \left(0.004e^{-45.84t} - 2259.51e^{-0.94t} + 8.84e^{-0.18t} + 380.39e^{-1.1t} + 1870.41e^{-0.91t}\right)u(t) \tag{7.122}$$

which is plotted in Figure 7.34. Note that if the oral dose involved T4 instead of T3, the model would need to be changed by adding three more compartments for T4 (lower part of Figure 7.31). We still need the three T3 compartments, since T4 transforms into T3.

Another way to solve for the response is to directly analyze the system using the D-Operator matrix approach on Eqs. (7.113)–(7.117), which appears easier, since the input is 0 and there is no forced response. However, considerable additional work is required to calculate the two extra initial conditions ($\dddot{q}_1(0)$ and $\ddddot{q}_1(0)$) needed to solve for the extra two terms in the natural response, which is not trivial. Thus, we have

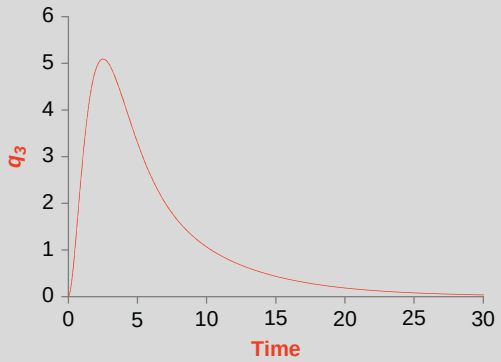


FIGURE 7.34 Illustration of the quantity in compartment 3 in Example Problem 7.18.

```
>> syms D
>> A=[-1.1 0 0 0 0;1.1 -.91 0 0 0;0.9 -15 .5 30 .4;0 0 15 -31 0;0 0 .5 0 -.45];
>> det(D*eye(5)-A)
ans =
  D^5 + (1224*D^4)/25 + (293191*D^3)/2000 + (787421*D^2)/5000 +
  (2656059*D)/40000 + 301301/40000
>> eig(A)
ans =
  -0.1748
  -0.9392
  -45.8360
  -0.9100
  -1.1000
```

The reconstructed differential equation for q_3 is

$$\dddot{\dddot{q}}_3 + 48.92\dddot{q}_3 + 146.6\ddot{q}_3 + 157.5\ddot{q}_3 + 66.4\dot{q}_3 + 7.53q_3 = 0$$

From the roots, the response is written as

$$q_3 = B_1e^{-0.18t} + B_2e^{-0.94t} + B_3e^{-45.84t} + B_4e^{-0.91t} + B_5e^{-1.1t} \tag{7.123}$$

which is the same form as Eq. (7.122). To calculate B_1 through B_5, we use the initial conditions, $q_1(0) = 25$, $q_2(0) = 0$, $q_3(0) = 0$, $q_4(0) = 0$, and $q_5(0) = 0$, to find, after considerable effort, that $q_3(0) = 0, \dot{q}_3(0) = 0, \ddot{q}_3(0) = 24.75$, $\dddot{q}_3(0) = -433.3752$, and $\dddot{\dddot{q}}_1(0) = 17,935$.

Using the initial conditions and Eq. (7.123), we have

$$q_3(0) = 0 = B_1 + B_2 + B_3 + B_4 + B_5$$

$$\dot{q}_3(0) = 0 = -0.18B_1 - 0.94B_2 - 45.84B_3 - 0.91B_4 - 1.1B_5$$

$$\ddot{q}_3(0) = 24.75 = (-0.18)^2B_1 + (-0.94)^2B_2 + (-45.84)^2B_3 + (-0.91)^2B_4 + (-1.1)^2B_5$$

$$\dddot{q}_3(0) = -433.3725 = (-0.18)^3B_1 + (-0.94)^3B_2 + (-45.84)^3B_3 + (-0.91)^3B_4 + (-1.1)^3B_5$$

$$\dddot{\dddot{q}}_3(0) = 17,935 = (-0.18)^4B_1 + (-0.94)^4B_2 + (-45.84)^4B_3 + (-0.91)^4B_4 + (-1.1)^4B_5$$

To solve for the unknown constants, we evaluate the unknown coefficients using MatLab:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -0.18 & -0.94 & -45.84 & -0.91 & -1.1 \\ 0.031 & 0.88 & 2101 & 0.83 & 1.21 \\ -0.005 & -0.83 & -96,298 & -.75 & -1.33 \\ 0.0009 & 0.78 & 4,4139,948 & 0.69 & 1.46 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 24.75 \\ -433.3725 \\ 17,934 \end{bmatrix}$$

Continued

which gives

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \end{bmatrix} = \begin{bmatrix} 8.84 \\ -2259.15 \\ 0.004 \\ 1870 \\ 380.3 \end{bmatrix}$$

and

$$q_3 = \left(0.004e^{-45.84t} - 2259.15e^{-0.94t} + 8.84e^{-0.18t} + 380.3e^{-1.1t} + 1870e^{-0.91t}\right)u(t)$$

7.9 EXERCISES

1. Determine the number of Na^+ and K^+ ions inside a cell with a volume of 1 nL and concentrations given in Table 7.2.
2. Suppose the concentrations of Na^+, Cl^-, and K^+ are 20, 52, and 158 mM/L, respectively. Determine the number of ions in a cell of 2 nL.
3. A cell with a volume of 1.5 nL contains 2×10^{14} molecules of K^+ and 1.5×10^{13} molecules of Na^+. What is the concentration for each ion?
4. Two compartments, with volumes V_1 and V_2, are separated by a thin membrane, and solute moves from one compartment to the other by diffusion. If an amount ζ of solute is dumped into compartment 1 at $t = 0$, then find the concentration in each compartment.
5. Two compartments, with volumes V_1 and V_2, are separated by a thin membrane, and solute moves from one compartment to the other by diffusion. If an amount ζ of solute is dumped into compartment 2 at $t = 0$, then find the concentration in each compartment.
6. Two compartments, with equal volumes of 0.0572 cm^3, are separated by a thin membrane, and solute moves from one compartment to the other by diffusion. If all of the solute is initially dumped into one compartment, then find the transfer rate if the time constant equals $27 \times 10^3 s^{-1}$.
7. A system is given by two compartments separated by a thin membrane, and solute moves from one compartment to the other by diffusion. The volume of compartment 1 is $0.1 cm^3$ and compartment 2 is 0.3 cm^3. The transfer rate is $2.0 \times 10^{-3} s^{-1}$. Suppose 3 g of solute is dumped into compartment 2. Solve for the concentration in both compartments.
8. A system is given by two compartments separated by a thin membrane, and solute moves from one compartment to the other by diffusion. Suppose the volume of compartment 1 equals 0.0572 cm^3 and is twice as large as compartment 2. If 100 g of solute is dumped into compartment 2, then solve for the concentration in both compartments for an arbitrary transfer rate K (where the solution is expressed in terms of K).
9. A system is given by two compartments separated by a thin membrane, and solute moves from one compartment to the other by diffusion. The volume of compartment 1 is 0.0572 cm^3 and compartment 2 is 0.0286 cm^3. Suppose 100 M of solute is dumped into compartment 2, and the concentration response for compartment 2 is $c_2(t) = \left(1165.67 + 2331.3e^{-0.015t}\right)u(t)\frac{M}{cm^3}$. (a) Find the transfer rate. (b) Solve for the concentration in compartment 1.

10. A system is given by two compartments separated by a thin membrane, and solute moves from one compartment to the other by diffusion. The volume of compartment 1 is 0.03 cm^3 and compartment 2 is 0.01 cm^3. Suppose 50 M of solute is dumped into compartment 2, and the concentration response for compartment 1 is $c_1(t) = 1250\left(1 - e^{-0.01t}\right)u(t)\frac{M}{cm^3}$. (a) Find the transfer rate. (b) Solve for the concentration in compartment 2.

11. Find the initial osmotic pressure at room temperature for a cell if the only ions present are *KCl* on either side of the membrane. Assume the concentrations for K^+ and Cl^- from Table 7.2 and that the ions cannot cross the membrane. The cell volume is 2 nL. Determine the final cell volume.

12. Find the initial osmotic pressure at room temperature for a cell if the only ions present are $CaCl_2$ on either side of the membrane. Assume the concentrations for Ca^{+2} and Cl^- from Table 7.2 and that the ions cannot cross the membrane. The cell volume is 2 nL. Determine the final cell volume.

13. Find the initial osmotic pressure at room temperature for a cell if all the ions present are listed in Table 7.2. Assume that the ions cannot cross the membrane. The cell volume is 2 nL. Determine the final cell volume.

14. Find the initial osmotic pressure at room temperature for a cell if the only ions present are *KCl* and *NaCl* on either side of the membrane. Assume the concentrations for K^+, Na^+, and Cl^- from Table 7.2, and that only K^+ can cross the membrane. The cell volume is 2 nL. Describe what happens to the ions. Determine the final cell volume.

15. Find the initial osmotic pressure at room temperature for a cell if the only ions present are *KCl* and *NaCl* on either side of the membrane, and 0.2×10^{-9} M of a protein inside the cell. Assume the concentrations for K^+, Na^+, and Cl^- from Table 7.2, and that only K^+ can cross the membrane. The cell volume is 2 nL. Describe what happens to the ions. Determine the final cell volume.

16. Find the osmolarity and osmotic pressure of 2 mM Na_2SO_4 at room temperature.

17. Find the osmolarity and osmotic pressure of a 9% solution of *NaCl* at room temperature.

18. Find the osmotic pressure at room temperature for a cell if the ions in Table 7.2 are present.

19. Consider a cell with an internal osmolarity of 300 mOsm and volume of 2 nL in a 30 nL solution of 300 mOsm. A 3 nL, 5% *NaCl* by weight solution is added to the extracellular space. Assuming that *NaCl* is impermeable and that the moles inside the cell do not change, describe the events that take place until steady state is achieved. What is the final osmolarity of the cell? What is the volume of the cell at steady state?

20. Consider a cell with an internal osmolarity of 300 mOsm and volume of 2 nL in a 30 nL solution of 300 mOsm. Three mM of $CaCl_2$ is added to the extracellular space. Assuming that $CaCl_2$ is impermeable and that the moles inside the cell do not change, describe the events that take place until steady state is achieved. What is the final osmolarity of the cell? What is the volume of the cell at steady state?

21. Consider a cell with an internal osmolarity of 300 mOsm and volume of 2 nL in a 30 nL solution of 300 mOsm. Five mM of urea is added to the extracellular space. Assuming that urea is permeable and that the moles originally inside the cell are impermeable, describe the events that take place until steady state is achieved.

Continued

22. Given the cell described in Figure 7.7 with $a = 500$ mM, $P_K = 1.0$, and $P_{Na} = 0.04$, at steady state, plot the relationship between $\vec{J}_p$ and V.

23. Suppose 500 mg of dye was introduced into the plasma compartment. After reaching steady state, the concentration in the blood is $0.0893 \frac{\text{mg}}{\text{cm}^3}$. Find the volume of the plasma compartment.

24. Suppose 1 g bolus of solute is injected into a plasma compartment of 3 L. The transfer rate out of the compartment equals $0.7\ \text{hr}^{-1}$. Solve for the solute concentration. What is the half-life of the solute in the plasma compartment?

25. An unknown quantity of radioactive iodine (I^{131}) is instantaneously passed into the plasma. The time dependence of the quantity of I^{131} in the plasma exhibits an exponential decay from 100 mg with a time constant of 1 day, while the urine shows an exponential rise from zero to 75 mg with a time constant of 1 day. Assuming the compartment model in Example Problem 7.5, determine the transfer rates and the half-life.

26. Suppose a patient ingested a small quantity of radioactive Iodine (I^{131}). A simple model describing the removal of I^{131} from the bloodstream into the urine and thyroid is given in Example Problem 7.5. (a) Sketch the response of the system. (b) Suppose the thyroid is not functioning and does not take up any I^{131}. Sketch the response of the abnormal system and compare to the result from (a).

27. A radioactive bolus of I^{131} is injected into a plasma compartment. The time dependence of the concentration of I^{131} in the plasma is $c_1 = 143e^{-1.6t}\frac{\text{mg}}{100\ \text{mL}}$. The amount of I^{131} is 10 K mg. Assuming the compartmental model in Example Problem 7.5, find (a) the volume of the plasma compartment, (b) $K = K_1 + K_2$, and (c) the half-life.

28. Find the half-life for the model given in Eq. (7.33) and Figure 7.8.

29. Use SIMULINK to simulate the model given in Eq. (7.33) and Figure 7.8. Use the parameters given in Figure 7.11.

30. Demonstrate for the one-compartment pharmacokinetic model given in Section 7.5.3 with Eq. (7.33) and Figure 7.8 that as γ increases, both $t_{\max}$ and $q_1(t_{\max})$ decrease.

31. An antibiotic is exponentially administered into the body, with $f(t) = 75e^{-2t}u(t)$. Assume the model given in Figure 7.8 with $K_{10} = 0.3$. (a) Analytically solve for the quantity of the antibiotic in the plasma. (b) Simulate the quantity of the antibiotic in the plasma using SIMULINK. (c) What is the time to maximum concentration, and what is the quantity in the plasma at that time?

32. For the one-compartment repeat dosage in Section 7.5.4, derive Eq. (7.42) from (7.41) and Eq. (7.44) from (7.33).

33. A 2 g bolus of antibiotic is administered to a person with a plasma volume of 3 L. The average impulse response for this drug is shown in Figure 7.35. Assuming a one-compartment model, determine the transfer rate. If the concentration of the drug is not to fall below 10 percent of the initial dosage at steady state, how often does the drug need to be given to maintain this minimum level?

34. A 4 g bolus of antibiotic is administered to a person with a plasma volume of 3 L. The average washout response for this drug in a plasma volume of 3 L is shown in Figure 7.36. Assuming a one-compartment model, determine the transfer rate. If the concentration of the drug is not to fall below 25 percent of the initial dosage at steady state, how often does the drug need to be given to maintain this minimum level?

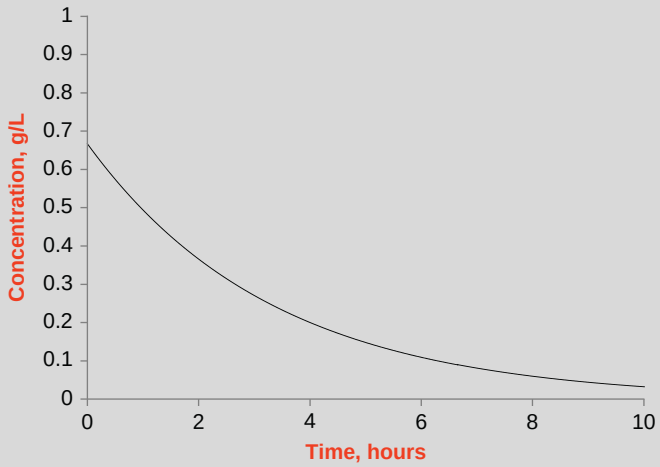


FIGURE 7.35 Illustration for Exercise 33.

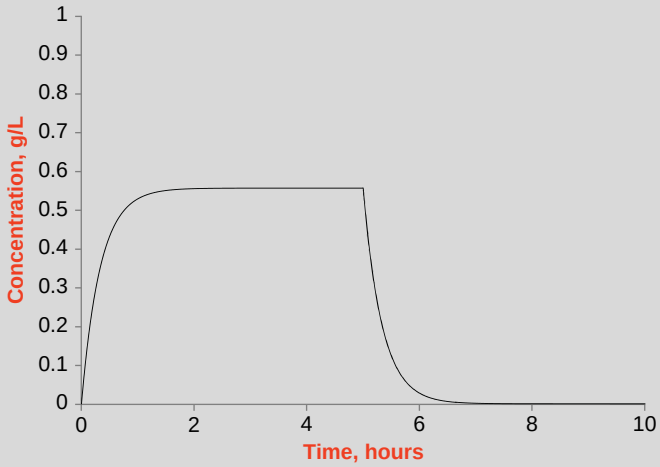


FIGURE 7.36 Illustration for Exercise 34.

35. Given the two-compartment model in Figure 7.16 and $q_1(0) = 0$ and $q_2(0) = \zeta$, solve for the concentration in compartment 2.

36. Given the two-compartment model in Figure 7.16 and $q_1(0) = \alpha$ and $q_2(0) = \zeta$, solve for the concentration in each compartment.

37. Given the two-compartment model shown in Figure 7.17 with a pulse ingestion of solute in the digestive system and removal of solute via metabolism and excretion in urine, solve for the plasma concentration.

Continued

38. Given the two-compartment model shown in Figure 7.17 with a $\zeta\delta(t) + (1-\zeta)(u(t) - u(t-t_0))$ ingestion of solute in the digestive system and removal of solute via metabolism and excretion in urine, solve for the plasma concentration.

39. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0, K_{10} = 0.5,\ K_{21} = 0.3,$ $K_{20} = 0.9,\ f_1(t) = 0,$ and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. (a) Solve for the quantity in each compartment. (b) Find the maximum amount of solute in compartment 1.

40. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.2, K_{10} = 0.7,$ $K_{21} = 0,\ K_{20} = 1,\ f_1(t) = 2u(t),$ and $f_2(t) = 0$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

41. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0, K_{10} = 0.6,$ $K_{21} = 0.1,\ K_{20} = 0.8,\ f_1(t) = 0,$ and $f_2(t) = \delta(t) + 5u(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

42. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.3, K_{10} = 0.2,$ $K_{21} = 0,\ K_{20} = 0.4,\ f_1(t) = 4u(t),$ and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

43. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.3, K_{10} = 0.5,$ $K_{21} = 0, K_{20} = 1,\ f_1(t) = 2u(t),$ and $f_2(t) = 3\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

44. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0, K_{10} = 0.6,$ $K_{21} = 0.1,\ K_{20} = 0.5,\ f_1(t) = 0,$ and $f_2(t) = \delta(t) + 5u(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

45. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.3, K_{10} = 0.7,$ $K_{21} = 0, K_{20} = 0.1,\ f_1(t) = 3\delta(t),$ and $f_2(t) = 0$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

46. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.4, K_{10} = 1.0,$ $K_{21} = 0, K_{20} = 0.3,\ f_1(t) = 2u(t),$ and $f_2(t) = 0$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

47. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.2, K_{10} = 0.8,$ $K_{21} = 0, K_{20} = 0.3,\ f_1(t) = \delta(t) + 2u(t),$ and $f_2(t) = 0$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

48. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.3, K_{10} = 0.5,$ $K_{21} = 0.1, K_{20} = 0.4,\ f_1(t) = 2u(t),$ and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

49. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.3, K_{10} = 0.7,$ $K_{21} = 3, K_{20} = 1,\ f_1(t) = 2\delta(t),$ and $f_2(t) = 5u(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

50. Consider the two-compartment model shown in Figure 7.15 with $K_{12} = 0.2, K_{10} = 0.6,$ $K_{21} = 0.1, K_{20} = 0.5,\ f_1(t) = 3\delta(t),$ and $f_2(t) = \delta(t) + 5u(t)$. Assume that the initial conditions are zero. Solve for the quantity in each compartment.

51. Suppose 1 g of solute is ingested into the digestive system that has a transfer rate of $1.4\ \text{hr}^{-1}$ into the plasma. The plasma compartment is 3 L and has transfer rate of $0.7\ \text{hr}^{-1}$ into the environment. (a) Solve for the solute concentration in the plasma. (b) When is the maximum solute concentration observed in the plasma compartment? (c) What is the maximum solute in the plasma compartment?

52. Consider the model in Figure 7.37. The time dependence of the concentration of a radioactively labeled solute in the plasma is

$$c_1(t) = 143e^{-1.6t} + 57e^{-2.8t}\,\frac{\text{mg}}{100\text{ mL}}$$

after injecting a bolus of 10 g into the plasma. (a) Find the volume of the plasma compartment. (b) Find the transfer rates K_{12}, K_{21}, and K_{13}. (c) Suppose the input is changed to $5u(t)\,\frac{\text{mg}}{100\text{ mL}}$, and solve for $c_1(t)$ and $c_2(t)$. (d) Suppose the input is changed to $5(u(t) - u(t-2))\,\frac{\text{mg}}{100\text{ mL}}$, and solve for $c_1(t)$ and $c_2(t)$.

53. A unit step input is applied to the compartmental system in Figure 7.38. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 2$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

54. A unit step input is applied to the compartmental system in Figure 7.38. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 2$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

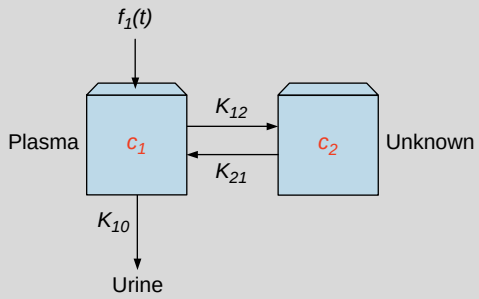


FIGURE 7.37 Illustration for Exercise 52.

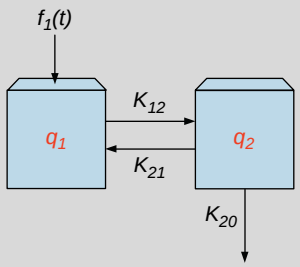


FIGURE 7.38 Illustration for Exercises 53–68.

Continued

55. The input to the compartmental system in Figure 7.38 is $2u(t) - 2u(t-1)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 2$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

56. The input to the compartmental system in Figure 7.38 is $2u(t-1)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

57. The input to the compartmental system in Figure 7.38 is $2e^{-0.5562t}u(t)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

58. The input to the compartmental system in Figure 7.38 is $2e^{-0.5562t}u(t)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 1$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

59. The input to the compartmental system in Figure 7.38 is $2e^{-0.1438t}u(t)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

60. The input to the compartmental system in Figure 7.38 is $2e^{-0.1438t}u(t)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 2$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

61. The input to the compartmental system in Figure 7.38 is $2e^{-0.1438t}u(t) - 2e^{-0.1438(t-10)t}u(t-10)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

62. The input to the compartmental system in Figure 7.38 is $3e^{-t}u(t)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 2$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

63. The input to the compartmental system in Figure 7.38 is $3e^{-t}u(t) - 3e^{-(t-3)}u(t-3)$. The transfer rates are $K_{20} = 0.2, K_{21} = 0.1$, and $K_{12} = 0.4$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1;

(b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

64. The input to the compartmental system in Figure 7.38 is $0.5e^{-2t}u(t)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 4$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

65. The input to the compartmental system in Figure 7.38 is $0.5e^{-2(t-2)}u(t-2)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

66. The input to the compartmental system in Figure 7.38 is $0.5e^{-2t}u(t) - 0.5e^{-2(t-1.5)}u(t-1.5)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

67. The input to the compartmental system in Figure 7.38 is $3\cos 4tu(t)$. The transfer rates are $K_{20} = 0.3, K_{21} = 1.0$, and $K_{12} = 0.6$. The initial conditions are $q_1(0) = 0$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

68. The input to the compartmental system in Figure 7.38 is $3\sin 2tu(t)$. The transfer rates are $K_{20} = 3, K_{21} = 5$, and $K_{12} = 7$. The initial conditions are $q_1(0) = 1$ and $q_2(0) = 0$. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2. (e) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1 and q_2.

69. For the compartmental system in Figure 7.39, a bolus of solute $(f_3(t) = \delta(t))$ is ingested into the digestive system (compartment 3). Assume that the initial conditions are zero. Write a single differential equation involving only variable (a) q_1; (b) q_2. For $t > 0$, solve the system for (c) q_1; (d) q_2.

70. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.4, K_{10} = 0.5,\ K_{21} = 0.6, K_{31} = 0.9,\ K_{32} = 0.7,\ K_{23} = 0.2,\ K_{13} = 0.8,\ f_1(t) = 3u(t)$, and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

71. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.5, K_{10} = 0.3,\ K_{21} = 0.6, K_{31} = 0.9,\ K_{32} = 0.7,\ K_{23} = 0.2,\ K_{13} = 0.8$, and $f_3(t) = 3\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

Continued

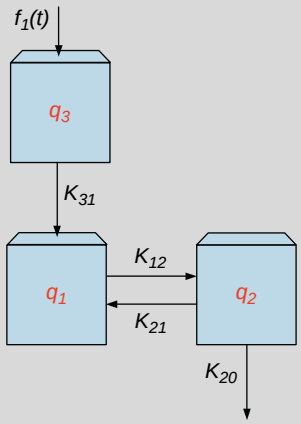


FIGURE 7.39 Illustration for Exercise 69.

72. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.4, K_{10} = 0.5,\ K_{21} = 0.6, K_{31} = 0.9,\ K_{32} = 0.7,\ K_{23} = 0.2,\ K_{13} = 0.8,$ and $f_2(t) = 5u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

73. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.6, K_{20} = 0.2,\ K_{21} = 0.3, K_{31} = 0.5,\ K_{32} = 0.6,\ K_{23} = 0.4,\ K_{13} = 0.8,$ and $f_1(t) = 2u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

74. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.3, K_{20} = 0.4,\ K_{21} = 0.8, K_{31} = 0.5,\ K_{32} = 0.3,\ K_{23} = 0.4,\ K_{13} = 0.8,\ f_1(t) = 3\delta(t),$ and $f_3(t) = 3u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

75. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.6, K_{20} = 0.2,\ K_{21} = 0.3, K_{31} = 0.5,\ K_{32} = 0.6,\ K_{23} = 0.4,\ K_{13} = 0.8,$ and $f_2(t) = \delta(t) + 5u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

76. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.3, K_{10} = 0.5, K_{21} = 0.2, K_{23} = 0.4, K_{32} = 0.6,$ and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the

input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

77. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.4, K_{30} = 0.5, K_{21} = 0.7, K_{23} = 0.8, K_{32} = 0.2,$ and $f_3(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

78. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.3, K_{20} = 0.2, K_{21} = 0.3, K_{23} = 0.4, K_{32} = 0.6,$ and $f_1(t) = 2u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

79. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.7, K_{10} = 0.3, K_{21} = 0.4, K_{23} = 0.5, K_{32} = 0.6,$ and $f_2(t) = \delta(t) + 5u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

80. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.1, K_{10} = 0.2, K_{23} = 4.0, K_{31} = 0.4,$ and $f_3(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

81. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.3, K_{20} = 0.2, K_{23} = 2.0, K_{31} = 0.6,$ and $f_2(t) = 4\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

82. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.4, K_{10} = 0.2, K_{23} = 5.0, K_{31} = 1.0,$ and $f_3(t) = 2u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

83. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.6, K_{30} = 0.2, K_{23} = 5.0, K_{31} = 1.0,$ and $f_1(t) = u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and

Continued

only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

84. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.3, K_{10} = 0.1, K_{23} = 0.4, K_{31} = 0.6,$ and $f_1(t) = 3\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

85. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.3, K_{20} = 0.1, K_{23} = 0.2, K_{31} = 0.4,$ and $f_2(t) = 4\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

86. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.4, K_{10} = 0.2, K_{23} = 0.5, K_{31} = 1.0,$ and $f_3(t) = 8u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

87. Consider the unilateral three-compartment model shown in Figure 7.22 with nonzero parameters and inputs $K_{12} = 0.6, K_{30} = 0.2, K_{23} = 0.8, K_{31} = 0.3,$ and $f_1(t) = 4u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

88. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.3, K_{10} = 0.5, K_{21} = 0.2, K_{23} = 0.4,$ and $f_2(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

89. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{30} = 0.5, K_{21} = 0.7, K_{23} = 0.8, K_{32} = 0.2,$ and $f_3(t) = 5\delta(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

90. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.3$, $K_{20} = 0.2$, $K_{23} = 0.4$, $K_{32} = 0.6$, and $f_1(t) = 2u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

91. Consider the mammillary three-compartment model shown in Figure 7.21 with nonzero parameters and inputs $K_{12} = 0.7$, $K_{10} = 0.3$, $K_{21} = 0.4$, $K_{32} = 0.6$, and $f_3(t) = 3\delta(t) + 5u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the

input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

92. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.4$, $K_{10} = 0.5$, $K_{21} = 0.6$, $K_{31} = 0.9$, $K_{21} = 0.4$, $K_{32} = 0.7$, $K_{23} = 0.2$, $K_{13} = 0.8$, and $f_2(t) = 3e^{-t}\,u(t)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$,solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

93. Consider the three-compartment model shown in Figure 7.20 with nonzero parameters and inputs $K_{12} = 0.4$, $K_{10} = 0.5$, $K_{21} = 0.6$, $K_{31} = 0.9$, $K_{21} = 0.4$, $K_{32} = 0.7$, $K_{23} = 0.2$, $K_{13} = 0.8$, and $f_2(t) = 3e^{-(t-1)}\,u(t–1)$. Assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3. For $t > 0$, solve the system for (d) q_1; (e) q_2; (f) q_3. (g) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, and q_3.

94. Consider the following three-compartment model in Figure 7.40. A 5 g radioactively labeled bolus is injected into compartment 2. The time dependence of solute concentration in compartment 2 is

$$c_2(t) = -6.6271e^{-3.1069t} + 106.6271e^{-0.1931t}\,\frac{\text{mg}}{100\text{ mL}}$$

(a) What is the volume of compartment 2? (b) Determine the transfer rates K_{21}, K_{23}, and K_{32}.

95. Given a mammillary four-compartment model as described in Figure 7.26, with nonzero parameters and inputs $K_{12} = 0.3$, $K_{10} = 0.2$, $K_{21} = 0.4$, $K_{31} = 0.8$, $K_{13} = 0.7$, $K_{14} = 0.2$, $K_{41} = 0.5$, and $f_2(t) = 5\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

96. Given a mammillary four-compartment model as described in Figure 7.26, with nonzero parameters and inputs $K_{12} = 0.5$, $K_{10} = 0.1$, $K_{21} = 0.3$, $K_{20} = 0.3$, $K_{31} = 0.2$, $K_{13} = 0.5$, $K_{14} = 0.7$, $K_{41} = 0.2$, and $f_1(t) = 5u(t)$, assume that the initial conditions are zero. Write a single differential

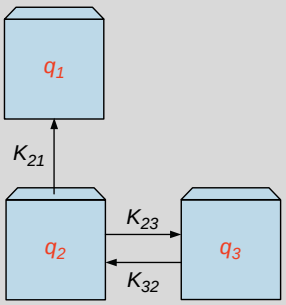


FIGURE 7.40 Illustration for Exercise 94.

Continued

equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$,solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

97. Given a catenary four-compartment model as described in Figure 7.27, with nonzero parameters and inputs $K_{12} = 0.3$, $K_{10} = 0.1$, $K_{21} = 0.5$, $K_{30} = 0.4$; $K_{32} = 0.6$, $K_{23} = 0.4$, $K_{34} = 0.2$, $K_{43} = 0.7$, and $f_1(t) = 10\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

98. Given a catenary four-compartment model as described in Figure 7.27, with nonzero parameters and inputs $K_{12} = 0.7$, $K_{10} = 0.2$, $K_{21} = 0.4$, $K_{32} = 0.2$, $K_{23} = 0.7$, $K_{34} = 0.3$, $K_{43} = 0.5$, and $f_3(t) = 20u(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

99. Given a unilateral four-compartment model as described in Figure 7.28, with nonzero parameters and inputs $K_{12} = 0.4$, $K_{10} = 0.1$, $K_{23} = 0.6$, $K_{34} = 0.7$, $K_{41} = 0.4$, $K_{40} = 0.2$, and $f_3(t) = 20\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

100. Given a unilateral four-compartment model as described in Figure 7.28, with nonzero parameters and inputs $K_{12} = 0.4$, $K_{10} = 0.1$, $K_{23} = 0.6$, $K_{34} = 0.7$, $K_{41} = 0.4$, $K_{40} = 0.2$, and $f_3(t) = 20\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

101. Given a unilateral four-compartment model as described in Figure 7.28, with nonzero parameters and inputs $K_{12} = 0.4$, $K_{23} = 0.4$, $K_{34} = 0.4$, $K_{41} = 0.4$, and $f_3(t) = 10\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

102. Given a unilateral five-compartment model as described in Figure 7.28, with nonzero parameters and inputs $K_{12} = 0.5$, $K_{23} = 0.5$, $K_{34} = 0.5$, $K_{41} = 0.5$, $K_{51} = 0.5$, $K_{40} = 0.1$, and $f_2(t) = 10\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

103. Given a unilateral five-compartment model as described in Figure 7.28, with nonzero parameters and inputs $K_{12} = 0.5$, $K_{23} = 0.5$, $K_{34} = 0.5$, $K_{41} = 0.5$, $K_{51} = 0.5$, and $f_1(t) = 5\delta(t)$, assume that the initial conditions are zero. Write a single differential equation involving the input and only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2;

(g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph the quantity in each compartment.

104. Suppose 1 g of solute is dumped into compartment 1 as shown in Figure 7.41. The transfer rates are $K_{12} = 0.4$, $K_{23} = 0.6$, $K_{24} = 0.3$, $K_{32} = 1.2$, $K_{34} = 0.8$, and $K_{42} = 0.7$. Write a single differential equation involving only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and each variable.

105. For the compartmental system in Figure 7.42, a radioactively labeled bolus of solute, with magnitude of 1, was injected into compartment 3. Let $K_{21} = 0.2$, $K_{32} = 0.3$, $K_{31} = 0.7$, $K_{13} = 0.4$, $K_{34} = 0.9$, $K_{43} = 0.1$, and $K_{14} = 0.6$. Write a single differential equation involving only variable (a) q_1; (b) q_2; (c) q_3; (d) q_4. For $t > 0$, solve the system for (e) q_1; (f) q_2; (g) q_3; (h) q_4. (i) Using SIMULINK, simulate the system from the original set of differential equations and graph q_1, q_2, q_3, and q_4.

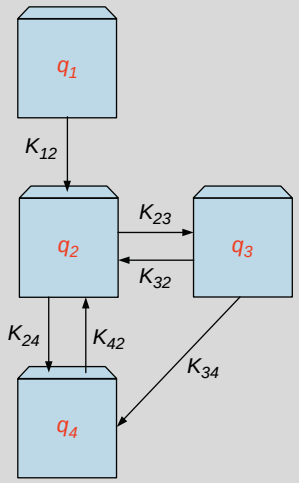


FIGURE 7.41 Illustration for Exercise 104.

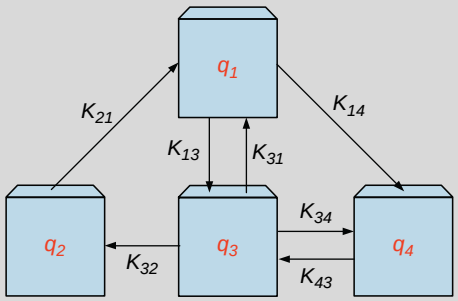


FIGURE 7.42 Illustration for Exercise 105.

Continued

106. Consider the model illustrated in Figure 7.33 for the oral input of T3 thyroid hormone replacement therapy. Assume T3 exists in compartment 1 (gut) in solid form and in compartment 2 (still the gut) in liquid form. Compartment 3 represents the plasma, and compartments 4 and 5 are the fast and slow tissues, respectively. Assume that the input is bolus, $f_1(t) = 5\delta(t)$, and that the initial conditions are zero. Further, assume that $K_{12} = 1.1$, $K_{20} = 0.01$, $K_{23} = 0.9$, $K_{34} = 7.0$, $K_{43} = 30$, $K_{40} = 0.8$, $K_{35} = 2.0$, $K_{53} = 0.3$, and $K_{50} = 0.1$. Note that K_{40} and K_{50} represent T3 metabolism. Solve for the quantity of T3 in compartment 3.

107. Consider the model illustrated in Figure 7.43 for the oral input of thyroid hormone replacement therapy using T4. Assume T4 exists in compartment 7 (gut) in solid form and in compartment 8 (still the gut) in liquid form. Compartments 2 (for T3) and 5 (for T4) represent the plasma, compartments 1 and 4 are the fast tissues, and compartments 3 and 6 are the slow tissues. Assume that the input is bolus, $f_7(t) = 25\delta(t)$, and that the initial conditions are zero. Further, $K_{78} = 1.1$, $K_{80} = 0.01$, $K_{85} = 0.62$, $K_{21} = 15$, $K_{12} = 30$, $K_{10} = 1.0$, $K_{23} = 0.5$, $K_{32} = 0.4$, $K_{30} = 0.05$, $K_{40} = 0.08$, $K_{45} = 0.45$, $K_{54} = 0.28$, $K_{56} = 0.05$, $K_{65} = 0.017$, and $K_{60} = 0.018$. Note that K_{10}, K_{30}, K_{40}, and K_{60} represent T3 and T4 metabolism. Solve for the quantity of T3 in compartment 2.

108. Consider the model illustrated in Figure 7.44 for oral input of T4 thyroid hormone replacement therapy. Assume T4 exists in compartment 7 (gut) in solid form and in compartment 8 (still the gut) in liquid form. Compartments 2 (for T3) and 5 (for T4) represent the plasma, compartments 1 and 4 are the fast tissues, and compartments 3 and 6 are the slow tissues. Assume that the input is bolus, $f_7(t) = 25\delta(t)$, and that the initial conditions are zero. Further, assume that $K_{78} = 1.1$, $K_{80} = 0.01$, $K_{85} = 0.62$, $K_{21} = 7.0$, $K_{12} = 10$, $K_{10} = 0.8$, $K_{23} = 2.0$, $K_{32} = 0.3$, $K_{30} = 0.1$, $K_{40} = 0.06$, $K_{45} = 1.0$, $K_{54} = 0.3$, $K_{56} = 0.0$, $K_{65} = 0.03$, and $K_{60} = 0.02$. Note that K_{10}, K_{30}, K_{40}, and K_{60} represent T3 and T4 metabolism. Solve for the quantity of T3 in compartment 2.

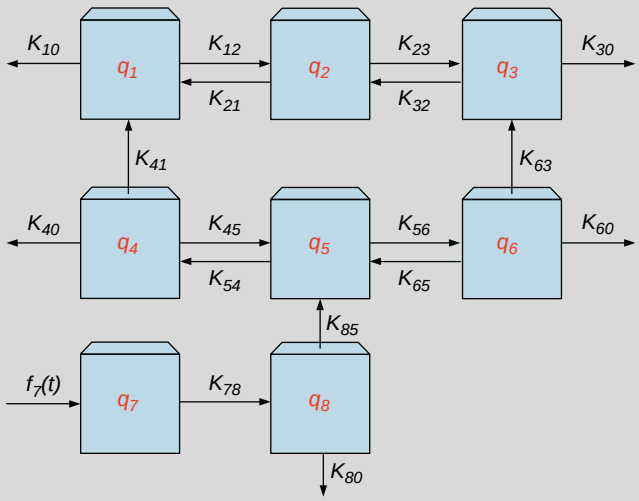


FIGURE 7.43 Illustration for Exercise 107.

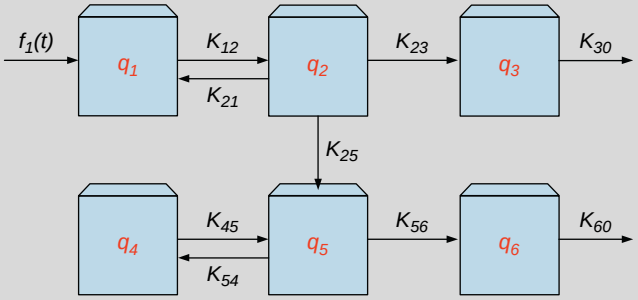


FIGURE 7.44 Illustration for Exercises 108 and 109.

109. Solve for the quantity in each compartment shown in Figure 7.44 given $K_{12} = 1.6$, $K_{21} = 0.5$, $K_{23} = 2.0$, $K_{30} = 0.5$, $K_{25} = 2.5$, $K_{45} = 0.4$, $K_{54} = 1.5$, $K_{60} = 0.5$, $K_{56} = 0.4$, and $f_1(t) = 10\delta(t)$.

Suggested Reading and References

E. Ackerman, L.C. Gatewood, Mathematical Models in the Health Sciences, University of Minnesota Press, Minneapolis, 1979.

E.S. Allman, J.A. Rhodes, Mathematical Models in Biology, An Introduction, Cambridge University Press, Cambridge, UK, 2004.

S.A. Berger, W. Goldsmith, E.R. Lewis, Bioengineering, Oxford University Press, Oxford, UK, 1996.

G.E. Briggs, J.B.S. Haldane, A note on the kinetics of enzyme action, Biochem. J. 19 (1925) 338–339.

N.F. Britton, Essential Mathematical Biology, Springer, London, 2003.

J.H.U. Brown, J.E. Jacobs, L. Stark, Biomedical Engineering, F.A. Davis Company, Philadelphia, 1971.

E. Carson, C. Cobelli, Modeling Methodology for Physiology and Medicine, Academic Press, London, 2001.

J.R. Cameron, J.G. Skofroinick, R. Grant, Physics of the Body, Medical Physics Publishing, Madison, WI, 1999.

J.J. DiStefano, F. Mori, Am. J. Physiol. Regul. Integr. Comp. Physiol. 233 (1977) 134–144.

J.J. DiStefano, P.H. Mak, Am. J. Physiol. Regul. Integr. Comp. Physiol. 236 (1979) 137–141.

R. Fisher, Compartmental Analysis, in: J.D. Enderle, S.M. Blanchard, J.D. Bronzino (Eds.), Introduction to Biomedical Engineering, Academic Press, San Diego, California, 2000, pp. 1062.

M.E. Fisher, A Semiclosed-Loop Algorithm for the Control of Blood Glucose Levels in Diabetics, IEEE Trans. Biomed. Eng. 38 (1) (1991).

K. Godfrey, Compartmental Models and Their Applications, Academic Press, San Diego, California, 1983.

A.C. Guyton, Textbook of Medical Physiology, eighth ed., W.B. Saunders Company, Philadelphia, 1991.

W.M. Haddad, V.S. Chellaboina, E. August, Stabilitiy and Dissipativity Theory for Discrete-Time Non-Negative and Compartmental Dynamical Systems, International Journal of Control 76 (18) (2003) 1845–1861.

V. Henri, Lois Générales de l'Action des Diastases, Hermann, Paris, 1903.

F.C. Hoppensteadt, C.S. Peskin, Mathematics in Medicine and the Life Sciences, Springer-Verlag, New York, 1990.

J.A. Jacquez, Modeling with Compartments, BioMedware, Ann Arbor, MI, 1999.

J.A. Jacquez, Compartmental Analysis in Biology and Medicine, third ed, BioMedware, Ann Arbor, MI, 1996.

J. Keener, J. Sneyd, Mathematical Physiology, Springer, New York, 1998.

L. Michaelis, M. Menten, Die Kinetik der Invertinwirkung, Biochem. Z. 49 (1913) 333–369.

J.D. Murray, Mathematical Biology, third ed., Springer, New York, 2001.

R.B. Northrop, Endogenous and Exogenous Regulation and Control of Physiological Systems, CRC Press, 1999.

A. Ritter, S. Reisman, B. Michniak, Biomedical Engineering Principles, CRC Press, Boca Raton, FL, 2005.

S. Schnell, C. Mendoza, Closed form solution for time-dependent enzyme kinetics, J. Theor. Biol. 187 (1997) 207–212.

This page intentionally left blank

CHAPTER

8

Biochemical Reactions and Enzyme Kinetics

John D. Enderle, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Using the law of mass action, quantitatively describe a chemical reaction using differential equations.
- Using the law of mass action, quantitatively describe enzyme kinetics using differential equations.
- Describe and calculate the quasi-steady-state approximation for chemical reactions.
- Simulate chemical reactions and enzyme kinetics using SIMULINK.
- Use the quasi-steady-state approximation to describe metabolism in a compartmental model.
- Model biochemical reactions, enzyme kinetics, diffusion, carrier-mediated transport and active transport in a physiological system.
- Quantitatively describe the Na-K pump.
- Qunatitatively describe cellular respiration.
- Explain and model activators and inhibitors in a chemical reaction.

This chapter focuses on chemical reactions that occur inside and outside a cell following the law of mass action—that is, the rate of accumulation is proportional to the product of the reactants. These chemical reactions support all functions needed to support life and involve such activities as the synthesis of hormones and proteins, muscle contraction, respiration, reproduction, neural signaling, and many other reactions. While the law of mass action is useful, it is not appropriate for all chemical reactions. In some cases, the exact chemical reaction mechanism is not known, and the law of mass action doesn't work.

The law of mass action creates models that are nonlinear and possibly time-varying. In addition, the rate of accumulation is profoundly impacted by temperature, such that an increase in temperature increases the reaction rate. The reactants are assumed to be uniformly distributed in the compartment, the probability of a collision depends on the concentration of the reactants, and such collisions are sufficient to create the products. Since the models here are nonlinear, we typically use SIMULINK to solve these problems. However, we will present analytical solutions to some special nonlinear cases in this chapter.

Catalysts are substances that dramatically change reaction rates. These substances are generally present in small amounts and are not consumed in the reaction. Their function is to decrease the amount of time to reach steady state in a chemical reaction. Consider two reactants that spontaneously create a product at room temperature but do so at a very low rate. In the presence of an appropriate catalyst, the speed of the reaction is dramatically increased, and the time to reach steady state is decreased. Enzymes are protein catalysts used in biological reactions that regulate and control most processes in the body. Enzymes increase the reaction rate by thousands or even trillions and are reactant specific. Typically, the enzyme concentration is rather small in relation to the reactant. While the reaction rate is profoundly increased in the presence of an enzyme, the overall energy used to form the product does not change.

In this chapter, we first examine simple chemical reactions and then enzyme kinetics. Next, we look at quasi-steady-state approximations, the Michaelis-Menten elimination, and enzyme regulation. We finish by examining important processes, such as the transport of oxygen and carbon dioxide through the circulatory system to the cell, the Na-K pump, and cellular respiration.

8.1 CHEMICAL REACTIONS

Consider the following single-stage chemical reaction

$$A + B \xrightarrow{K} P \tag{8.1}$$

in which chemicals A and B react to form the product P, with reaction rate constant K. Equation (8.1) is known as a stoichiometric equation that lists the number of reactants on the left side necessary to form the product on the right side. Conservation of mass requires that the total quantity of reactant A must equal the quantity of A in the reactant and the product, and likewise, the same is true with reactant B. The stoichiometric equation does not describe the dynamics or kinetics of the chemical reaction—that is, the time course of the reaction that may be very fast or very slow. The kinetics of the reaction is written

according to the law of mass action. The arrow in Eq. (8.1) shows the direction of the reaction that occurs spontaneously, and the reaction rate constant describes how quickly the reaction occurs. The reaction rate constant is a function of temperature, whereby an increase in temperature generally increases K. For our purposes, temperature is constant and so is K.

8.1.1 Chemical Bonds

Individual atoms rarely appear in nature, but they appear within a molecule. Molecules are collections of atoms that are held together by strong chemical bonds in which electron(s) are shared or electron(s) are transferred from one atom to another. Forces between molecules are relatively weak, allowing molecules to act independently of one another. A compound is a molecule that contains at least two different atoms, whereas a molecule can contain just one type of atom. Molecules such as hydrogen, H_2, and oxygen, O_2, are not compounds because they consist of a single type of atom. For simplicity in presentation, we will use the term *molecules* to describe both molecules and compounds.

An atom is an electrically neutral particle that consists of an equal number of protons and electrons. An atom with an unbalanced number of electrons or protons is called an ion, which makes it a positively or negatively charged particle. Examples of positively charged ions, called cations, include Na^+, K^+, and Ca^{+2}; Na^+ and K^+ have each lost one electron, and Ca^{+2} has lost two electrons. An example of a negatively charged ion, called an anion, is Cl^-, which has gained an extra electron.

There are many types of chemical bonds that join atoms and molecules together. They vary in strength, with ionic and covalent bonds displaying the strongest bonds, and the hydrogen bond displaying the weakest bond.

Many molecules are composed of positively and negatively charged ions that are bound by an ionic bond. These bonds are extremely strong due to the electrostatic attraction between the oppositely charged ions. Ionic bonds usually involve an atom that has few electrons in the outer shell, with another atom that has an almost complete set of electrons in the outer shell. Ionic bonds involve a transfer of electrons from one atom to another atom. Consider sodium chloride, where sodium has one electron in the outer shell and chloride has seven electrons in its outer shell. The bond is formed by sodium transferring its electron to chloride, resulting in a molecule consisting of Na^+Cl^- (usually written as $NaCl$). Consider magnesium and oxygen atoms forming an ionic bond. Magnesium has two electrons in its outer field that are transferred to oxygen, forming $Mg^{+2}O^{-2}$.

Covalent bonds involve two atoms sharing an electron pair that increases the stability of the molecule. The atoms do not have to be the same type, but they must be the same electronegativity. For instance, H_2 is formed by a covalent bond. Consider carbon dioxide, a molecule that consists of one carbon and two oxygen atoms. These atoms share the electrons, with carbon having four electrons in its outer shell, and oxygen having six electrons in its outer shell. The four electrons in carbon are used in the outer shell of the two oxygen atoms.

A hydrogen bond involves the force between a hydrogen atom in one molecule and an electronegative atom in another molecule, typically with oxygen in H_2O, nitrogen in NH_3, and fluorine in HF. Note that the force experienced in a hydrogen bond is not within a molecule but between molecules. The small size of hydrogen allows it to become very close to another molecule with a small atomic radii like the ones previously mentioned.

An example of a hydrogen bond is found among water molecules. The oxygen in water, which is negatively charged, binds to two positively charged hydrogens from two other water molecules. These bonds connect the water molecules together and give it some unusual properties, such as high surface tension and viscosity. Keep in mind that this bond is a relatively weak bond compared to that holding the water molecule together. As water is heated to the boiling point, all of the hydrogen bonds are dissolved as it becomes a gas. When water freezes, the hydrogen bonds form a structured, less dense orientation. Between freezing and boiling, water is oriented in very dense, random configuration. Hydrogen also plays an an important role within proteins and nucleic acids, which allows bonding within the molecule to achieve different shapes.

A chemical reaction in which a molecule loses electrons is called oxidation, and a gain of electrons is called reduction. Some multistage chemical reactions involve both oxidation and reduction. Energy is released when a chemical bond is formed, which is the same amount of energy necessary to break that chemical bond. Thus, chemical bonds store energy. The first step in a chemical reaction requires energy to break the bonds holding the reactant molecules together so they can be rearranged to form an activated complex, an unstable high-energy intermediate. From the activated complex, the molecules join together to form the product and typically release energy. In chemical reactions, if the energy stored in the chemical bonds of the product is less than the total energy stored in the reactants, then net energy is released. The amount of energy required in a chemical reaction is inversely proportional to the reaction rate constant.

8.1.2 Kinetics of a Single-Stage Chemical Reaction

To understand the law of mass action, a compartment model visualization of Eq. (8.1) is shown in Figure 8.1, with the understanding that the volumes for the two compartments are identical (i.e., the reaction occurs in a single container of constant volume). The differential equation describing Eq. (8.1) or Figure 8.1 is

$$\dot{q}_A = -Kq_Aq_B \tag{8.2}$$

where q_A and q_B are the quantities of reactants A and B, and K is the transfer rate or reaction rate constant. The transfer rate includes the volume in converting the concentrations to quantities in Eq. (8.2). Note that the right-hand side of Eq. (8.2) involves the product q_Aq_B, instead of a single variable as before. In addition, we could have written Eq. (8.2) in terms of q_B, $\dot{q}_B = -Kq_Aq_B$ or in terms of q_P, $\dot{q}_P = Kq_Aq_B$. Since drawing the compartmental model is not essential in writing the differential equation that describes the kinetics of the

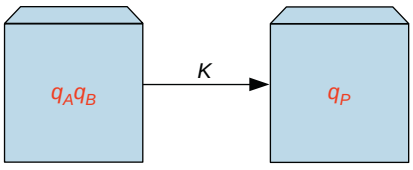


FIGURE 8.1 Compartmental model of the single-stage chemical reaction in Eq. (8.1). Another way to describe this system is to remove one of the reactants in the left box and include it as part of the reaction rate—that is, Kq_B.

chemical reaction, we will write the differential equations directly from the stoichiometric equation. Of course, Eq. (8.2) can be written in terms of concentrations by substituting (concentration×volume) for quantity.

Eq. (8.2) is a nonlinear differential equation. In general, nonlinear equations cannot be solved directly, but they must be simulated using a program like SIMULINK. Keep in mind that there are some special nonlinear differential equations that can be solved, but these are few in number, with a few presented in this chapter. Since $\dot{q}_A = \dot{q}_B$, we can integrate both sides of $\int_0^t dq_A = \int_0^t dq_b$ and have $q_A - q_A(0) = q_B - q_B(0)$, which, when substituted into Eq. (8.2), gives

$$\dot{q}_A = -Kq_Aq_B = -Kq_A(q_A - q_A(0) + q_B(0)) \tag{8.3}$$

Rearranging Eq. (8.3) yields

$$\frac{dq_A}{q_A(q_A - q_A(0) + q_B(0))} = -Kdt \tag{8.4}$$

The left-hand side of Eq. (8.4) is rewritten using a technique called partial fraction expansion (details are briefly described here for the unique roots case)[1] as

$$\begin{gathered}\left(\frac{B_1 dq_A}{(q_A - q_A(0) + q_B(0))} + \frac{B_2 dq_A}{q_A}\right) = -Kdt \\ \left(\frac{1}{q_A(0) - q_B(0)}\right)\left(\frac{dq_A}{(q_A - q_A(0) + q_B(0))} - \frac{dq_A}{q_A}\right) = -Kdt\end{gathered} \tag{8.5}$$

where

$$B_1 = \left[(q_A - q_A(0) + q_B(0))\frac{1}{q_A(q_A - q_A(0) + q_B(0))}\right]\Bigg|_{q_A = q_A(0) - q_B(0)} = \left[\frac{1}{q_A}\right]\Bigg|_{q_A = q_A(0) - q_B(0)} = \frac{1}{q_A(0) - q_B(0)}$$

[1]In general, if

$$\frac{1}{s^n + a_{n-1}s^{n-1} + \cdots a_1 s + a_0} = \frac{1}{(s - s_1)\cdots(s - s_n)}$$

and the roots are real, then

$$\frac{1}{(s - s_1)\cdots(s - s_n)} = \frac{B_1}{s - s_1} + \cdots \frac{B_n}{s - s_n}$$

where

$$B_i = \left[(s - s_i)\frac{1}{(s - s_1)\cdots(s - s_n)}\right]\Bigg|_{s = s_i}$$

It should be clear when evaluating the coefficient B_i in the previous equation that the factor $(s - s_i)$ always cancels with the same term in the denominator before calculating B_i.

$$B_2 = \left[q_A \frac{1}{q_A(q_A - q_A(0) + q_B(0))}\right]\Big|_{q_A=0} = \left[\frac{1}{(q_A - q_A(0) + q_B(0))}\right]\Big|_{q_A=0} = -\frac{1}{q_A(0) - q_B(0)}$$

Integrating both sides of Eq. (8.5) gives

$$\left(\frac{1}{q_A(0) - q_B(0)}\right)\left(\ln\left(\frac{q_A - q_A(0) + q_B(0)}{q_B(0)}\right) - \ln\left(\frac{q_A}{q_A(0)}\right)\right) = -Kt \tag{8.6}$$

and after algebraically manipulating, we have

$$q_A = \frac{(q_A(0) - q_B(0))}{\left(1 - \left(\frac{q_B(0)e^{-K(q_A(0)-q_B(0))t}}{q_A(0)}\right)\right)} \tag{8.7}$$

One final note regarding the solution of Eq. (8.2) is to consider the case in which $q_B(0) \gg q_A(0)$. Here, the change in q_B is small compared to $q_B(0)$, and q_B can be treated essentially as a constant—that is, $q_B = q_B(0)$. Thus, an approximation to Eq. (8.2) is

$$\dot{q}_A = -Kq_Bq_A = -Kq_B(0)q_A \tag{8.8}$$

which can be straightforwardly solved as

$$q_A = q_A(0)e^{-Kq_B(0)t}u(t) \tag{8.9}$$

Substituting Eq. (8.9) into $\dot{q}_P = Kq_Bq_A = Kq_B(0)q_A = Kq_B(0)q_A(0)e^{-Kq_B(0)t}$ and integrating gives

$$q_P = q_A(0)\left(1 - e^{-Kq_B(0)t}\right)u(t) \tag{8.10}$$

It should be clear that the steady-state value of q_P equals $q_A(0)$.

Let us take another look at the solution in Eq. (8.7) for q_A, and compare it to that of Eq. (8.9) by assuming that $q_B(0) \gg q_A(0)$, which gives the approximation $q_B = q_B(0)$, and thus $q_A - q_A(0) = q_B - q_B(0) = 0$. Hence, we have $\ln\left(\frac{q_A - q_A(0) + q_B(0)}{q_B(0)}\right) = \ln\left(\frac{q_B(0)}{q_B(0)}\right) = 0$, and therefore Eq. (8.7) becomes

$$\left(\frac{1}{q_B(0) - q_A(0)}\right)\left(\ln\left(\frac{q_A}{q_A(0)}\right)\right) = \frac{1}{q_B(0)}\ln\left(\frac{q_A}{q_A(0)}\right) = -Kt \tag{8.11}$$

Rearranging and taking the exponential of both sides gives

$$q_A = q_A(0)e^{-Kq_B(0)t}u(t) \tag{8.12}$$

which is the result we found in Eq. (8.9). In terms of the body, many small molecules like glucose and other nutrients are available in large quantities as compared with other reactants. Thus, this simplification is often quite appropriate.

8.1.3 Single-Stage Reversible Chemical Reaction

Next, consider the single-stage reversible chemical reaction as given in Eq. (8.13):

$$A + B \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} P \tag{8.13}$$

Here, the chemicals A and B react to form the product P, and P has a reverse reaction to form A and B. The law of mass action describing this system is given by

$$\begin{aligned} \dot{q}_P &= K_1 q_A q_B - K_{-1} q_P \\ \dot{q}_A &= -K_1 q_A q_B + K_{-1} q_P \\ \dot{q}_B &= -K_1 q_A q_B + K_{-1} q_P \end{aligned} \tag{8.14}$$

Equation (8.14) is nonlinear, which can be solved using SIMULINK or also mathematically, as shown in the following example.

EXAMPLE PROBLEM 8.1

Consider the reaction given in Eq. (8.14), with $q_A(0) = 10, q_B(0) = 15$, and $q_P(0) = 0$, $K_1 = 2$, and $K_{-1} = 3$. Solve for q_P.

Solution

From Eq. (8.14), it is clear that $\dot{q}_P = -\dot{q}_A = -\dot{q}_B$ and after integrating, we have

$$q_P - q_P(0) = q_A(0) - q_A = q_B(0) - q_B \tag{8.15}$$

To solve for q_P, we eliminate q_a and q_B using Eq. (8.15) by substituting $q_A = q_A(0) + q_P(0) - q_P$ and $q_B = q_B(0) + q_P(0) - q_P$ into Eq. (8.14), giving

$$\dot{q}_P = K_1 q_A q_B - K_{-1} q_P = K_1 (q_P(0) + q_A(0) - q_P)(q_P(0) + q_B(0) - q_P) - K_{-1} q_P \tag{8.16}$$

Since $q_P(0) = 0$, we have

$$\begin{aligned} \dot{q}_P &= K_1 (q_A(0) - q_P)(q_B(0) - q_P) - K_{-1} q_P \\ &= 2(10 - q_P)(15 - q_P) - 3q_P \\ &= 2\left(q_P^2 - \frac{53}{2} q_P + 150 \right) \end{aligned} \tag{8.17}$$

and after rearranging terms

$$\frac{dq_P}{q_P^2 - \frac{53}{2} q_P + 150} = \frac{dq_P}{(q_P - 18.3)(q_P - 8.2)} = 2dt \tag{8.18}$$

Once again, partial fraction expansion is used to rewrite Eq. (8.18) as

$$0.0989 \left(\frac{dq_P}{(q_P - 18.3)} - \frac{dq_P}{(q_P - 8.2)} \right) = 2dt \tag{8.19}$$

Continued

where

$$B_1 = \left[(q_P - 18.3)\frac{1}{(q_P - 18.3)(q_P - 8.2)}\right]\Bigg|_{q_P=18.3} = \left[\frac{1}{(q_P - 8.2)}\right]\Bigg|_{q_P=18.3} = 0.0989$$

and

$$B_2 = \left[(q_P - 8.2)\frac{1}{(q_P - 18.3)(q_P - 8.2)}\right]\Bigg|_{q_P=8.2} = \left[\frac{1}{(q_P - 18.3)}\right]\Bigg|_{q_P=8.2} = -0.0989$$

Integrating Eq. (8.19) and rearranging yields

$$\ln\left(2.2\frac{(q_P - 8.2)}{(q_P - 18.3)}\right) = -20.2t$$

and after solving for q_P, we have

$$q_P = \left(\frac{8.2(1 - e^{-20.2t})}{1 - 0.45e^{-20.2t}}\right)u(t) \tag{8.20}$$

The solution for the reactants are found from $q_A = q_A(0) + q_P(0) - q_P$ and $q_B = q_B(0) + q_P(0) - q_P$, yielding

$$q_A = \left(\frac{1.8059 + 3.7e^{-20.2t}}{1 - 0.45e^{-20.2t}}\right)u(t)$$

and

$$q_B = \left(\frac{6.8059 + 1.4798e^{-20.2t}}{1 - 0.4476e^{-20.2t}}\right)u(t)$$

These results are shown in Figure 8.2. Note that conservation of mass is still maintained, since q_P contains both q_A and q_B, so $q_A + q_B + 2 \times q_P = 25$.

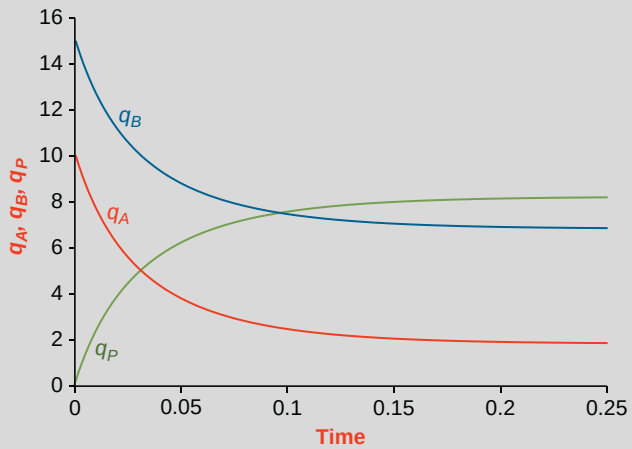


FIGURE 8.2 Illustration of the amount of product, q_P, in Example 8.1.

In general, it can be shown that the solution for q_P in Eq. (8.14) is

$$q_P = \left(\frac{\beta\left(1 - e^{-K_1(\alpha-\beta)t}\right)}{1 - \frac{\beta}{\alpha} e^{-K_1(\alpha-\beta)t}} \right) u(t) \tag{8.21}$$

where

$$\lambda = \sqrt{\left((q_A(0) + q_B(0) + 2q_P(0)) + \frac{K_2}{K_1} \right)^2 - 4(q_A(0) + q_P(0))(q_B(0) + q_P(0))}$$

$$\alpha = \frac{q_A(0) + q_B(0) + 2q_P(0) + \frac{K_2}{K_1} + \lambda}{2}$$

$$\beta = \frac{q_A(0) + q_B(0) + 2q_P(0) + \frac{K_2}{K_1} - \lambda}{2}$$

Next, consider the case in which $q_B(0) \gg q_A(0)$, where the change in q_B is small and can be treated as a constant $q_B(0)$. Thus, an approximation for Eq. (8.14) is

$$\dot{q}_P = K_1 q_B(0) q_A - K_{-1} q_P \tag{8.22}$$

Next, we remove q_A in Eq. (8.22) by letting $q_A = q_P(0) + q_A(0) - q_P$, giving

$$\dot{q}_P = K_1 q_B(0)(q_A(0) + q_P(0)) - (K_1 q_B(0) + K_{-1}) q_P \tag{8.23}$$

which can be straightforwardly solved as

$$q_P = \left(\frac{(K_{-1} q_P(0) - K_1 q_B(0) q_A(0))}{(K_1 q_B(0) + K_{-1})} e^{-(K_1 q_B(0) + K_{-1})t} + \frac{(K_1 q_B(0)(q_A(0) + q_P(0)))}{(K_1 q_B(0) + K_{-1})} \right) u(t) \tag{8.24}$$

8.1.4 Higher-Order Chemical Reactions and Sequential Reactions

Consider the higher-order chemical reaction in Eq. (8.25):

$$\alpha A + \beta B \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} P \tag{8.25}$$

in which αA of reactant A and βB of reactant B form product P, and P has a reverse reaction to form $\alpha A + \beta B$. The law of mass action for Eq. (8.25) is

$$\begin{aligned} \dot{q}_P &= K_1 q_A^\alpha q_B^\beta - K_{-1} q_P \\ \dot{q}_A &= -K_1 q_A^\alpha q_B^\beta + K_{-1} q_P \\ \dot{q}_B &= -K_1 q_A^\alpha q_B^\beta + K_{-1} q_P \end{aligned} \tag{8.26}$$

Note that when there are α molecules of A reacting, the q_A term in Eq. (8.26) is raised to the α power, and similarly, when β molecules of B react, then the q_B term in Eq. (8.26) is raised to the β power. In problems like this, we resort to using SIMULINK for the solution.

Now consider the sequential reaction in Eq. (8.27)

$$A \xrightarrow{K_1} B \xrightarrow{K_2} P \tag{8.27}$$

in which reactant A produces reactant B, which then produces reactant P. The law of mass action for Eq. (8.27) is

$$\begin{aligned} \dot{q}_A &= -K_1 q_A \\ \dot{q}_B &= K_1 q_A - K_2 q_B \\ \dot{q}_P &= K_2 q_B \end{aligned} \tag{8.28}$$

Equation (8.28) is easily solved, assuming nonzero initial condition for reactant A, $q_A(0)$, and zero for the other two reactants, as

$$\begin{aligned} q_A &= q_A(0) e^{-K_1 t} u(t) \\ q_B &= \frac{K_1 q_A(0)}{K_2 - K_1} \left(e^{-K_1 t} - e^{-K_2 t}\right) u(t) \\ q_P &= \frac{K_1 K_2 q_A(0)}{K_2 - K_1} \left(\frac{1 - e^{-K_1 t}}{K_1} - \frac{1 - e^{-K_2 t}}{K_2} \right) u(t) \end{aligned} \tag{8.29}$$

Sequential reactions are typical of those occurring in the body and can be quite complex, as described in Sections 8.4–8.6. As we will see, enzymes play a role in sequential reactions, and reactants can be of higher orders. Some sequential reactions involve many sequences in which some branch backward to previous reactions.

When there are profound differences in the transfer rates in a sequential reaction, simplifications occur that make analysis much easier. For instance, suppose $K_2 \gg K_1$. Here, the second reaction is much faster than the first reaction, and it appears that reactant A immediately produces reactant P. The term *rate limiting* is often used to describe this behavior—that is, the first reaction slows the creation of the product.

For a rate limiting reaction, an approximation to Eq. (8.29) for q_B and q_P is to eliminate the $e^{-K_2 t}$ term, since it goes to zero almost immediately, as compared to the $e^{-K_1 t}$ term, giving

$$\begin{aligned} q_B &\approx \frac{K_1 q_A(0)}{\cancel{K_2 - K_1}^{\,K_2}} \left(e^{-K_1 t} - \cancel{e^{-K_2 t}}^{\,0}\right) = \frac{K_1 q_A(0) e^{-K_1 t}}{K_2} = \frac{K_1}{K_2} q_A \\ q_P &\approx \frac{K_1 K_2 q_A(0)}{\cancel{K_2 - K_1}^{\,K_2}} \left(\frac{1 - e^{-K_1 t}}{K_1} - \cancel{\frac{1 - e^{-K_2 t}}{K_2}}^{\,0} \right) = q_A(0) \left(1 - e^{-K_1 t}\right) \end{aligned} \tag{8.30}$$

8.1.5 Quasi-Steady-State

Another way to look at a rate limiting reaction is to assume that reactant B is in a quasi-steady-state mode—that is $\dot{q}_B = 0$. From $\dot{q}_B = 0$ and Eq. (8.28), we have that $0 = K_1 q_A - K_2 q_B$,

and therefore $q_B = \frac{K_1}{K_2} q_A$. Since $\dot{q}_P = K_1 q_B$, we eliminate q_B by substituting $q_B = \frac{K_1}{K_2} q_A$, which gives $\dot{q}_P = K_1 q_A$ and $q_P = q_A(0)\left(1 - e^{-K_1 t}\right)$.

While quasi-steady-state assumes that reactant B is immediately in steady state and reactant A creates product P directly, there is a period of time, $\left(\frac{5}{K_2}\right)$, in which q_B moves from 0 to $\frac{K_1}{K_2} q_A$. Note also that steady state for reactant B is quite small and equals $\left(\frac{K_1}{K_2} q_A\right)$.

Figure 8.3 illustrates the approximation in Eq. (8.30), with the true solution for Eq. (8.29) given with $q_A(0) = 10, q_B(0) = 0$ and $q_P(0) = 0$, $K_1 = 2$, and $K_2 = 500$. For q_B, the approximation is quite accurate after it reaches quasi-steady-state, $t = \frac{5}{K_2} = 0.01$. For q_P, the approximation is quite accurate for the entire duration.

Now suppose $K_1 \gg K_2$. By a similar rational, the second reaction is now slower as compared to the first reaction and is rate limiting. For the rate limiting second reaction, an approximation to Eq. (8.29) for q_B and q_P is to eliminate the $e^{-K_1 t}$ term, since it goes to zero almost immediately, as compared to the $e^{-K_2 t}$ term, giving

$$
\begin{aligned}
q_A &= q_A(0) e^{-K_1 t} \\
q_B &\approx \frac{K_1 q_A(0)}{K_2 - K_1} \left(e^{-K_1 t} - e^{-K_2 t} \right) = q_A(0) e^{-K_2 t} \\
q_P &\approx \frac{K_1 K_2 q_A(0)}{K_2 - K_1} \left(\frac{1 - e^{-K_1 t}}{K_1} - \frac{1 - e^{-K_2 t}}{K_2} \right) = q_A(0) \left(1 - e^{-K_2 t}\right)
\end{aligned}
\tag{8.31}
$$

Here, reactant A disappears almost immediately and reactant B increases almost immediately to $q_A(0)$.

Another way to look at this rate limiting reaction is to assume that reactant A is in a quasi-steady-state mode—that is $\dot{q}_A = 0$. From $\dot{q}_A = 0$ and Eq. (8.28), we have $q_A = 0$ and

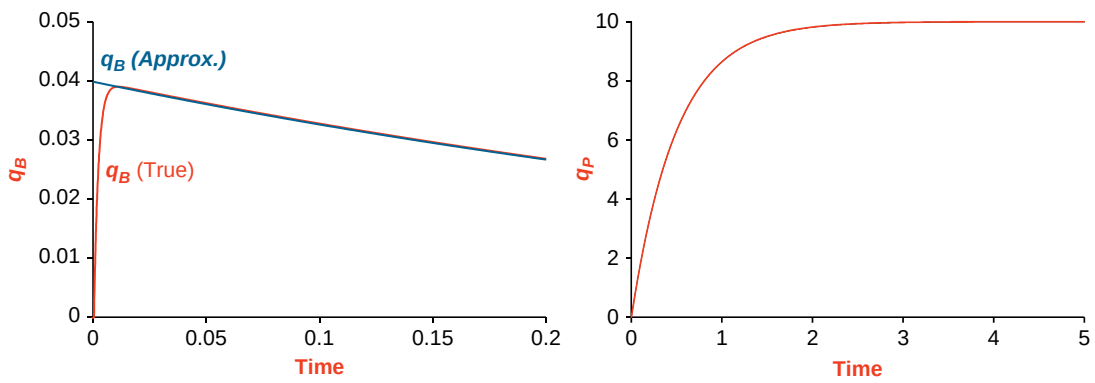


FIGURE 8.3 A rate limiting sequential reaction for K_1. Note that both the approximation and true solution for q_P are drawn in the figure on the right.

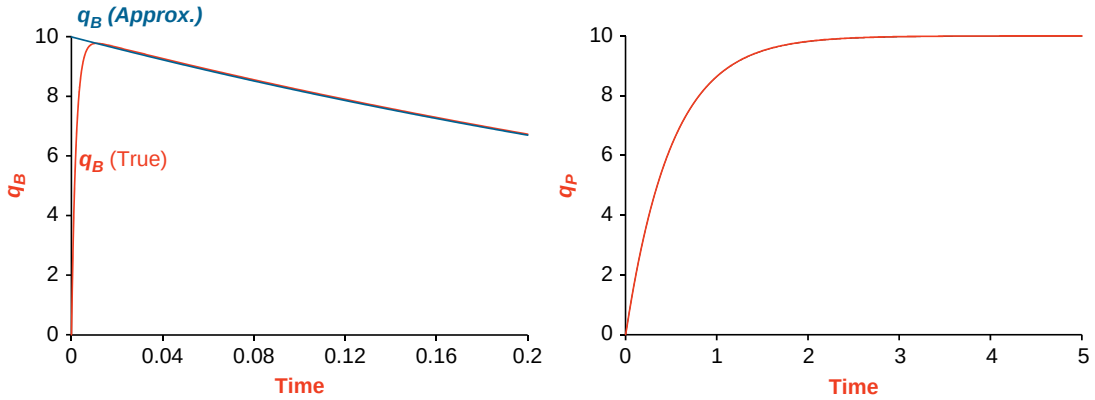


FIGURE 8.4 A rate limiting sequential reaction for K_2. Note that both the approximation and true solution for q_P are drawn in the figure on the right.

$\dot{q}_B = -K_2 q_B$, which has the solution $q_B = q_A(0)e^{-K_2 t}$. Next, $\dot{q}_P = K_2 q_B = q_A(0)K_2 e^{-K_2 t}$, which has the solution $q_P = q_A(0)\left(1 - e^{-K_2 t}\right)$. While quasi-steady-state assumes that reactant A is immediately in steady state at time 0 and reactant B rises to $q_A(0)$ immediately, there is a period of time, $\left(\frac{5}{K_1}\right)$, in which q_A falls to 0 and q_B rises.

Figure 8.4 illustrates the approximation in Eq. (8.31), with the true solution in Eq. (8.29) given with $q_A(0) = 10, q_B(0) = 0$ and $q_P(0) = 0$, $K_1 = 500$, and $K_2 = 2$. For q_B, the approximation is quite accurate after it reaches quasi-steady-state, $t = \frac{5}{K_1} = 0.01$. For q_P, the approximation is quite accurate the entire duration.

8.2 ENZYME KINETICS

As described earlier, catalysts are substances that accelerate reactions but are not consumed or changed by the reaction. Most chemical reactions in the body can occur without the presence of catalysts, but they occur at a very low rate. An enzyme is a large protein that catalyzes biochemical reactions in the body. These reactions convert a reactant, now called a substrate because it involves an enzyme, into a product by lowering the free energy of activation. Enzymes can increase the rate of the reaction by an order of thousands to trillions. Each enzyme is highly specific and only allows a particular substrate to bind to its active site. Many enzymes are used in the control and regulation of functions of the body.

In general, an enzyme reaction involves a series of reactions. Binding the enzyme with the substrate is the first step in creating an intermediate complex, which increases the ability of the substrate to react with other molecules. The next step is when the substrate-enzyme complex breaks down to form the free enzyme and product. Both the first step and the second step are reversible, but in the second step, the reverse reaction is so small that it is often omitted.

The overall rate of the enzyme catalyzed reaction with a single substrate is a function of the amount of enzyme and substrate and is given as

$$\text{Reaction Rate} = \frac{K_2 \times \text{Enzyme} \times \text{Substrate}}{K_M + \text{Substrate}} \tag{8.32}$$

where K_2 and K_M are reaction rate constants (K_M is called the Michaelis constant). We will derive and discuss this equation in more detail later in this section.

In an enzyme catalyzed reaction with much more substrate than enzyme, the reaction rate depends linearly on enzyme concentration according to Eq. (8.32), since the substrate concentration is essentially a constant. When there is much more enzyme than substrate, only a small portion of the enzyme is combined with the substrate, and the reaction rate is determined by both the enzyme and substrate levels. The typical enzyme catalyzed reaction consists of a series of reactions, each step with its own reaction rate. As we will see, the overall reaction rate is determined by the slowest reaction in the chain of reactions. The slowest reaction is called the capacity-limited reaction.

Enzyme catalyzed reactions serve a regulatory role, as well as accelerating biochemical reactions. Consider the relationship between adenosine diphosphate (ADP) and ATP inside the cell. ATP is created in the mitochondria where oxidation of nutrients (carbohydrates, proteins, and fats) produces carbon dioxide, water, and energy. The energy from the oxidation of nutrients converts ADP into ATP. ATP is used as fuel for almost all activities of the body, such as the $Na - K$ pump, action potentials, synthesis of molecules, creation of hormones, and contractions of muscles. At steady state, the concentration of ADP is very low in the cell, and thus the creation of ATP in the mitochondria is at a low rate. During periods of high cell activity, ATP is consumed, releasing energy through the loss of one phosphate radical, leaving ADP. The increased concentration of ADP causes an increase in the oxidation of nutrients in the mitochondria, producing more ATP. Thus, the ADP-ATP cycle is balanced and based on the needs of the cell. The reactions necessary to synthesize ATP are described in Section 8.5.

Enzyme reactions do not appear to follow the law of mass action; that is, as the substrate increases, the reaction rate does not increase without bound but reaches a saturation level (that is, it is capacity-limited). Capacity-limited reactions are quite prevalent and describe most metabolic reactions and functions of the body, such as the movement of molecules across the cell membrane and how substrates are removed from the body through the kidneys.

8.2.1 Michaelis-Menten Kinetics and the Quasi-Steady-State Approximation

In 1903, Victor Henri first described the relationship between a substrate and an enzyme, followed by Michaelis and Menten in 1913, and Briggs and Haldane in 1925, with a capacity-limited elimination rate for the chemical reaction. The first assumption is that the enzyme and substrate quickly reach steady state in the formation of the complex, and then the complex more slowly dissociates into the product and enzyme. We also assume that the amount of enzyme is much smaller than the substrate. We refer to the combined work here

as Michaelis-Menten kinetics. The models presented here assume that temperature and other conditions remain constant unless otherwise indicated.

This section introduces a standard approximation that greatly simplifies the analysis of biochemical reactions and predictions of how fast a reaction will occur. As shown, the quasi-steady-state approximation provides an excellent representation of the system's differential equations, which are problematic, since they involve stiff differential equations. By assuming a quasi-steady-state approximation, the set of differential equations is reduced to a set of algebraic equations. When the early pioneers developed the theory of enzyme reactions, the solution of the differential equations was not possible by either simulation or direct solution. Thus, the quasi-steady-state approximation allowed a rather complete description of the enzyme reaction except for the initial stage involving the formation of the complex. The set of algebraic equations also provides a mechanism to measure the parameters of the reaction.

Stiff differential equations require simulation solutions with an extremely small step size using the standard integrators. Consider Figure 8.4 and the graph for q_B. From 0 to .02 s, q_B changes quickly, and after that, it changes slowly. The step size needs to be very small, after which a small step size is not needed, since q_B is slowly changing. Thus, simulations take a very long time to run using the standard integrator. If the step size is too large, the simulation solution is incorrect because small errors amplify as it reaches steady state. The default integrator `ode45` in SIMULINK is not a good choice for stiff problems because it is inefficient. The integrator `ode23tb` is a better choice for stiff problems, which manipulates the step size using an efficient algorithm. We will explore this issue later in this section.

The capacity-limited reaction model uses a two-step process given by Eq. (8.33):

$$S + E \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} ES \xrightarrow{K_2} E + P \tag{8.33}$$

The enzyme mediated reaction first has substrate S combining with enzyme E to form the unstable complex ES. Then the complex ES breaks down into the product P and E.

The law of mass action equation for Eq. (8.33) is

$$\begin{aligned} \dot{q}_S &= -K_1 q_S q_E + K_{-1} q_{ES} \\ \dot{q}_{ES} &= K_1 q_S q_E - (K_{-1} + K_2) q_{ES} \\ \dot{q}_P &= K_2 q_{ES} \end{aligned} \tag{8.34}$$

Since $q_E(0) = q_E + q_{ES}$ or $q_E = q_E(0) - q_{ES}$, we have $\dot{q}_E = -\dot{q}_{ES}$. Thus,

$$\dot{q}_E = -K_1 q_S q_E + (K_{-1} + K_2) q_{ES} \tag{8.35}$$

To begin with the classical description, we assume that the system is closed—that is, $q_S(0) = q_S + q_{ES} + q_P$ and $q_E(0) = q_E + q_{ES}$. Note that E is not consumed in the reaction. Moreover, it is assumed that the complex ES increases quickly to a maximum, $q_{ES_{\max}}$ and then changes very slowly after that. Therefore, we assume complex ES is in a quasi-steady-state mode shortly after the reaction starts—that is, $\dot{q}_{ES} = 0$. We also assume that

the amount of enzyme is very small in comparison to the substrate and product and that $K_2 \gg K_{-1}$. From $\dot{q}_{ES} = 0$ and Eq. (8.34), we have

$$0 = K_1 q_S q_E - (K_{-1} + K_2) q_{ES} \tag{8.36}$$

Since $q_E(0) = q_E + q_{ES}$, we substitute $q_E = q_E(0) - q_{ES}$ into Eq. (8.36) to give

$$0 = K_1 q_S (q_E(0) - q_{ES}) - (K_{-1} + K_2) q_{ES} \tag{8.37}$$

and after rearranging,

$$q_{ES} = \frac{K_1 q_S q_E(0)}{(K_1 q_S + K_{-1} + K_2)} = \frac{q_S q_E(0)}{\left(q_S + \frac{K_{-1}+K_2}{K_1}\right)} = \frac{q_S q_E(0)}{(q_S + K_M)} = \frac{q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)} \tag{8.38}$$

where $K_M = \frac{K_{-1}+K_2}{K_1}$. The constant, K_M, is called the Michaelis constant, as mentioned in the beginning of this chapter.

The maximum complex ES from Eq. (8.38), with $q_s = q_s(0)$, is approximately

$$q_{ES_{\max}} \approx \left.\frac{q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)}\right|_{q_S = q_S(0)} = \frac{q_E(0)}{\left(1 + \frac{K_M}{q_S(0)}\right)} \tag{8.39}$$

Further, since $q_E = q_E(0) - q_{ES}$, we have

$$q_E = q_E(0) - \frac{q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)} = \frac{q_E(0)}{\left(1 + \frac{q_S}{K_M}\right)} \tag{8.40}$$

To find a quasi-steady-state approximation for q_S, we use Eq. (8.34); $\dot{q}_S = -K_1 q_S q_E + K_{-1} q_{ES} = -\left(\dot{q}_{ES} + K_2 q_{ES}\right)$. With $\dot{q}_{ES} = 0$ and q_{ES} from Eq. (8.38), we have

$$\dot{q}_S = -K_2 q_{ES} = -\frac{K_2 q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)} \tag{8.41}$$

Now, we have only two parameters that describe the change in substrate in Eq. (8.41), which takes the place of the set of differential equations in Eq. (8.34). Moreover, we will show that these two parameters can be estimated directly from data. Rearranging Eq. (8.41) gives

$$\left(1 + \frac{K_M}{q_S}\right) dq_S = -K_2 q_E(0) dt \tag{8.42}$$

and after integrating both sides of Eq. (8.42) yields

$$(q_S(0) - q_S) + K_M \ln\left(\frac{q_S(0)}{q_S}\right) = K_2 q_E(0) t$$

or

$$t = \frac{1}{K_2 q_E(0)} \left((q_S(0) - q_S) + K_M \ln\left(\frac{q_S(0)}{q_S}\right) \right) \tag{8.43}$$

Note that Eq. (8.43) provides a nonlinear expression for t as a function of q_S and that, until recently, it was thought impossible to write an expression of q_S as a function of t. In 1997, Schnell and Mendoza used computer algebra to solve Eq. (8.43) for q_S as a function of t using the Lambert W function. While Eq. (8.43) can be used to solve for t by substituting values of q_S from $q_S(0)$ to zero, it is far easier to simulate Eq. (8.41) for q_S.

The product is given by

$$\dot{q}_P = K_2 q_{ES} = \frac{K_2 q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)} \tag{8.44}$$

To eliminate q_S, we have $q_S = q_S(0) - q_{ES} - q_P$. Since q_{ES} is small compared to the other quantities after the initial phase of the response, we have $q_S = q_S(0) - q_P$ and

$$\dot{q}_P = \frac{K_2 q_E(0)}{\left(1 + \frac{K_M}{q_S}\right)} = \frac{K_2 q_E(0)}{\left(1 + \frac{K_M}{q_S(0) - q_P}\right)} \tag{8.45}$$

Solving Eq. (8.45) using the same approach to find Eq. (8.43), we have

$$t = \frac{1}{K_2 q_E(0)} \left(q_P - K_M \ln\left(\frac{q_S(0) - q_P}{q_S(0)}\right)\right) \tag{8.46}$$

This approximation works well for the reaction, except during the initial quick phase when both q_{ES} and q_P are small.

Biochemists define the reaction rate as either the rate of disappearance of the substrate or the appearance of the product. Using the Michaelis-Menten approximation, the reaction rates are equal and are given by $V = -\dot{q}_S = \dot{q}_P$. However, based on the true differential equations in Eq. (8.34), these two definitions are not equal. Using the approximation in Eq. (8.41), we have

$$V = -\dot{q}_S = \frac{K_2 q_S q_E(0)}{(q_S + K_M)} = \frac{V_{\max} q_S}{(q_S + K_M)} = \frac{V_{\max}}{\left(1 + \frac{K_M}{q_S}\right)} \tag{8.47}$$

where $V_{\max} = K_2 q_E(0)$ is the maximum velocity of the substrate disappearance. Equation (8.47) is plotted in Figure 8.5. Note that at high substrate levels, V approaches $V_{\max}$, since all of the enzyme is engaged and the velocity saturates. In this region, the reaction rate is independent of q_s. At low substrate levels, the reaction rate is approximately linearly dependent on substrate level, $V = \frac{V_{\max}}{K_M} q_s$.

Now consider when $V = \frac{V_{\max}}{2}$. Substituting into Eq. (8.47), we have

$$\frac{V_{\max}}{2} = \frac{V_{\max} q_S}{(q_S + K_M)}$$

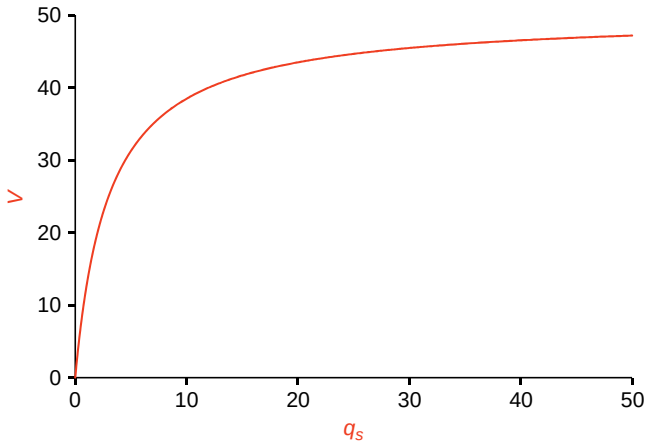


FIGURE 8.5 Velocity of substrate disappearance. Note that the reaction velocity is for $t = 0$, and q_S is the intial quantity of substrate. $V_{\max} = 50$ and $K_M = 3$.

and after dividing both sides by $V_{\max}$, we have

$$\frac{1}{2} = \frac{q_S}{(q_S + K_M)}$$

Simplifying the previous equation gives

$$q_S = K_M$$

when $V = \frac{V_{\max}}{2}$.
Given the lack of computer simulation capability in the early 1900s, the quasi-steady-state approximation gave an excellent and efficient solution to enzyme kinetics. However, with the current computer power and the availability of stiff differential equation simulators, it is far easier to simulate enzyme kinetics using the computer rather than solving a set of algebraic equations.

EXAMPLE PROBLEM 8.2

Simulate the reaction given in Eq. 8.34 and compare with the quasi-steady-state approximation for q_S, q_E, q_{ES}, and q_P. Assume that $K_1 = 8$, $K_{-1} = 0.01$, $K_2 = 5$, $q_S(0) = 1$, $q_E(0) = 0.08, q_{ES}(0) = 0$, and $q_P(0) = 0$.

Solution

The SIMULINK model is shown in Figures 8.6 and 8.7, both executed using the `ode23tb` integrator. Because of the ease in solution, the quasi-steady-state approximation for q_S, q_E, q_{ES}, and q_P is carried out using the differential equation for $\dot{q}_S$, Eq. (8.41), and the algebraic equations,

Continued

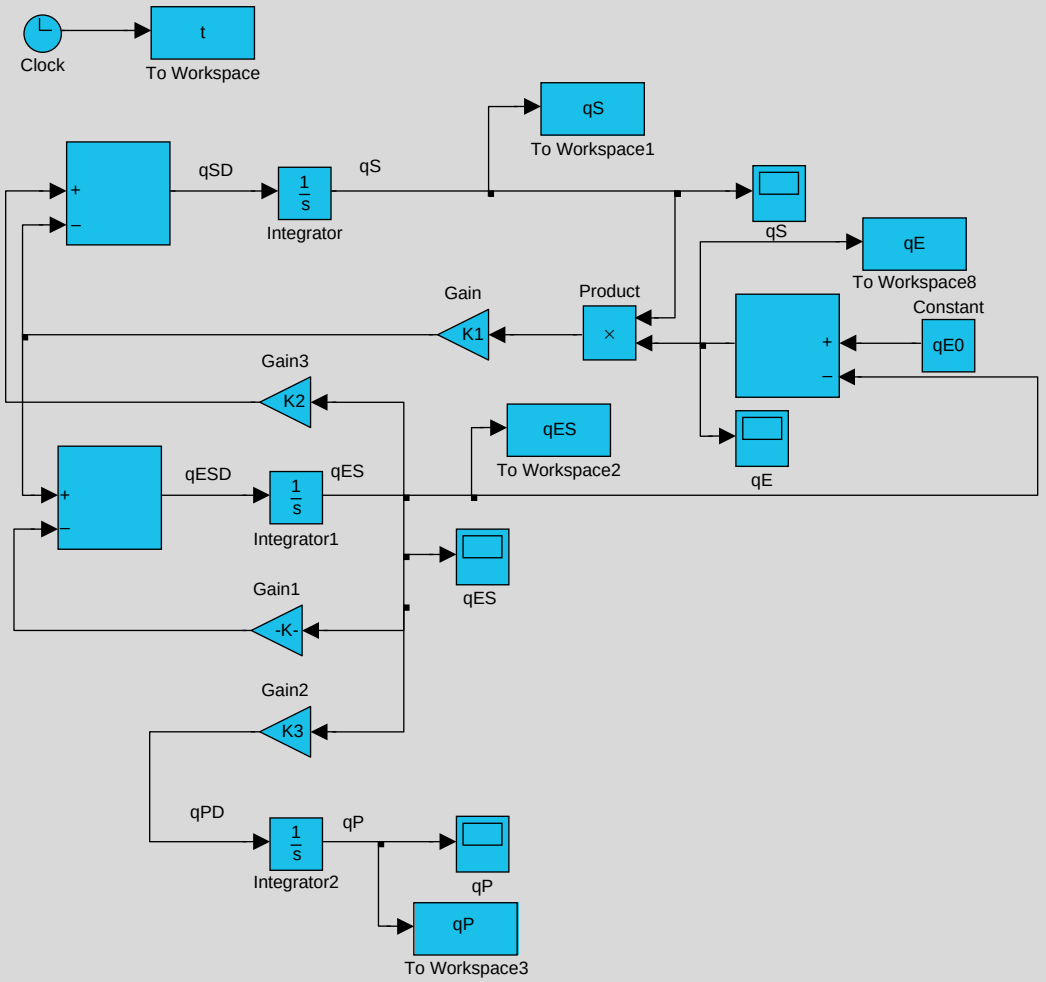


FIGURE 8.6 SIMULINK model for Eq. (8.34) in Example Problem 8.2.

Eqs. (8.38) and (8.40), and $q_P = q_S(0) - q_S - q_{ES}$. Figure 8.8 presents the simulation solution with both models. It is clear that there is an excellent agreement between the two solutions after the quick phase is completed at 0.5 s. Notice the expected error between q_E and q_{EMM}, and q_{ES} and q_{ESMM}. Also note that the initial product predicted by the quasi-steady-state approximation is negative.

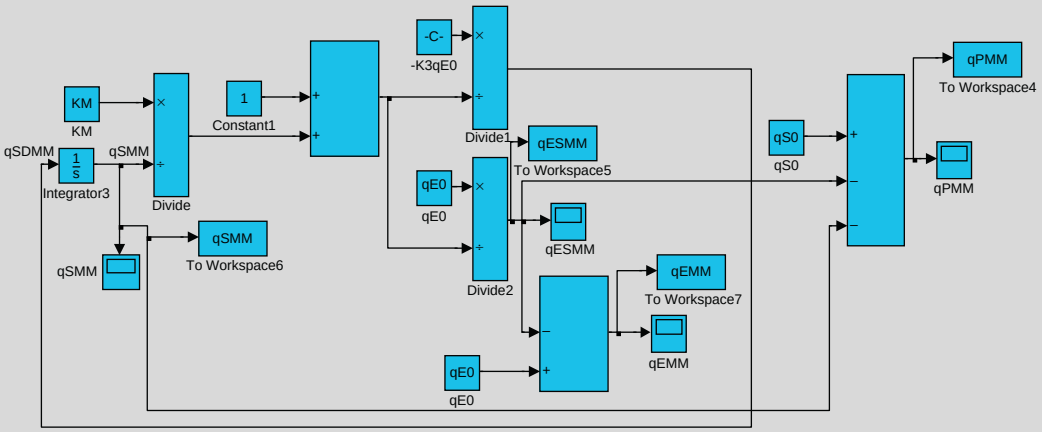


FIGURE 8.7 SIMULINK model in Example Problem 8.2 using the quasi-steady-state approximation.

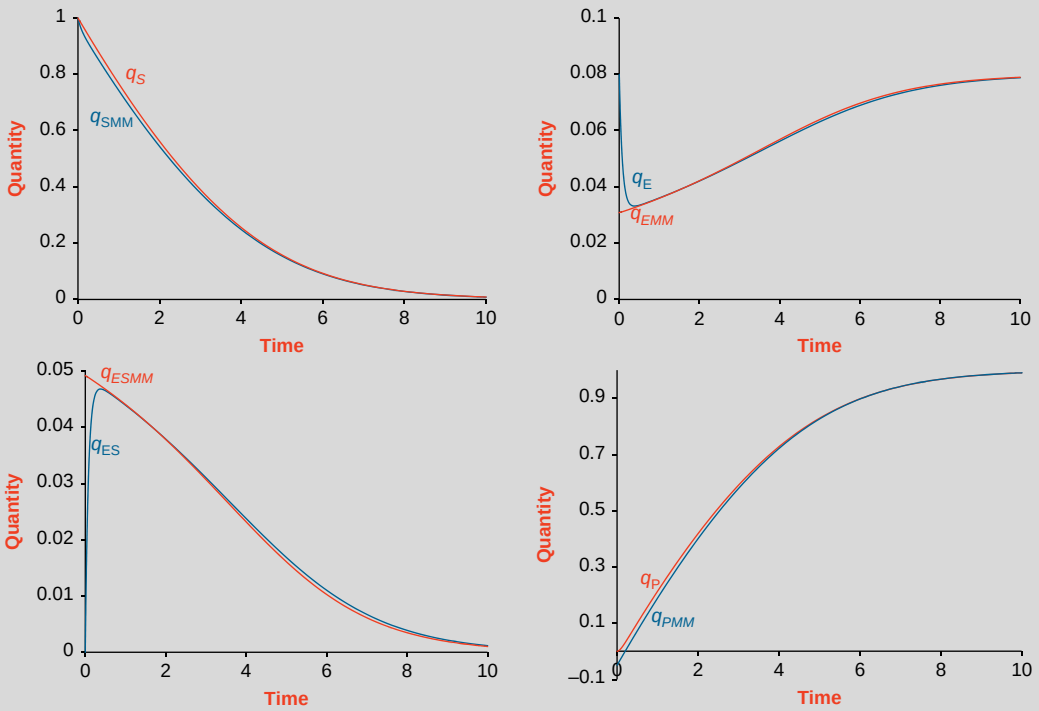


FIGURE 8.8 Simulations for Example Problem 8.2.

8.2.2 Estimation of the Michaelis-Menten Parameters

The reaction rate given by Eq. (8.47) contains two parameters, $V_{\max}$ and K_M, and is given by

$$V = \frac{V_{\max} q_S}{(q_S + K_M)} = \frac{V_{\max}}{\left(1 + \frac{K_M}{q_S}\right)} \tag{8.48}$$

One can estimate $V_{\max}$ and K_M by first taking the reciprocal of Eq. (8.48), giving

$$\left(\frac{1}{V}\right) = \frac{K_M}{V_{\max}}\left(\frac{1}{q_S}\right) + \frac{1}{V_{\max}} \tag{8.49}$$

which is called the Lineweaver-Burk equation. By plotting $\frac{1}{V}$ against $\frac{1}{q_S}$, a straight line results as shown in Figure 8.9, with the ordinate-intercept as $\frac{1}{V_{\max}}$ and the slope as K_M. Since the substrate changes as a function of time, the Lineweaver-Burk equation is usually carried out with the initial value of substrate, $q_S(0)$—that is,

$$\frac{1}{V(0)} = \frac{K_M}{V_{\max}}\left(\frac{1}{q_S(0)}\right) + \frac{1}{V_{\max}} \tag{8.50}$$

The method to generate data for the Lineweaver-Burk equation is to perform a series of experiments that increase the quantity of substrate, while keeping the amount of enzyme constant. Recall at $V = \frac{V_{\max}}{2}$, $q_S = K_M$. If $q_S \gg K_M$, then $V = V_{\max}$. While estimation of the parameters from the Lineweaver-Burk plot looks attractive, any measurement error is amplified by the transformation, thus giving poor estimates of the parameters.

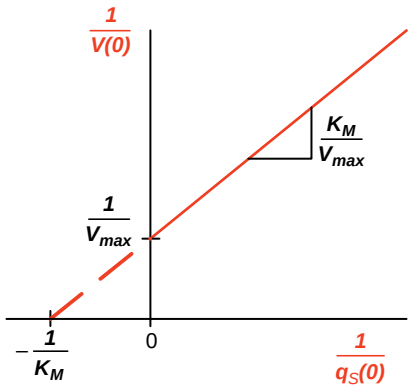


FIGURE 8.9 A Lineweaver-Burk plot.

8.3 ADDITIONAL MODELS USING THE QUASI-STEADY-STATE APPROXIMATION

The quasi-steady-state approximation for Michaelis-Menten kinetics can be used for more than just enzyme reactions.[2] The quasi-steady-state approximation is useful when describing the elimination of substances from the body with capacity-limited rates, such as renal excretion and metabolism, and even for linearized models of muscle using the Hill equation. While the quasi-steady-state approximation was developed for enzyme reactions with variables $q_S, q_E, q_{ES},$ and q_P, we will use it as a nonlinear transfer rate in a compartment model, where the substrate, q_S, in Eq. (8.47) is the quantity in a compartment. Here, we first consider a one-compartment model with a variety of inputs and then a two-compartment model.

8.3.1 One-Compartment Model

Consider a one-compartment model in which the elimination is characterized by the quasi-steady-state approximation

$$\dot{q}_S = -\frac{V_{\max}}{(q_S + K_M)} q_S + f(t) \tag{8.51}$$

Impulse Input

Consider first an impulse input. Previous examples have used $f(t) = \zeta\delta(t)$, where ζ is the strength of the impulse function—a problem that is handled most simply by a change in the initial condition. To solve Eq. (8.51) with an impulse input, we have

$$\dot{q}_S = -\frac{V_{\max}}{(q_S + K_M)} q_S \tag{8.52}$$

with $q_S(0) = \zeta$. As before, Eq. (8.52) is rearranged to give

$$\left(1 + \frac{K_M}{q_S}\right) dq_S = -V_{\max} dt$$

and after integrating, we have

$$t = \frac{1}{V_{\max}} \left(\zeta - q_S + K_M \ln\left(\frac{\zeta}{q_S}\right)\right) \tag{8.53}$$

The same comments about the solution of Eq. (8.43) apply to Eq. (8.53).

[2]Material in this section based on Godfrey.

Step Input

We now draw our attention to the case in which the input to the system in Eq. (8.51) is a step input, $f(t) = \zeta u(t)$. With initial condition $q_S(0)$, we have for $t \geq 0$

$$\dot{q}_S = -\frac{V_{\max}}{(q_S + K_M)} q_S + \zeta = \frac{(\zeta - V_{\max})q_S + \zeta K_M}{(q_S + K_M)} \tag{8.54}$$

Separating q_S and t in Eq. (8.54) gives

$$\begin{aligned} dt &= \frac{K_M dq_S}{(\zeta - V_{\max})q_S + \zeta K_M} + \frac{q_S dq_S}{(\zeta - V_{\max})q_S + \zeta K_M} \\ &= \frac{K_M dq_S}{(\zeta - V_{\max})q_S + \zeta K_M} + \left(\frac{1}{(\zeta - V_{\max})}\right)\left(\frac{q_S dq_S}{q_S + \dfrac{\zeta K_M}{(\zeta - V_{\max})}}\right) \\ &= \frac{K_M dq_S}{(\zeta - V_{\max})q_S + \zeta K_M} + \left(\frac{1}{(\zeta - V_{\max})}\right)\left(1 - \frac{\dfrac{\zeta K_M}{(\zeta - V_{\max})}}{q_S + \dfrac{\zeta K_M}{(\zeta - V_{\max})}}\right) dq_S \end{aligned} \tag{8.55}$$

Then we integrate Eq. (8.55)

$$\int_0^t dt = \int_{q_S(0)}^{q_S} \left(\frac{K_M dq_S}{(\zeta - V_{\max})q_S + \zeta K_M} + \left(\frac{1}{\zeta - V_{\max}}\right)\left(1 - \frac{\dfrac{\zeta K_M}{(\zeta - V_{\max})}}{q_S + \dfrac{\zeta K_M}{(\zeta - V_{\max})}}\right)\right) dq_S \tag{8.56}$$

which gives

$$\begin{aligned} t &= \left(\left(\frac{K_M}{(\zeta - V_{\max})}\right) \ln\left(\frac{(\zeta - V_{\max})q_S + \zeta K_M}{(\zeta - V_{\max})q_S(0) + \zeta K_M}\right) + \left(\frac{q_S - q_S(0)}{\zeta - V_{\max}}\right) \right. \\ &\quad \left. - \frac{\zeta K_M}{(\zeta - V_{\max})^2} \ln\left(\frac{q_S + \dfrac{\zeta K_M}{(\zeta - V_{\max})}}{q_S(0) + \dfrac{\zeta K_M}{(\zeta - V_{\max})}}\right) \right) \\ &= \frac{q_S - q_S(0)}{\zeta - V_{\max}} - \frac{V_{\max} K_M}{(\zeta - V_{\max})^2} \ln\left(\frac{(\zeta - V_{\max})q_S + \zeta K_M}{(\zeta - V_{\max})q_S(0) + \zeta K_M}\right) \end{aligned} \tag{8.57}$$

To determine q_S at steady state for this one-compartment system, we set $\dot{q}_S = 0$ in Eq. (8.54), which gives

$$q_S(\infty) = \frac{\zeta K_M}{V_{\max} - \zeta} \tag{8.58}$$

Since q_S needs to be positive, this requires $V_{\max} > \zeta$. Thus, the maximum removal rate must be larger than the input for a bound solution for q_S.

To examine the case when $V_{\max} = \zeta$, we substitute these values into Eq. (8.54), giving

$$\dot{q}_S = \frac{\zeta K_M}{(q_S + K_M)} \tag{8.59}$$

Separating q_S and t in Eq. (8.59) gives

$$dt = \frac{1}{\zeta K_M}(q_S + K_M)dq_S$$

and after integrating

$$t = \frac{1}{2\zeta K_M}\left(q_S^2 - q_S^2(0)\right) + \frac{1}{\zeta}\left(q_S - q_S(0)\right) \tag{8.60}$$

which increases without bound, $q_S \to \infty$. If $\zeta \geq V_{\max}$, then the substrate is not eliminated quickly enough, continuously increasing, and $q_S \to \infty$.

Exponential Input

Next, consider an exponential input to a one-compartment model of Eq. (8.51), a type of input previously observed in Example Problem 7.8, when a substrate is digested and moves into the plasma. For $t \geq 0$, the model is

$$\dot{q}_S = -\frac{V_{\max}}{(q_S + K_M)}q_S + q_2(0)K_{21}e^{-K_{21}t} \tag{8.61}$$

where $q_2(0)$ is the initial amount of substrate in the digestive system and K_{21} is the transfer rate from the digestive system to the compartment. Simulating Eq. (8.61) is the only way to solve this problem.

Before simulation became an easy solution method, the quasi-steady-state approximation was linearized with a lower and upper bound, with the actual solution falling in between these two bounds. For the lower bound, the quasi-steady-state approximation is $\frac{V_{\max}}{K_M}$, where the quasi-steady-state approximation is a constant. As we will see, the lower bound is a good approximation for small values of q_S. For the upper bound, the quasi-steady-state approximation is $\dfrac{V_{\max}}{q_{S_{\max}} + K_M}$, where $q_{S_{\max}}$ is the maximum q_S. As before with the lower bound, the quasi-steady-state approximation is a constant for the upper bound. The upper bound is a good approximation for large values of q_S. This approach works for all inputs.

For example, consider the exponential input with $q_S(0) = 0$, where the equation for the lower bound is

$$\dot{q}_S = -\frac{V_{\max}}{K_M} q_S + q_2(0) K_{21} e^{-K_{21} t} \tag{8.62}$$

with a solution of

$$q_S = \left(\frac{q_2(0) K_{21}}{\frac{V_{\max}}{K_M} - K_{21}} \right) \left(e^{-K_{21} t} - e^{-\frac{V_{\max}}{K_M} t} \right) \tag{8.63}$$

and the equation for the upper bound is

$$\dot{q}_S = -\frac{V_{\max}}{(q_{S_{\max}} + K_M)} q_S + q_2(0) K_{21} e^{-K_{21} t} \tag{8.64}$$

with a solution of

$$q_S = \left(\frac{q_2(0) K_{21}}{\left(\frac{V_{\max}}{q_{S_{\max}} + K_M} \right) - K_{21}} \right) \left(e^{-K_{21} t} - e^{-\left(\frac{V_{\max}}{q_{S_{\max}} + K_M} \right) t} \right) \tag{8.65}$$

To illustrate the approximations, Figure 8.10 plots the simulation of Eq. (8.61), with the analytic solution of Eqs. (8.63) and (8.65). For small values of q_S, the lower bound

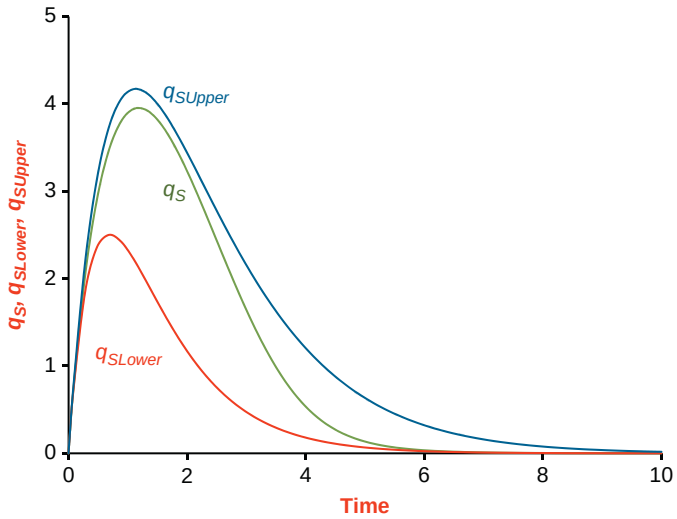


FIGURE 8.10 Responses for a one-compartment model with quasi-steady-state approximation (q_S), lower bound approximation (q_{SLower}), and upper bound approximation (q_{SUpper}), with an exponential input. Parameters are $V_{\max} = 3, K_M = 2.5, K_{21} = 1.0, q_S(0) = 0, q_{S_{\max}} = 4.0$, and $q_2(0) = 10$.

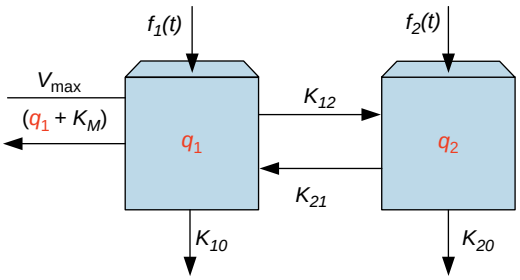


FIGURE 8.11 A two-compartment model with a quasi-steady-state approximation transfer rate.

approximation is quite good, as is the upper bound for large values of q_S. The accuracy of the lower and upper bound approximations depends quite heavily on the closeness of K_M to $q_{S_{\max}}$. The farther apart the two are, the poorer the approximation.

8.3.2 Two-Compartment Model

Next, consider a two-compartment model with a quasi-steady-state approximation transfer rate as shown in Figure 8.11. Such models are useful in describing some capacity-limited biochemical reactions that are more complex than a one-compartment model. Note that the quasi-steady-state approximation is for compartment 1 only, and the other transfer rates are the usual constants. The equations describing the system are

$$\begin{aligned}\dot{q}_1 &= -\left(K_{10} + K_{12} + \frac{V_{\max}}{(q_1 + K_M)}\right) q_1 + K_{21} q_2 + f_1(t) \\ \dot{q}_2 &= K_{12} q_1 - (K_{20} + K_{21}) q_2 + f_2(t)\end{aligned} \tag{8.66}$$

EXAMPLE PROBLEM 8.3

Simulate the model in Eq. (8.66), given that $K_{12} = 1, K_{21} = 2,\ V_{\max} = 3,\ K_M = 0.5,\ q_1(0) = 0,\ q_2(0) = 0,\ K_{10} = 0$, and $K_{20}(0) = 0$. The inputs are $f_1(t) = u(t) - u(t-4)$, and $f_2(t) = 0$.

Solution

The equations describing this system are

$$\dot{q}_1 = -\left(3 + \frac{1}{(q_1 + 0.5)}\right) q_1 + 2q_2 + u(t) - u(t-4)$$

$$\dot{q}_2 = 3q_1 - 2q_2$$

The SIMULINK model shown in Figure 8.12 was executed using the `ode23tb` integrator. The solution is shown in Figure 8.13.

Continued

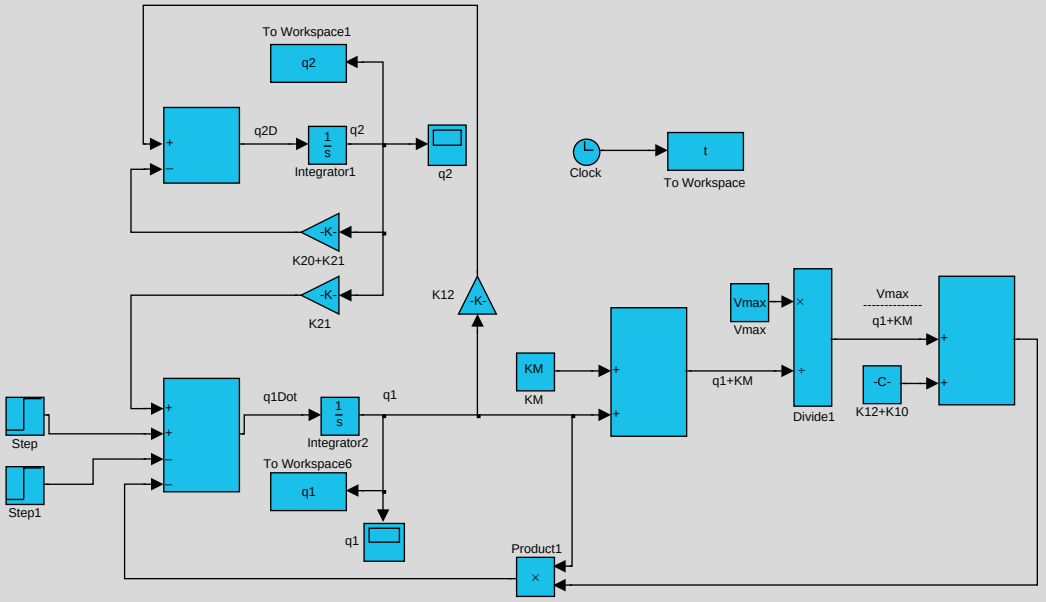


FIGURE 8.12 SIMULINK model for Example Problem 8.3.

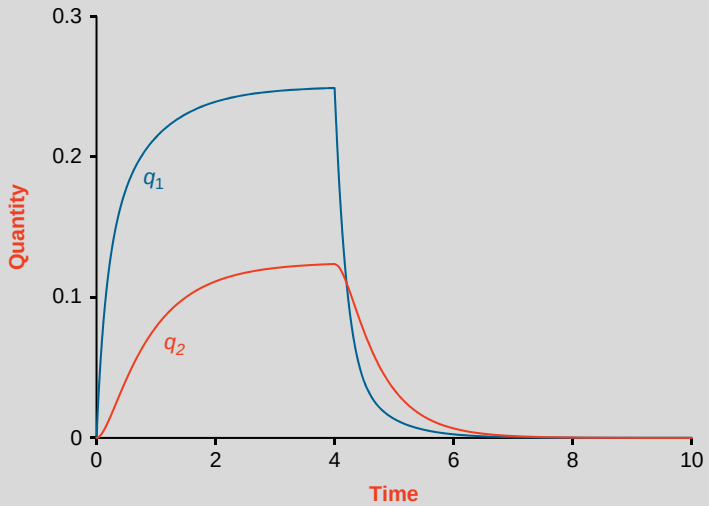


FIGURE 8.13 Response for Example Problem 8.3.

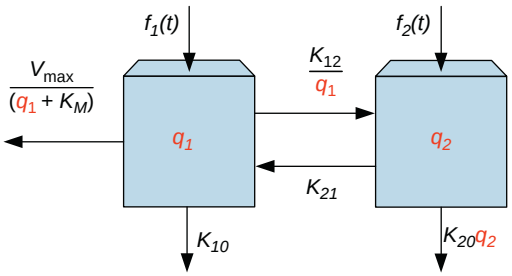


FIGURE 8.14 A two-compartment model with a quasi-steady-state approximation and biochemical reaction transfer rates. In addition, a constant input from compartment 1 into compartment 2 is used, with $\frac{K_{12}}{q_1}$ as the transfer rate.

Consider the model shown in Figure 8.14 that contains a quasi-steady-state approximation and a biochemical reaction as transfer rates. We also have a constant transfer of substrate from compartment 1 into compartment 2 by using $\frac{K_{12}}{q_1}$ as the transfer rate. The equations that describe this system are

$$\begin{aligned}\dot{q}_1 &= -K_{12} - \left(K_{10} + \frac{V_{\max}}{(q_1 + K_M)}\right) q_1 + K_{21} q_2 + f_1(t) \\ \dot{q}_2 &= K_{12} - K_{20} q_2^2 + f_2(t)\end{aligned} \tag{8.67}$$

Note that the constant loss from compartment 1 and input to compartment 2 could describe an active pump. Also note that a chemical reaction in compartment 2 consumes the substrate with rate K_{20}. We will expand on these concepts with multicompartment models in the next section that includes diffusion.

8.4 DIFFUSION, BIOCHEMICAL REACTIONS, AND ENZYME KINETICS

In Chapter 7, we discussed diffusion as a flow of ions down the concentration gradient. Up to now, we have approached biochemical reactions and enzyme kinetics occurring in a homogeneous volume. Now, we include diffusion from another compartment, biochemical reactions, and enzyme kinetics in our analysis. As we will see, the movement of a substrate or an enzyme into the cell by diffusion allows a product to be created. This product then can be used inside the cell or diffused out of the cell to be used by another cell or tissue. Additionally, the same situation occurs within the organelles of the cell. These reactions can serve a regulatory role as well as accelerating biochemical reactions; recall the reaction involving ADP and ATP in the mitochondria, where the availability of ADP regulates the production of ATP.

8.4.1 Diffusion and Biochemical Reactions

Consider the movement of a substrate A into a cell by diffusion, which then reacts with B to form product P, as shown in Figure 8.15. Product P then leaves the cell by diffusion. Assume that the quantity of B_i is regulated by another system. Let subscript i denote inside the cell, and let o denote outside the cell. The chemical reaction is

$$A_i + B_i \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} P_i$$

$$\begin{aligned} \dot{q}_{A_i} &= -K_1 q_{A_i} q_{B_i} + K_{-1} q_{P_i} \\ \dot{q}_{P_i} &= K_1 q_{A_i} q_{B_i} - K_{-1} q_{P_i} \end{aligned} \tag{8.68}$$

where K_1 and K_{-1} are the reaction rates, q_{A_o} and q_{A_i} are the quantities of substrate A, q_{B_i} is the quantity of substrate B_i, and q_{P_o} and q_{P_i} are the quantities of product P. Diffusion across the membrane is given by

$$\begin{aligned} \dot{q}_{A_i} &= C_{oi} q_{Ao} - C_{io} q_{A_i} \\ \dot{q}_{P_i} &= D_{oi} q_{P_o} - D_{io} q_{P_i} \end{aligned} \tag{8.69}$$

where C_{oi}, C_{io}, D_{oi}, and D_{io} are the diffusion transfer rates that depend on the volume, as described in Example Problem 7.5. The equations describing the complete system (biochemical reaction and diffusion) are

$$\begin{aligned} \dot{q}_{A_o} &= -C_{oi} q_{A_o} + C_{io} q_{A_i} \\ \dot{q}_{A_i} &= -K_1 q_{A_i} q_{B_i} + K_{-1} q_{P_i} + C_{oi} q_{A_o} - C_{io} q_{A_i} \\ \dot{q}_{P_o} &= -D_{oi} q_{P_o} + D_{io} q_{P_i} \\ \dot{q}_{P_i} &= K_1 q_{A_i} q_{B_i} - K_{-1} q_{P_i} + D_{oi} q_{P_o} - D_{io} q_{P_i} \end{aligned} \tag{8.70}$$

Transport of Oxygen into a Slow Muscle Fiber

Consider the delivery of oxygen into a slow muscle fiber. Let's begin at the lung where oxygen first diffuses through the alveoli membrane into the capillaries and from the capillaries into the red blood cell, as shown in Figure 8.16. Let q_{O_A} be the quantity of O_2 in the

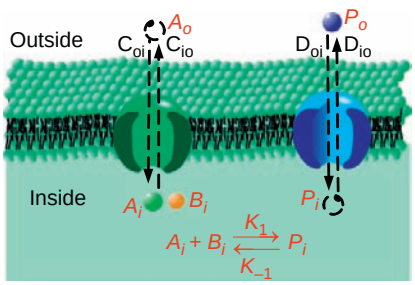


FIGURE 8.15 Diffusion and a biochemical reaction.

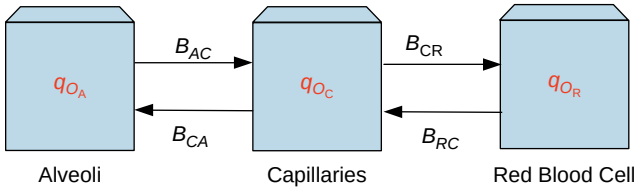


FIGURE 8.16 The diffusion of oxygen from the alveoli into the red blood cell.

alveoli, q_{O_C} be the quantity of O_2 in the capillaries, and q_{O_R} be the quantity of O_2 in the red blood cells. The equation that describes the movement of oxygen is given by

$$\begin{aligned}
\dot{q}_{O_A} &= B_{CA}q_{O_C} - B_{AC}q_{O_A} \\
\dot{q}_{O_C} &= B_{AC}q_{O_A} + B_{RC}q_{O_R} - B_{CA}q_{O_C} - B_{CR}q_{O_C} \\
\dot{q}_{O_R} &= B_{CR}q_{O_C} - B_{RC}q_{O_R}
\end{aligned} \tag{8.71}$$

Once inside the red blood cell, oxygen then binds with hemoglobin (Hb), forming oxyhemoglobin (HbO_8). This is a reversible reaction that allows oxygen to be taken up by the red blood cell and released into the tissues. The binding of O_2 with hemoglobin allows 100 times more oxygen in the blood than if it had just dissociated into the blood. The overall chemical reaction is

$$Hb + 4O_2 \rightleftharpoons HbO_8$$

Hemoglobin has four polypeptide subunits (proteins), with each polypeptide subunit attached to a heme group. Each heme group can bind with a molecule of O_2. The four molecules of O_2 that bind to Hb do not simultaneously react with heme groups but occur in four steps, with each step facilitating the next step. Figure 8.17 illustrates the five states of hemoglobin based on the number of O_2 molecules bound to it, ranging from 0 to 4. Let q_{H_0} be the quantity of Hb, q_{H_1} be the quantity of HbO_2, and so on, up to q_{H_4} be the quantity of HbO_8. Equation (8.72) describes the chemical reactions that take place to create the oxyhemoglobin:

$$\begin{aligned}
\dot{q}_{H_0} &= K_{10}q_{H_1} - K_{01}q_{H_0}q_{O_R} \\
\dot{q}_{H_1} &= K_{01}q_{H_0}q_{O_R} + K_{21}q_{H_2} - K_{10}q_{H_1} - K_{12}q_{H_1}q_{O_R} \\
\dot{q}_{H_2} &= K_{12}q_{H_1}q_{O_R} + K_{32}q_{H_3} - K_{21}q_{H_2} - K_{23}q_{H_2}q_{O_R} \\
\dot{q}_{H_3} &= K_{23}q_{H_2}q_{O_R} + K_{43}q_{H_4} - K_{32}q_{H_3} - K_{34}q_{H_3}q_{O_R} \\
\dot{q}_{H_4} &= K_{34}q_{H_3}q_{O_R} - K_{43}q_{H_4}
\end{aligned} \tag{8.72}$$

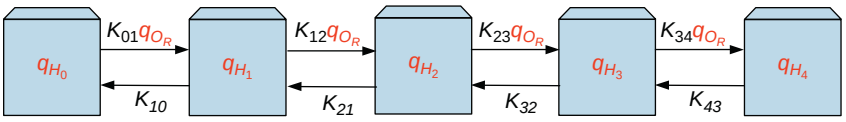


FIGURE 8.17 The five states of hemoglobin.

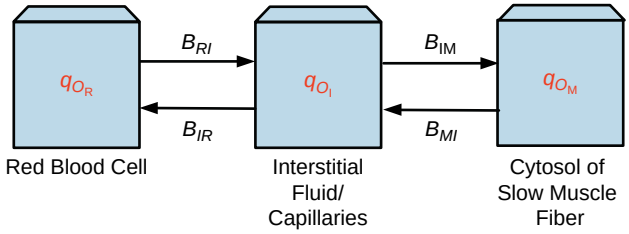


FIGURE 8.18 The diffusion of oxygen from the red blood cell into the cytosol of the slow muscle fiber.

Note that we have assumed that the reverse reactions do not involve oxygen and that oxygen is treated as a molecule and not two oxygen atoms, which would have introduced a q_O^2 term in Eq. (8.72).

Oxygen is transported through the arterial system to the capillaries, where it diffuses out of the red blood cell into the interstitial fluid. It then moves into the cytosol (the liquid part of the cytoplasm that does not contain any organelles) of slow muscle fibers, as shown in Figure 8.18, where q_{O_I} is the quantity of O_2 in the interstitial fluid, q_{O_M} is the quantity of O_2 in the cytosol of the slow muscle fiber, and the other quantities are defined as before. The equation describing the diffusion process and reactions with Hb is given by

$$\begin{aligned}
\dot{q}_{O_R} &= B_{IR}q_{O_C} - B_{RI}q_{O_R} + K_{10}q_{H_1} - K_{01}q_{H_0}q_{O_R} + K_{21}q_{H_2} - K_{12}q_{H_1}q_{O_R} \\
&\quad + K_{32}q_{H_3} - K_{23}q_{H_2}q_{O_R} + K_{43}q_{H_4} - K_{34}q_{H_3}q_{O_R} \\
\dot{q}_{O_I} &= B_{RI}q_{O_R} + B_{MI}q_{O_M} - B_{IR}q_{O_I} - B_{IM}q_{O_I} \\
\dot{q}_{O_M} &= B_{IM}q_{O_I} - B_{MI}q_{O_M}
\end{aligned} \tag{8.73}$$

On the arterioles side of the capillary membrane, PO_2 is approximately 100 mm Hg and 98 percent saturated. On the venule side of the capillary membrane, PO_2 is approximately 40 mm Hg and 75 percent saturated. When PO_2 is high, oxygen quickly binds with hemoglobin, and when PO_2 is low, oxygen is quickly released from hemoglobin. As the red blood cell moves through the capillary, the PO_2 gradient causes oxygen to be released into the interstitial fluid.

Once inside the slow muscle fiber, oxygen is moved to the mitochondria using a different mechanism than that used in other cells. After oxygen diffuses across the cell membrane, it quickly binds to myoglobin (Mb), a protein-like hemoglobin whose function is to transport and store oxygen, and forms oxymyoglobin (MbO_2). By storing oxygen in oxymyoglobin, oxygen is driven into the cell by a large concentration gradient until it binds with all available myoglobin. At this point, the oxygen concentration on either side of the membrane equilibrates. Slow muscle fibers also have many more mitochondria than other cells, which allows higher levels of oxidative metabolism. Thus, the muscle fiber is able to store a large quantity of oxygen in oxymyoglobin, and when needed, it is readily available to the mitochondria to create ATP. This greatly increases oxygen transport to the mitochondria than if the cell just depended on oxygen diffusion in the absence of myoglobin.

When oxygen is bound to myoglobin to create oxymyoglobin in the cytosol, oxymyoglobin then diffuses from the cytosol into the mitochondria, and once in the mitochondria, a

reverse reaction occurs, releasing O_2 and Mb. Oxygen in the mitochondria is consumed with sugar to create ATP during cell respiration, as described in the next section. The myoglobin then diffuses back to the cytosol, where the process repeats itself. The first reaction in the cytosol is given by

$$O_2 + Mb \xrightarrow{D_1} MbO_2$$

and in the mitochondria

$$MbO_2 \xrightarrow{D_2} Mb + O_2$$

which is described by the following equations that includes diffusion:

$$\begin{aligned}
\dot{q}_{O_M} &= -D_1 q_{O_M} q_{Mb_M} + B_{IM} q_{O_I} \\
\dot{q}_{MbO_M} &= D_1 q_{O_M} q_{Mb_M} - K_{MT} q_{MbO_M} \\
\dot{q}_{Mb_M} &= -D_1 q_{O_M} q_{Mb_M} + K_{TM} q_{Mb_T} \\
\dot{q}_{Mb_T} &= D_2 q_{MbO_T} - K_{TM} q_{Mb_T} \\
\dot{q}_{MbO_T} &= -D_2 q_{MbO_T} + K_{MT} q_{MbO_M}
\end{aligned} \tag{8.74}$$

where K_{MT} and K_{TM} are the diffusion transfer rates from the cytosol into the mitochondria and the mitochondria into the cytosol, respectively; q_{O_M} is the quantity of O_2 inside the cytosol; q_{Mb_M} is the quantity of Mb in the cytosol; q_{MbO_M} is the quantity of MbO_2 in the cytosol; q_{MbO_T} is the quantity of MbO_2 inside the mitochondria; and q_{Mb_T} is the quantity of Mb inside the mitochondria. We assume that none of the MbO_2 or Mb leaves the cell. We assume that no O_2 leaves the cell and no Mb_i diffuses into the mitochondria.

Carbon dioxide is created in the cells during cell respiration, as discussed shortly. It then diffuses out of the cell into the interstitial fluid and then moves to the capillaries. Once inside the capillaries, 90 percent of the carbon dioxide moves into the red blood cell, and 10 percent dissolves into the fluid of the blood. Carbon dioxide is then transported to the lungs.

When inside the red blood cell, carbon dioxide almost instantaneously goes through the following reactions:

1. Approximately 70 percent of the carbon dioxide reacts with water to form carbonic acid (H_2CO_3), using the enzyme carbonic anhydrase, and then dissociates into hydrogen (H^+) and biocarbonate $\left(HCO_3^-\right)$ ions. These reactions occur within a fraction of a second and are given by

$$CO_2 + H_2O \xrightarrow{carbonic\ anhydrase} H_2CO_3 \longrightarrow HCO_3^- + H^+$$

2. The hydrogen then binds with the hemoglobin in the red blood cell.

$$H^+ + HbH^- \rightarrow (H)Hb$$

3. The biocarbonate diffuses out of the red blood cell, replaced by chloride ions via a biocarbonate-chloride carrier protein in the cell membrane.

4. The remaining 20 percent of the carbon dioxide in the red blood cell combines with hemoglobin to form carbaminohemoglobin. This is a weak bond that is easily broken:

$$Hb + CO_2 \rightarrow HbCO_2$$

After the red blood cell reaches the lungs, the oxygen that diffused across the alveoli membrane displaces the carbon dioxide in the blood and binds with the hemoglobin. Carbon dioxide then diffuses through the alveoli membrane and is then exhaled. The entire process then repeats itself.

8.4.2 Diffusion and Enzyme Kinetics

Consider the movement of a substrate into a cell by diffusion, which then reacts with an enzyme to ultimately form a product that leaves the cell by diffusion, as shown in Figure 8.19. The chemical reaction is

$$S_i + E \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} ES_i \xrightarrow{K_2} E + P_i \tag{8.75}$$

and diffusion by

$$\begin{aligned} \dot{q}_{S_i} &= B_{oi} q_{S_o} - B_{io} q_{S_i} \\ \dot{q}_{P_i} &= D_{oi} q_{P_o} - D_{io} q_{P_i} \end{aligned} \tag{8.76}$$

where B_{oi}, B_{io}, D_{oi}, and D_{io} are the diffusion transfer rates. The equations describing the complete system are

$$\begin{aligned} \dot{q}_{S_i} &= -K_1 q_{S_i} q_E + K_{-1} q_{ES_i} + B_{oi} q_{S_o} - B_{io} q_{S_i} \\ \dot{q}_{ES_i} &= K_1 q_{S_i} q_E - (K_{-1} + K_2) q_{ES_i} \\ \dot{q}_{P_i} &= K_2 q_{ES_i} + D_{oi} q_{P_o} - D_{io} q_{P_i} \end{aligned} \tag{8.77}$$

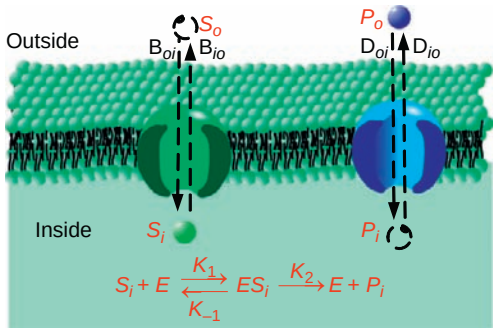


FIGURE 8.19 Diffusion and enzyme kinetics.

To remove q_{S_o} in Eq. (8.77), we assume a constant total substrate $q_{ST} = q_{S_i} + q_{S_o}$, and with $q_{S_o} = q_{ST} - q_{S_i}$, we have

$$\begin{aligned}
\dot{q}_{S_i} &= -K_1 q_{S_i} q_E + K_{-1} q_{ES_i} + B_{oi}(q_{ST} - q_{Si}) - B_{io} q_{S_i} \\
&= -K_1 q_{S_i} q_E + K_{-1} q_{ES_i} - (B_{oi} + B_{io}) q_{S_i} + B_{oi} q_{ST} \\
\dot{q}_{ES_i} &= K_1 q_{S_i} q_E - (K_{-1} + K_2) q_{ES_i} \\
\dot{q}_{P_i} &= K_2 q_{ES_i} + D_{oi} q_{P_o} - D_{io} q_{P_i}
\end{aligned} \tag{8.78}$$

We can substitute the quasi-steady-state approximation, $-K_1 q_{S_i} q_E + K_{-1} q_{ES_i} = -\frac{V_{\max}}{(q_{S_i} + K_M)} q_{S_i}$ (based on Eq. (8.47)), into Eq. (8.78) and get

$$\begin{aligned}
\dot{q}_{S_i} &= -\frac{V_{\max}}{(q_{S_i} + K_M)} q_{S_i} - (B_{oi} + B_{io}) q_{S_i} + B_{oi} q_{ST} \\
&= -\left(\frac{V_{\max}}{(q_{S_i} + K_M)} + B_{oi} + B_{io} \right) q_{S_i} + B_{oi} q_{ST}
\end{aligned} \tag{8.79}$$

where $q_{ES_i} = \frac{q_E(0)}{\left(1 + \frac{K_M}{q_{S_i}}\right)}$ and $q_{P_i} = q_{S_i}(0) - q_{S_i} - q_{ES_i}$.

8.4.3 Carrier-Mediated Transport

Now consider carrier-mediated transport, where an enzyme carrier in the cell membrane has a selective binding site for a substrate, which, when bound, transports the substrate through the membrane to be released inside the cell. Many also refer to this process as facilitated diffusion. Carrier-mediated transport does not use energy to transport the substrate but depends on the concentration gradient. Without carrier-mediated transport, the substrate cannot pass through the membrane.

Carrier-mediated transport differs from diffusion, since it is capacity-limited and diffusion is not. That is, as the quantity of the substrate increases, the carrier-mediated transport reaction rate increases and then saturates, where regular diffusion increases linearly without bound, as shown in Figure 8.20.

Figure 8.21 illustrates carrier-mediated transport, described by

$$\begin{aligned}
&S_o + C_o \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} P_o \underset{K_2}{\overset{K_2}{\rightleftarrows}} P_i \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} C_i + S_i \\
&C_o \underset{K_2}{\overset{K_2}{\rightleftarrows}} C_i
\end{aligned} \tag{8.80}$$

where S_o and S_i are the substrate outside and inside the cell, C_o is the carrier on the outside of the membrane, C_i is the carrier on the inside of the membrane, P_o is the bound substrate and carrier complex on the outside of the membrane, and P_i is the bound substrate and

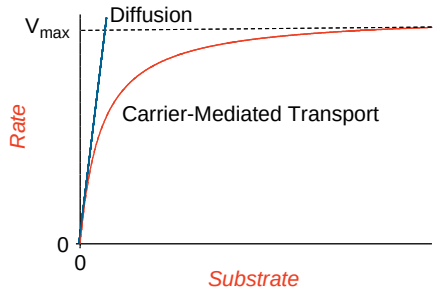


FIGURE 8.20 Comparison of the effect of the amount of substrate and the rate of diffusion and carrier-mediated transport.

carrier complex on the inside of the membrane. As shown in Figure 8.21, the substrate first moves to the binding site on the outside of the membrane and binds with the carrier to form the carrier-substrate complex, P_o (left). Next, the carrier-substrate complex moves from the outside to the inside of the membrane, given by the carrier-substrate complex P_i (right). It is not known how the carrier-substrate complex moves through the membrane, but we are fairly certain it happens.

The last step is when the carrier-substrate complex, P_i, dissociates into the substrate, S_i, and carrier, C_i. We have assumed that the reaction rates are the same for creation and the dissociation of the complex, $\underset{K_2}{\overset{K_2}{\rightleftharpoons}}$. In addition, we assume that the reaction of

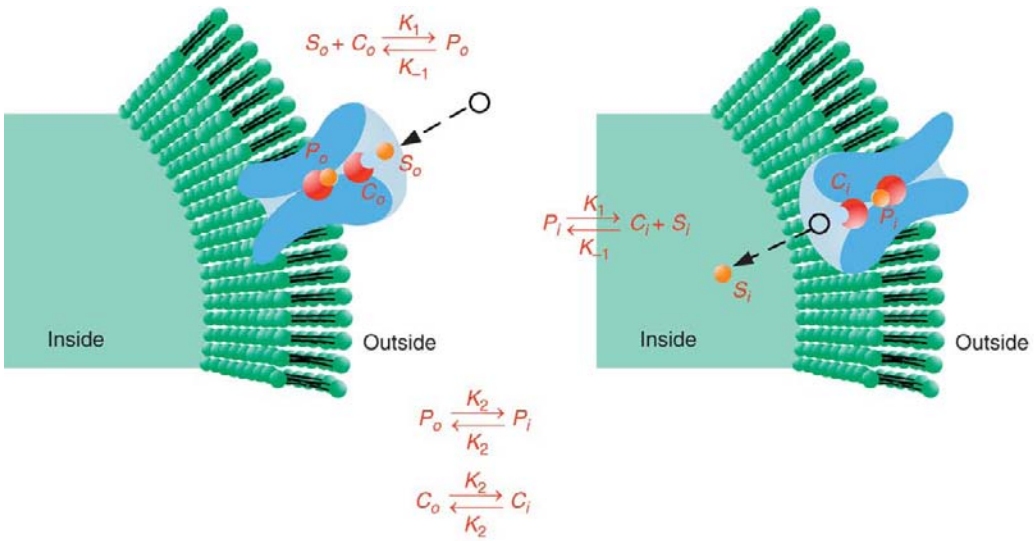


FIGURE 8.21 Carrier-mediated transport, with the carrier within the membrane. On the left side is the binding of the carrier with the substrate. On the right side is the substrate released from the binding site into the cell.

carrier-substrate complex from outside to inside has the same reaction rate, $\underset{K_2}{\overset{K_2}{\rightleftarrows}}$. Using the law of mass action, we get

$$\begin{aligned}
\dot{q}_{S_o} &= -K_1 q_{S_o} q_{C_o} + K_{-1} q_{P_o} \\
\dot{q}_{C_o} &= -K_1 q_{S_o} q_{C_o} + K_{-1} q_{P_o} + K_2 q_{C_i} - K_2 q_{C_o} \\
\dot{q}_{P_o} &= K_1 q_{S_o} q_{C_o} - (K_{-1} + K_2) q_{P_o} + K_2 q_{P_i} \\
\dot{q}_{P_i} &= -(K_1 + K_2) q_{P_i} + K_2 q_{P_o} + K_{-1} q_{S_i} q_{C_i} \\
\dot{q}_{S_i} &= -K_{-1} q_{S_i} q_{C_i} + K_1 q_{P_i} \\
\dot{q}_{C_i} &= -K_{-1} q_{S_i} q_{C_i} + K_1 q_{P_i} + K_2 q_{C_o} - K_2 q_{C_i}
\end{aligned} \tag{8.81}$$

Since the carrier is not consumed in the reaction, then the total carrier is a constant, given as $q_{C_o} + q_{C_i} + q_{P_o} + q_{P_i} = \zeta$.

Naturally, we can add an input to the system in Eq. (8.81) or simplify using the quasi-steady-state approximation as before. In addition, the substrate can be involved in other reactions inside the cell, such as moving into an organelle (mitochondria) via diffusion and then experiencing an enzyme reaction.

Glucose Transport

Consider the transport of glucose across the cell membrane. We know that glucose does not diffuse across the cell membrane but is transported across the cell membrane by a carrier-mediated transport process. Glucose binds to a protein that transports it across the membrane, allowing it to pass into the cytosol. This process does not use any energy. Using the model illustrated in Figure 8.21, we have

$$\begin{aligned}
&G_0 + C_0 \xrightarrow{K_1} P_o \underset{K_2}{\overset{K_2}{\rightleftarrows}} P_i \xrightarrow{K_1} G_i + C_i \\
&C_o \underset{K_2}{\overset{K_2}{\rightleftarrows}} C_i
\end{aligned} \tag{8.82}$$

where G_o and G_i is glucose outside and inside the cell, respectively; C_o is the carrier on the outside of the membrane; C_i is the carrier on the inside of the membrane; P_o is the bound substrate and carrier complex on the outside of the membrane; and P_i is the bound substrate and carrier complex on the inside of the membrane. Note that there is no reverse reaction for glucose in Eq. (8.82), since glucose does not leave the cell. We assume that glucose is consumed at a constant rate J_i inside the cell during cell respiration (described in the next section) and that glucose is available in the interstitial fluid at a rate of J_o. The equations that describe this system are given by

$$\begin{aligned}
\dot{q}_{G_o} &= -K_1 q_{G_o} q_{C_o} + J_o \\
\dot{q}_{C_o} &= -K_1 q_{G_o} q_{C_o} + K_2 q_{C_i} - K_2 q_{C_o} \\
\dot{q}_{P_o} &= K_1 q_{G_o} q_{C_o} - K_2 q_{P_o} + K_2 q_{P_i} \\
\dot{q}_{G_i} &= K_1 q_{P_i} - J_i \\
\dot{q}_{C_i} &= K_1 q_{P_i} + K_2 q_{C_o} - K_2 q_{C_i} \\
\dot{q}_{P_i} &= K_2 q_{P_o} - K_1 q_{P_i}
\end{aligned} \tag{8.83}$$

As described earlier, transport of glucose through the cell membrane is a capacity-limited reaction because of the enzyme carrier. Glucose concentration in the blood varies from a typical value of $90\frac{\text{mg}}{100\text{ mL}}$, up to $200\frac{\text{mg}}{100\text{ mL}}$ after eating, and down to $40\frac{\text{mg}}{100\text{ mL}}$ three hours after eating. Thus, the input, J_o, is a function of eating. The body uses the following two mechanisms to control glucose concentration:

1. Automatic feedback involving insulin secretion by the pancreas
2. The liver

The pancreas, in addition to digestive functions, secretes insulin directly into the blood. Insulin facilitates diffusion of glucose across the cell membrane. The rate of insulin secretion is regulated so glucose is maintained at a constant level.

The liver acts as a storage vault for glucose. When excess amounts of glucose are present, almost immediately two-thirds is stored in the liver. Conversely, when the glucose level in the blood falls, the stored glucose in the liver replenishes the blood glucose. Insulin has a moderating effect on the function of the liver. The binding of the carrier enzyme with glucose is a function of the transfer rate, K_1, and is a function of the insulin level, which can increase the transport of glucose by as much 20 times the base rate without insulin. We will ignore the aspects of insulin and liver storage in our model.

8.4.4 Active Transport

Now consider active transport, which is similar to carrier-mediated transport but operates against the concentration gradient and uses energy to move the substrate across the cell membrane. Active transport uses an enzyme carrier in the cell membrane that has a selective binding site for a substrate, which, when bound, transports the substrate through the membrane to be released inside the cell.

Active transport uses energy to run, typically by the hydrolysis of ATP. Here, the energy from ATP is used in the transport of the substrate, leaving ADP and an inorganic phosphate $\left(PO_4^{-2}\right)$ in the cytosol. The ADP is then recycled in the mitochondria to create more ATP using glucose as described in the next section. Active transport is capacity-limited like carrier-mediated diffusion: as the quantity of the substrate increases, the transport reaction rate increases and then saturates, as shown in Figure 8.20.

The Na-K pump is the most important active transport process, which pumps Na^+ out of the cell against the concentration gradient and replaces it inside the cell with K^+ against the concentration gradient. This pump is used to maintain the ion gradients and resting membrane potential, as described in Chapter 12, and is also required to maintain cell volume, as described in the previous chapter.

Another important active transport process is the Na-Ca ATP-ase pump that keeps Ca^{+2} levels low inside the cell. It is vitally important that the concentration of Ca^{+2} be kept low inside the cell $\left(\text{approximately } 10^{-7}\text{M}\right)$ as compared to the outside concentration $\left(\text{approximately } 10^{-3}\text{M}\right)$. The concentration gradient drives Ca^{+2} into the cell, and the Na-Ca ATP-ase pump drives Ca^{+2} out of the cell.

Na-K Pump

The Na-K pump is an integral part of the cell membrane that exists in all cells in the body. Approximately 70 percent of all ATP in the neuron and 25 percent of all ATP in all other cells

is used to fuel the Na-K pump at rest. This pump is vital to maintain the cell's resting membrane potential. In this section, we focus on the enzyme reactions. The Na-K pump was discovered by Jens Skou in 1957, who subsequently received a Nobel Prize for his work in 1997. Using radioactive ions, Skou showed that the concentrations of the ions are interdependent, implying the involvement of a common mechanism using an ATP-ase carrier.

The overall reaction for the Na-K pump is given by

$$ATP + 3Na_i^+ + 2K_o^+ \xrightarrow[ATP-ase]{} ADP + P_i + 3Na_o^+ + 2K_i^+$$

where the subscripts i for inside the cell and o for outside the cell are used as before.

The pump has 3 Na^+ binding sites and 2 K^+ binding sites in its two conformations. There is a higher concentration of Na^+ outside the cell than inside and a higher concentration of K^+ inside the cell than outside; left unchecked by the pump, this gradient would drive Na^+ into the cell and K^+ out of the cell, and thus change the resting membrane potential. Any change in the concentration gradient of K^+ and Na^+ is prevented by the Na-K pump. The pump transports a steady stream of Na^+ out of the cell and K^+ into the cell.

The Na-K pump uses six steps to move 3 Na^+ ions out of the cell and 2 K^+ ions into the cell at a total cost of 1 ATP molecule. Figure 8.22 illustrates the six steps that are continually repeated at a rate of 100 s^{-1}:

1. Three Na_i^+ ions in the cytosol move into and bind to the carrier in the pump, $(Na_3C)_i$. Note that the pump has a bound ATP molecule.
2. ATP is hydrolyzed. ADP is released, and the inorganic phosphate, P, binds with $(Na_3C)_i$ to create $(Na_3CP)_i$.
3. Using the energy gained by the hydrolyzation, a conformational change occurs in the pump that moves $(Na_3CP)_o$ to the outside of the cell membrane, exposing Na_o^+ ions to the outside. The 3 Na_o^+ ions exit the pump.
4. On the outside of the cell, 2 K_o^+ ions then bind to the carrier and inorganic phosphate in the pump, creating $(K_2CP)_o$.
5. Dephosphorylation of the pump occurs, releasing the inorganic phosphate. Following this, a conformational change occurs in the pump that moves $(K_2C)_i$, exposing 2 K_i^+ ions to the inside of the cell.
6. ATP binds to the pump, and the 2 K_i^+ ions are released into the cell.

The following equations list the reactions that describe the Na-K pump, where C is the carrier and P is the inorganic phosphate. Note that we have eliminated the reverse reactions as they are relatively small.

$$\xrightarrow{K_1} ATP_i$$
$$3Na_i^+ + C_i \xrightarrow{K_2} (Na_3C)_i$$
$$(Na_3C)_i + ATP_i \xrightarrow{K_3} (Na_3CP)_i + ADP_i$$
$$(Na_3CP)_i \xrightarrow{K_4} (Na_3CP)_o \xrightarrow{K_5} 3Na_o^+ + (CP)_o$$
$$(CP)_o + 2K_o^+ \xrightarrow{K_6} (K_2CP)_o \xrightarrow{K_7} P_i + (K_2C)_i$$
$$(K_2C)_i \xrightarrow{K_8} 2K_i^+ + C_i$$
$$ADP_i \xrightarrow{K_9}$$

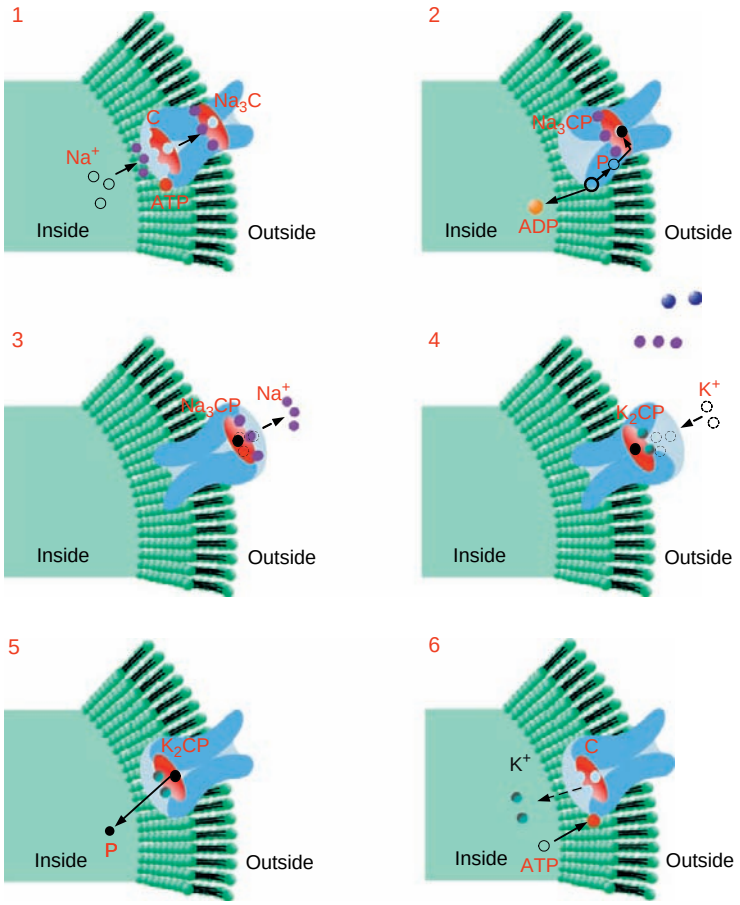


FIGURE 8.22 The six steps that characterize the Na-K pump.

Using the law of mass action, the Na-K pump is characterized by the following set of differential equations:

$$
\begin{aligned}
\dot{q}_{Na_i^+} &= J_{Na_D} - K_2 q_{Na_i^+}^3 q_{C_i} \\
\dot{q}_{(Na_3C)_i} &= K_2 q_{Na_i^+}^3 q_{C_i} - K_3 q_{(Na_3C)_i} q_{ATP_i} \\
\dot{q}_{ATP_i} &= K_1 - K_3 q_{(Na_3C)_i} q_{ATP_i} \\
\dot{q}_{(Na_3CP)_i} &= K_3 q_{(Na_3C)_i} q_{ATP_i} - K_4 q_{(Na_3CP)_i} \\
\dot{q}_{(Na_3CP)_o} &= K_4 q_{(Na_3CP)_i} - K_5 q_{(Na_3CP)_o} \\
\dot{q}_{Na_o^+} &= K_5 q_{(Na_3CP)_o} - J_{Na_D} \\
\dot{q}_{(CP)_o} &= K_5 q_{(Na_3CP)_o} - K_6 q_{K_o^+}^2 q_{(CP)_o} \\
\dot{q}_{K_o^+} &= J_{K_D} - K_6 q_{K_o^+}^2 q_{(CP)_o}
\end{aligned}
\tag{8.84}
$$

$$\begin{aligned}
\dot{q}_{(K_2CP)_o} &= K_7 q_{(K_2CP)_o} \\
\dot{q}_{P_i} &= K_7 q_{(K_2CP)_o} \\
\dot{q}_{(K_2C)_i} &= K_7 q_{(K_2CP)_o} - K_8 q_{(K_2C)_i} \\
\dot{q}_{ADP_i} &= -K_9 q_{ADP_i} \\
\dot{q}_{C_i} &= K_8 q_{(K_2C)_i} - K_2 q_{C_i} q_{Na_i^+}^3 \\
\dot{q}_{K_i^+} &= K_8 q_{(K_2C)_i} - J_{K_D}
\end{aligned}$$

where we assume a flow of K^+ out of the cytosol at a rate of J_{K_D} due to diffusion, a flow of Na^+ into the cell at a rate of J_{Na_D} from diffusion, ATP into the cytosol from the mitochondria at a rate of K_1, and ADP into the mitochondria from the cytosol at a rate of K_9. Note that there are other models that describe the Na-K pump based on different assumptions.

8.5 CELLULAR RESPIRATION: GLUCOSE METABOLISM AND THE CREATION OF ATP

At this time we wish to discuss cellular respiration involving glucose metabolism in more detail by describing the chemical processes that enable the body to create energy in the form of ATP. As previously noted, ATP is the fuel that supports life's processes, where energy is stored in the inorganic phosphate bonds. Energy is required to form these bonds in ATP, and energy is released when the bonds are broken.

Cellular respiration is among the best-known metabolic pathways, with detailed models going back to the 1960s. While much progress has been made, much uncertainty exists, and the system is still under considerable research. Here we will provide sufficient coverage based on the techniques in this chapter and focus on aerobic respiration that uses oxygen to create ATP in the mitochondria.

Cellular respiration consists of three major steps: glycolysis, Krebs cycle, and the electron transport chain, as illustrated in Figure 8.23. The first step occurs in the cytosol, and the two other steps occur in the mitochondria. Each step is very complex, with an overall reaction given by

$$C_6H_{12}O_6 + 6O_2 + 36ADP + 36P \longrightarrow 6CO_2 + 6H_2O + 36ATP \tag{8.85}$$

which consists of the oxidation of glucose $(C_6H_{12}O_6)$ to CO_2 and O_2, and a reduction of O_2 to H_2O. Once inside the cell, 1 M of glucose is transformed into 36 M of ATP by adding an inorganic phosphate P to each ADP. While there is some adenosine monophosphate (AMP) that, when combined with two inorganic phosphates, creates ATP, we will ignore this part of the reaction in this section.

Mitochondria are the powerhouses of the cell that produces the ATP used by the cell. A mitochondrion has two lipid bilayer membranes. The outer membrane is smooth and acts like a typical membrane. The inner membrane has many infoldings that contain the oxidative enzymes. Within the inner membrane is a cavity called the matrix that contains a nucleus with DNA, ribosomes, and dissolved enzymes. The matrix enzymes work with the oxidative enzymes to create ATP. The mitochondria's nucleus allows it to self-replicate

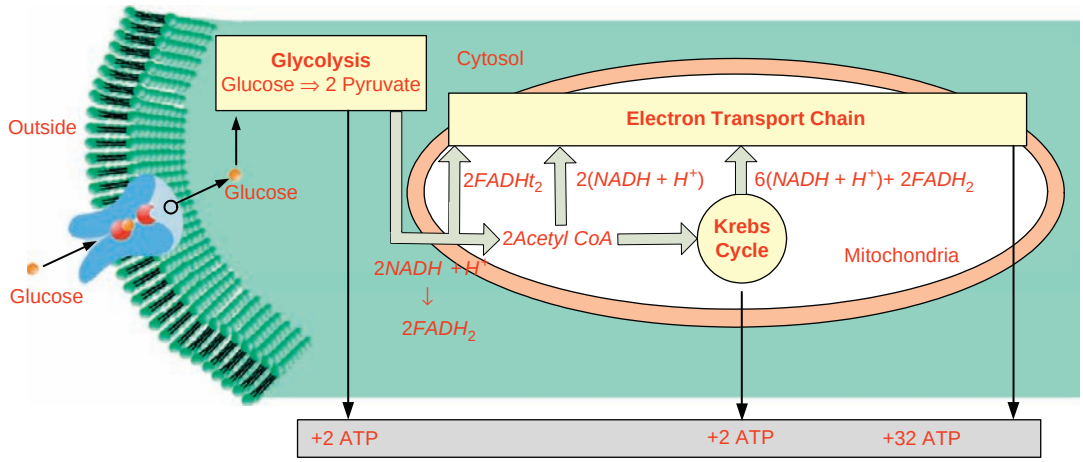


FIGURE 8.23 Cellular respiration consisting of three steps: glycolysis, Krebs cycle, and the electron transport chain. The amount of ATP created or lost in each step is indicated at the bottom of the figure. A total of 36 molecules of ATP are created from 1 molecule of glucose.

based on the needs of the cell for ATP. If there is an increased need for ATP, more mitochondria are produced.

We call the process of adding an inorganic phosphate to a molecule phosphorylation. We call the use of O_2 in the phosphorylation process oxidative phosphorylation. While the overall chemical reaction in Eq. (8.85) looks simple enough, it consists of many enzyme catalyzed reactions.

A carrier-mediated transport process is the mechanism postulated to move glucose into the cell from the interstitial fluid. Ten enzyme catalyzed reactions take place in the cytosol during glycolysis,[3] where a glucose molecule is broken down into two pyruvate $(C_3H_4O_3)$ molecules in which energy is consumed and produced. The overall reaction in glycolysis is given by

$$\begin{aligned} C_6H_{12}O_6 + 2NAD^+ + 2ADP + 2P \longrightarrow \\ 2C_3H_4O_3 + 2(NADH + H^+) + 2ATP + 2H_2O \end{aligned} \tag{8.86}$$

where NAD^+ is nicotinamide adenine dinucleotide and when combined with two hydrogen, forms $NADH + H^+$. During the electron transport chain, each molecule of $NADH + H^+$ is used to create three molecules of ATP.

The next two steps occur in the mitochondria. The second step begins with the conversion of pyruvate, $C_3H_4O_3$, into acetyl coenzyme A, followed by eight more major enzyme catalyzed reactions called the Krebs cycle.[4] The overall result of the Krebs cycle is the

[3]Glycolysis is also known as the Embden-Meyerhoff pathway.

[4]The Krebs cycle is also known as the citric acid and tricarboxylic acid cycle.

conversion of pyruvate into CO_2 and $NADH + H^+$. The overall reaction for the Krebs cycle is given by

$$\begin{aligned} &Acetyl\text{-}CoA + 3NAD^+ + FAD + 2H_2O + ADP + P \longrightarrow \\ &\quad CoA - SH + 2CO_2 + FADH_2 + ATP + 3(NADH + H^+) \end{aligned} \tag{8.87}$$

where FAD is flavin adenine dinucleotide, and when combined with two hydrogen forms $FADH_2$. During the electron transport chain, each molecule of $FADH_2$ is used to create two molecules of ATP.

The last step in cellular respiration is the electron transport chain.[5] Each of the large number of enzyme catalyzed reactions leading up to the electron transport chain contributes only two molecules of ATP, four molecules of $FADH_2$, and eight molecules of $NADH + H^+$. During the electron transport chain, $NADH - H^+$ and $FADH_2$ are used to create ATP by the oxidation of hydrogen atoms. Almost all of the energy created in cellular respiration occurs in the electron transport chain in the metabolism of glucose (32 molecules of ATP from one molecule of glucose). The overall reaction for the electron transport chain is given by

$$(NADH + H^+) + \frac{1}{2}O_2 + 3ADP + 3P \longrightarrow NAD^+ + 4H_2O + 3ATP$$

and

$$FADH_2 + \frac{1}{2}O_2 + 2ADP + 2P \longrightarrow FAD + 3H_2O + 2ATP \tag{8.88}$$

The analysis in this section illustrates some of the major steps in converting glucose into ATP. To maximize the amount of energy created from glucose, many reactions occur, rather than a direct reaction of glucose into one ATP, water, and carbon dioxide. Using a series of reactions in the cytosol and the mitochondria, a total of 36 molecules of ATP are created from one molecule of glucose. We focus more detail on glycolysis and the Krebs cycle and a broader treatment of the electron transport chain. We assume that the process of transporting glucose into the cell occurs via a carrier-mediated transport and begins the glycolysis of glucose in the cytosol.

8.5.1 Glycolysis

Glycolysis involves 10 enzyme catalyzed reactions in the cytosol, which transform one glucose molecule into two molecules of pyruvic acid, as shown in Figure 8.24. The intermediate complex reactions of the substrates and enzymes are not shown in Figure 8.24 for simplicity. However, they are included in the differential equations using a reaction rate constant for the intermediate complex given by B_i that corresponds to the reaction rate K_i. As shown in Figure 8.23, a net gain of 2 molecules of ATP and $2(NADH + H+)$ for each glucose molecule is seen during glycolysis. The major output of glycolysis is the creation of 2 pyruvate molecules from one molecule of glucose.

[5]The electron transport chain is also known as the chemiosmotic mechanism.

FIGURE 8.24 An overview of the 10 steps of glycolysis. Note that in steps 6–10, there are two molecules of everything, since 2 molecules of glyceraldehyde-3-phosphate are created for each glucose molecule. Also note that the intermediate complex reactions are not included in the illustration but are included in the differential equations describing glycolysis. Abreviations are listed next to each molecule.

Step 1 of glycolysis begins with the binding of glucose to a phosphate enzyme called hexokinase, resulting in glucose-6-phosphate. The immediate phosphorylation of glucose is required so glucose does not leave the cell. This reaction requires ATP and is given by

$$\begin{aligned} &G + Hx \xrightarrow{B_1} GHx \\ &GHx + ATP \xrightarrow{K_1} G6P + ADP + H^+ + Hx \\ &\dot{q}_G = -B_1 q_G q_{Hx} + J_G \\ &\dot{q}_{GHx} = B_1 q_G q_{Hx} - K_1 q_{ATP} q_{GHx} \end{aligned} \tag{8.89}$$

where J_G is the flow of glucose into the cell, as described in Section 8.4.3.

The second step involves the reaction of glucose-6-phosphate with the enzyme phosphoglucose isomerase enzyme to create fructose-6-phosphate. This reaction is given by

$$G6P + PI \underset{B_{-2}}{\overset{B_2}{\rightleftharpoons}} G6PPI \underset{K_{-2}}{\overset{K_2}{\rightleftharpoons}} FP6 + PI$$

$$\begin{aligned} \dot{q}_{GP6} &= K_1 q_{ATP} q_{GHx} + B_{-2} q_{G6PPI} - B_2 q_{GP6} q_{PI} \\ \dot{q}_{G6PPI} &= B_2 q_{GP6} q_{PI} + K_{-2} q_{FP6} q_{PI} - (B_{-2} + K_2) q_{G6PPI} \end{aligned} \tag{8.90}$$

The third step involves the reaction of fructose-6-phosphate and ATP with the enzyme phosphofructokinase to create fructose-1,6-bisphosphate and is given by

$$\begin{aligned} &F6P + PF \xrightarrow{B_3} F6PPF \\ &F6PPF + ATP \xrightarrow{K_3} F6BP + ADP + H^+ + PF \\ &\dot{q}_{F6P} = K_2 q_{G6PPI} - K_{-2} q_{F6P} q_{PI} - B_3 q_{F6P} q_{PF} \\ &\dot{q}_{F6PPF} = B_3 q_{F6P} q_{PF} - K_3 q_{F6PPF} q_{ATP} \end{aligned} \tag{8.91}$$

The fourth step involves the reaction of fructose-1,6-bisphosphate with the enzyme aldolase to create dihydroxyacetone-phosphate and glyceraldehyde-3-phosphate and is given by

$$\begin{aligned} &F6BP + ALD \underset{B_{-4}}{\overset{B_4}{\rightleftharpoons}} F6BPALD \underset{K_{-4}}{\overset{K_4}{\rightleftharpoons}} DP + G3P + ALD \\ &\dot{q}_{F6BP} = K_3 q_{FP6\,PF} q_{ATP} + B_{\ 4} q_{F6BPALD} - B_4 q_{F6BP} q_{ALD} \\ &\dot{q}_{F6BPALD} = B_4 q_{F6BP} q_{ALD} + K_{-4} q_{DP} q_{G3P} q_{ALD} - (B_{-4} + K_4) q_{F6BPALD} \end{aligned} \tag{8.92}$$

The fifth step involves the reaction of dihydroxyacetone-phosphate with the enzyme triose phosphate isomerases to create glyceraldehyde-3-phosphate and is given by

$$\begin{aligned} &DP + TPI \underset{B_{-5}}{\overset{B_5}{\rightleftharpoons}} DPTPI \underset{K_{-5}}{\overset{K_5}{\rightleftharpoons}} G3P + TPI \\ &\dot{q}_{DP} = K_4 q_{F6BPALD} + B_{-5} q_{DPTPI} - K_{-4} q_{DP} q_{G3P} q_{ALD} - B_5 q_{DP} q_{TPI} \\ &\dot{q}_{DPTPI} = B_5 q_{DP} q_{TPI} + K_{\ 5} q_{G3P} q_{TPI} - (B_{\ 5} + K_5) q_{DPTPI} \end{aligned} \tag{8.93}$$

The sixth step involves the reaction of glyceraldehyde-3-phosphate with the enzyme glyceraldehyde-3-phosphate dehydrogenase, an inorganic phosphate, and NAD^+ to create 1,3-bisphosphoglycerate and is given by

$$\begin{aligned} &G3P + GPD \underset{B_{-6}}{\overset{B_6}{\rightleftharpoons}} G3PD \\ &G3PD + P + NAD^+ \underset{K_{-6}}{\overset{K_6}{\rightleftharpoons}} BP + (NADH + H^+) + GPD \\ &\dot{q}_{G3P} = K_4 q_{F6BPALD} + K_5 q_{DPTPI} + B_{\ 6} q_{G3PD} - K_{\ 4} q_{DP} q_{G3P} q_{ADL} - K_{\ 5} q_{G3P} q_{TPI} \\ &\qquad - B_6 q_{G3P} q_{GPD} \\ &\dot{q}_{G3PD} = B_6 q_{G3P} q_{GPD} + K_{-6} q_{BP} q_{NADH+H^+} q_{GPD} - B_{-6} q_{G3PD} - K_6 q_{G3PD} q_P q_{NAD^+} \end{aligned} \tag{8.94}$$

The seventh step involves the reaction of 1,3-bisphosphoglycerate with the enzyme phosphoglycerate kinase and ADP to create 3-phosphoglycerate and ATP and is given by

$$\begin{aligned} &BP + PGK \xrightarrow{B_7} BPPGK \\ &BPPGK + ADP \xrightarrow{K_7} P3G + PGK + ATP \\ &\dot{q}_{BP} = K_6 q_{G3PD} - K_{-6} q_{BP} q_{NADH-H^+} q_{GPD} - B_7 q_{BP} q_{PGK} \\ &\dot{q}_{BPPGK} = B_7 q_{BP} q_{PGK} - K_7 q_{BPPGK} q_{ADP} \end{aligned} \tag{8.95}$$

Note that we have assumed only a forward reaction in this step.

The eighth step involves the reaction of 3-phosphoglycerate with the enzyme phosphoglycerate mutase to create 2-phosphoglycerate and is given by

$$\begin{aligned} &P3G + PM \underset{B_{-8}}{\overset{B_8}{\rightleftarrows}} P3GPM \underset{K_{-8}}{\overset{K_8}{\rightleftarrows}} P2G + PM \\ &\dot{q}_{P3G} = K_7 q_{BPPGK} q_{ADP} + B_{-8} q_{P3GPM} - B_8 q_{P3G} q_{PM} \\ &\dot{q}_{P3GPM} = B_8 q_{P3G} q_{PM} + K_{-8} q_{PGM} q_{PM} - (B_{-8} + K_8) q_{P3GPM} \end{aligned} \tag{8.96}$$

The ninth step involves the reaction of 2-phosphoglycerate with the enzyme enolase to create phosphoenolpyruvate and water, and is given by

$$\begin{aligned} &P2G + EN \underset{B_{-9}}{\overset{B_9}{\rightleftarrows}} P2GEN \underset{K_{-9}}{\overset{K_9}{\rightleftarrows}} PP + EN + H_2O \\ &\dot{q}_{P2G} = K_8 q_{P3GPM} + B_9 q_{P2GEN} - K_8 q_{P2G} q_{PM} - B_9 q_{P2G} q_{EN} \\ &\dot{q}_{P2GEN} = B_9 q_{P2G} q_{EN} + K_9 q_{PP} q_{EN} q_{H_2O} - (B_9 + K_9) q_{P2GEN} \end{aligned} \tag{8.97}$$

The last step involves the reaction of phosphoenolpyruvate, ADP, and H^+ with the enzyme pyruvate kinase to create pyruvate and ATP and is given by

$$\begin{aligned} &PP + PK \xrightarrow{B_{10}} PPPK \\ &PPPK + ADP \xrightarrow{K_{10}} PV + ATP + PK \\ &\dot{q}_{PP} = K_9 q_{P2GEN} - K_{-9} q_{PP} q_{EN} q_{H_2O} - B_{10} q_{PP} q_{PK} \\ &\dot{q}_{PPPK} = B_{10} q_{PP} q_{PK} - K_{10} q_{PPPK} q_{ADP} \\ &\dot{q}_{PV=} K_{10} q_{PPPK} q_{ADP} \end{aligned} \tag{8.98}$$

The equations for ATP and ADP in the cytosol are

$$\begin{aligned} \dot{q}_{ATP} &= -K_1 q_{ATP} q_{GHx} - K_3 q_{F6PPF} q_{ATP} + K_7 q_{BPPGK} q_{ADP} + K_{10} q_{PPPK} q_{ADP} + J_{ATP} \\ \dot{q}_{ADP} &= K_1 q_{ATP} q_{GHx} + K_3 q_{F6PPF} q_{ATP} - K_7 q_{BPPGK} q_{ADP} - K_{10} q_{PPPK} q_{ADP} + J_{ADP} \end{aligned} \tag{8.99}$$

where J_{ATP} and J_{ADP} are the net usage of ATP and ADP from other processes, such as the reactions in Sections 8.5.2 and 8.5.3 in the mitochondria and the Na-K pump described in Section 8.4.4.

ADP and ATP Movement in/out of the Mitochondria

Let's assume that the movement of ADP and ATP through the mitochondria's membranes is by diffusion from the cytosol through the outer membrane and carrier-mediated diffusion through the inner membrane. Further, we assume that ADP only enters the mitochondria and that ATP only leaves the mitochondria for simplicity. The equations that describe this transport are given in Eqs. (8.100) and (8.101).

$$
\begin{aligned}
&ADP_c \xrightarrow{D_1} ADP_o \\
&ADP_o + C_o \xrightarrow{B_1} T_o \underset{B_2}{\overset{B_2}{\rightleftarrows}} T_i \xrightarrow{B_1} C_i + ADP_i \\
&C_o \underset{B_2}{\overset{B_2}{\rightleftarrows}} C_i \\
&ATP_i + E_i \xrightarrow{B_3} R_i \underset{B_4}{\overset{B_4}{\rightleftarrows}} R_o \xrightarrow{B_3} E_o + ATP_o \\
&ATP_o \xrightarrow{D_2} ATP_c \\
&E_o \underset{B_5}{\overset{B_5}{\rightleftarrows}} E_i
\end{aligned}
\tag{8.100}
$$

where the subscript c is for the cytosol, o is for the outer membrane, and i is for the matrix; Ds are the diffusivity constants and Bs are the reaction rates; Cs and Es are enzymes; and Rs and Ts are the intermediate complexes. The equations that describe this system are

$$
\begin{aligned}
\dot{q}_{ADP_c} &= -D_1 q_{ADP_c} + J_{ADP_c} \\
\dot{q}_{ADP_o} &= D_1 q_{ADP_c} - B_1 q_{ADP_o} q_{C_o} \\
\dot{q}_{ADP_i} &= B_1 T_i - J_{ADP_i} \\
\dot{q}_{T_o} &= B_1 q_{ADP_o} q_{C_o} - B_2 q_{T_o} + B_2 q_{T_i} \\
\dot{q}_{T_i} &= B_2 q_{T_o} - (B_2 + B_1) q_{T_i} \\
\dot{q}_{C_o} &= -B_1 q_{ADP} q_{C_o} + B_2 q_{C_i} - B_2 q_{C_o} \\
\dot{q}_{C_i} &= B_1 q_{T_i} + B_2 q_{C_o} - B_2 q_{C_i} \\
\dot{q}_{ATP_i} &= -B_3 q_{ATP_i} q_{E_i} + J_{ATP_i} \\
\dot{q}_{R_i} &= B_3 q_{ATP_i} q_{E_i} + B_4 R_o - B_4 R_i \\
\dot{q}_{R_o} &= B_4 q_{R_i} - (B_4 + B_3) q_{R_o} \\
\dot{q}_{ATP_o} &= B_3 q_{R_o} - D_2 q_{ATP_o} \\
\dot{q}_{ATP_c} &= D_2 q_{ATP_o} - J_{ATP_c} \\
\dot{q}_{E_i} &= -B_3 q_{ATP_i} q_{E_i} + B_5 q_{E_o} - B_5 q_{E_i} \\
\dot{q}_{E_o} &= B_3 q_{R_o} + B_5 q_{E_i} - B_5 q_{E_o}
\end{aligned}
\tag{8.101}
$$

where J_{ATP_c} and J_{ATP_i} are the consumption/production of ATP in the cytosol and matrix, and J_{ADP_c} and J_{ADP_i} are the consumption/production of ADP in the cytosol and matrix.

Conversion of Pyruvate to Acetyl CoA

Pyruvate moves into the mitochondria by diffusion. Once inside the mitochondria, an enzyme catalyzed reaction occurs that converts pyruvate (PV) into acetyl coenzyme A (AC) as follows:

$$\begin{aligned} &PV + CoA - SH \xrightarrow{B_0} PVCoA \\ &PVCoA + NAD^+ \xrightarrow{K_0} AC + CoA - SH + CO_2 + (NADH + H^+) \end{aligned} \tag{8.102}$$

$$\begin{aligned} \dot{q}_{PV} &= J_{PV} - B_0 q_{PV} q_{CoA} \\ \dot{q}_{CoA} &= -B_0 q_{PV} q_{CoA} + K_0 q_{PVCoA} q_{NAD^+} \\ \dot{q}_{PVCoA} &= B_0 q_{PV} q_{CoA} - K_0 q_{PVCoA} q_{NAD^+} \\ \dot{q}_{AC} &= K_0 q_{PVCoA} q_{NAD^+} \end{aligned}$$

where CoA-SH (CoA) is coenzyme A, $PVCoA$ is the complex pyruvate dehydrogenase, and J_{PV} is the production of pyruvate given by Eq. (8.98). Keep in mind that from one glucose molecule, two pyruvate molecules are created that pass into the mitochondria. Also note that $2(NADH + H^+)$ are converted into $2FADH_2$ to transfer acetyl coenzyme A across the mitochondrial membrane, thus costing 2 ATP.

8.5.2 Krebs Cycle

The Krebs cycle involves a series of enzyme catalyzed reactions that reduce the acetyl portion of acetyl coenzyme A in the mitochondrial matrix, as shown in Figure 8.25. The Krebs cycle continuously recycles, reusing the substrates and enzymes with an overall reaction given by

$$\begin{aligned} &Acetyl\text{-}CoA + 3NAD^+ + FAD + 2H_2O + ADP + P \longrightarrow \\ &\quad CoA - SH + 2CO_2 + FADH_2 + ATP + 3(NADH + H^+) \end{aligned} \tag{8.103}$$

The reaction begins with the joining of *acetyl-coenzyme A* with *oxaloacetate* and water to form citrate and is given by

$$\begin{aligned} &AC + H_2O + CS + O \xrightarrow{B_1} ACOCS \xrightarrow{K_1} CT + CoA - SH + CS \\ &\dot{q}_{AC} = -B_1 q_{AC} q_{H_2O} q_{CS} q_O + J_{AC} \\ &\dot{q}_{ACOCS} = B_1 q_{AC} q_{H_2O} q_{CS} q_O - K_1 q_{ACOCS} \end{aligned} \tag{8.104}$$

where J_{AC} is the flow *of* acetyl coenzyme A based on Eq. (8.102).

The next step involves the reaction of *citrate* with the enzyme *aconitase* to create *isocitrate*, which is given by

$$CT + AT \underset{B_{-2}}{\overset{B_2}{\rightleftarrows}} CTAT \underset{K_{-2}}{\overset{K_2}{\rightleftarrows}} IT + AT$$

$$\begin{aligned} \dot{q}_{CT} &= K_1 q_{ACOCS} + B_{-2} q_{CTAT} - B_2 q_{CT} q_{AT} \\ \dot{q}_{CTAT} &= B_2 q_{CT} q_{AT} + K_{-2} q_{IT} q_{AT} - (B_{-2} + K_2) q_{CTAT} \end{aligned} \tag{8.105}$$

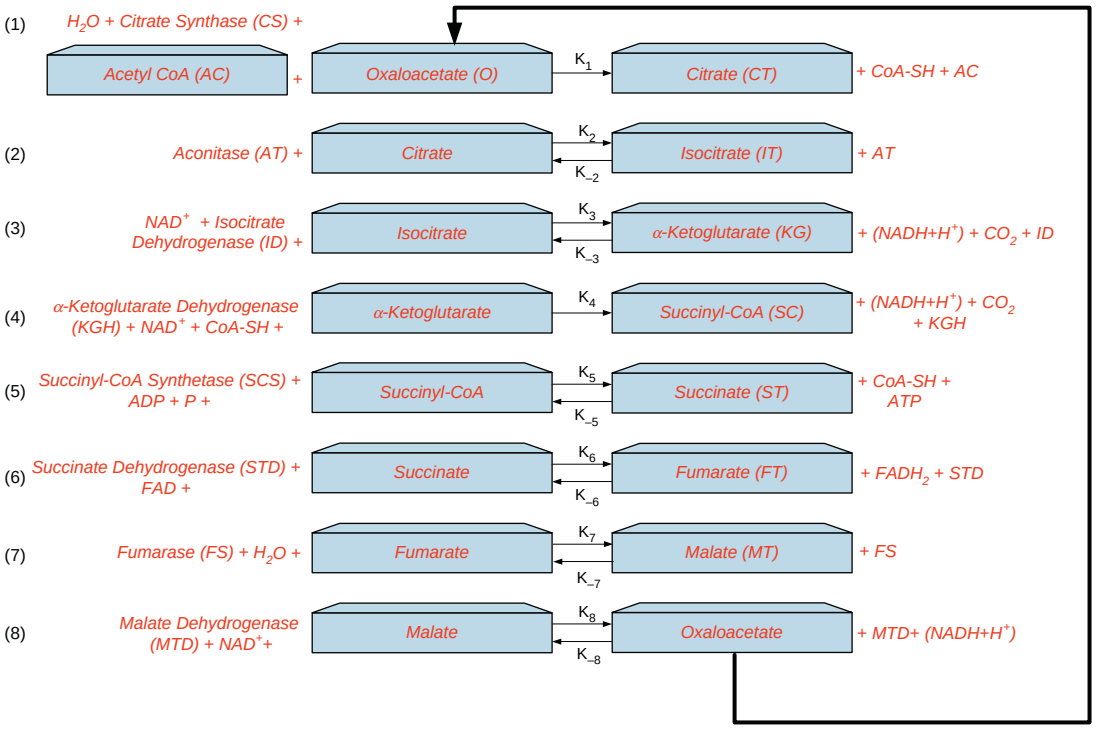


FIGURE 8.25 Overview of the Krebs cycle.

The third step involves the reaction of isocitrate and NAD^+ with the enzyme isocitrate dehydrogenase to create $\alpha-$ ketoglutarate, $(NADH+H^+)$ and CO_2, which is given by

$$\begin{aligned} &IT+ID \underset{B_{-3}}{\overset{B_3}{\rightleftharpoons}} ITID \\ &ITID+NAD^+ \underset{K_{-3}}{\overset{K_3}{\rightleftharpoons}} KG+ID+(NADH+H^+)+CO_2 \end{aligned} \tag{8.106}$$

$$\begin{aligned} \dot{q}_{IT} &= K_2 q_{CTAT} + B_{-3} q_{ITID} - K_{-2} q_{IT} q_{AT} - B_3 q_{IT} q_{ID} \\ \dot{q}_{ITID} &= B_3 q_{IT} q_{ID} + K_{-3} q_{KG} q_{ID} q_{NADH+H^+} q_{CO_2} - B_{-3} q_{ITID} - K_3 q_{ITID} q_{NAD^-} \end{aligned}$$

The fourth step involves the reaction of $\alpha-$ *ketoglutarate*, NAD^+ and *CoA* with the enzyme $\alpha-$ ketoglutarate dehydrogenase to create succinyl $-$ CoA, $(NADH+H^+)$, and CO_2. This reaction is given by

$$\begin{aligned} &KG+KGH \xrightarrow{B_4} KGKGH \\ &KGKGH+NAD^+ + CoA-SH \xrightarrow{K_4} SC+KGH+(NADH+H^+)+CO_2 \end{aligned} \tag{8.107}$$

$$\begin{aligned} \dot{q}_{KG} &= K_3 q_{ITID} q_{NAD^+} - B_4 q_{KG} q_{KGH} - K_{-3} q_{KG} q_{ID} q_{NADH+H^+} q_{CO_2} \\ \dot{q}_{KGKGH} &= B_4 q_{KG} q_{KGH} - K_4 q_{KGKGH} q_{NAD^+} q_{CoA} \end{aligned}$$

The fifth step involves the reaction of succinyl – CoA, ADP, and P with the enzyme succinyl – CoA synthetase, to create succinate, CoA and ATP. This reaction is given by

$$\begin{aligned} &SC + SCS \underset{B_{-5}}{\overset{B_5}{\rightleftharpoons}} SCSCS \\ &SCSCS + ADP + P \underset{K_{-5}}{\overset{K_5}{\rightleftharpoons}} ST + SCS + CoA - SH + ATP \end{aligned} \tag{8.108}$$

$$\begin{aligned} \dot{q}_{SC} &= K_4 q_{KGKGI1} q_{NAD^+} q_{CoA} + B_{-5} q_{SCSCS} - B_5 q_{SC} q_{SCS} \\ \dot{q}_{SCSCS} &= B_5 q_{SC} q_{SCS} + K_{-5} q_{ST} q_{SCS} q_{CoA} q_{ATP} - B_{-5} q_{SCSCS} - K_5 q_{SCSCS} q_{ADP} q_P \end{aligned}$$

The sixth step involves the reaction of succinate and FAD with the enzyme succinate dehydrogenase to create fumarate and $FADH_2$. This reaction is given by

$$\begin{aligned} &ST + STD \underset{B_{-6}}{\overset{B_6}{\rightleftharpoons}} STSTD \\ &STSTD + FAD \underset{K_{-6}}{\overset{K_6}{\rightleftharpoons}} FT + STD + FADH_2 \end{aligned} \tag{8.109}$$

$$\begin{aligned} \dot{q}_{ST} &= K_5 q_{SCSCS} q_{ADP} q_P + B_{-6} q_{STSTD} - K_{-5} q_{ST} q_{SCS} q_{CoA} q_{ATP} - B_6 q_{ST} q_{STD} \\ \dot{q}_{STSTD} &= B_6 q_{ST} q_{STD} + K_{-6} q_{FT} q_{STD} q_{FADH_2} - B_{-6} q_{STSTD} - K_6 q_{STSTD} q_{FAD} \end{aligned}$$

The seventh step involves the reaction of fumarate and H_2O with the enzyme fumarase to create malate. This reaction is given by

$$\begin{aligned} &FT + FS \underset{B_{-7}}{\overset{B_7}{\rightleftharpoons}} FTFS \\ &FTFS + H_2O \underset{K_{-7}}{\overset{K_7}{\rightleftharpoons}} MT + FS \end{aligned} \tag{8.110}$$

$$\begin{aligned} \dot{q}_{FT} &= K_6 q_{STSTD} q_{FAD} + B_{-7} q_{FTFS} - K_{-6} q_{FT} q_{STD} q_{FADH_2} - B_7 q_{FT} q_{FS} \\ \dot{q}_{FTFS} &= B_7 q_{FT} q_{FS} + K_{-7} q_{MT} q_{FS} - B_{-7} q_{STSTD} - K_7 q_{FTFS} q_{H_2O} \end{aligned}$$

The last step involves the reaction of malate and NAD^+ with the enzyme malate dehydrogenase to create oxaloacetate and $(NADH + H^+)$. This reaction is given by

$$\begin{aligned} &MT + MTD \underset{B_{-8}}{\overset{B_8}{\rightleftharpoons}} MTMTD \\ &MTMTD + NAD^+ \underset{K_{-8}}{\overset{K_8}{\rightleftharpoons}} O + MTD + (NADH + H^+). \end{aligned} \tag{8.111}$$

$$\begin{aligned} \dot{q}_{MT} &= K_7 q_{FTFS} q_{H_2O} + B_{-8} q_{FTFS} - K_{-7} q_{MT} q_{FS} - B_8 q_{MT} q_{MTD} \\ \dot{q}_{MTMTD} &= B_8 q_{MT} q_{MTD} + K_{-8} q_O q_{MTD} q_{NADH+H^+} - B_{-8} q_{MTMTD} - K_8 q_{MTMTD} q_{NAD^+} \end{aligned}$$

The final equation combines the oxaloacetate from the first reaction with that in the last reaction:

$$\dot{q}_O = K_8 q_{MTMTD} q_{NAD^+} - K_{-8} q_O q_{MTD} q_{NADH+H^+} - B_1 q_O \tag{8.112}$$

It should be clear that this reaction continually recycles and that it requires two cycles to process each glucose molecule.

8.5.3 Electron Transport Chain

The electron transport chain is the last step in the conversion of glucose into ATP, as illustrated in Figure 8.26. It involves a series of enzyme catalyzed chemical reactions that transfer electrons from $(NADH + H^+)$ and $FADH_2$ (donor molecules) to acceptor molecules. Ultimately the electron transport chain produces 32 molecules of ATP from one molecule of glucose through hydrogen oxidation, and also regenerates NAD and FAD for reuse in glycolysis. The overall reaction is given by

$$(NADH + H^+) + \frac{1}{2}O_2 + 3ADP + 3P \longrightarrow NAD^+ + 4H_2O + 3ATP$$

and

$$FADH_2 + \frac{1}{2}O_2 + 2ADP + 2P \longrightarrow FAD + 3H_2O + 2ATP \tag{8.113}$$

The electron transport chain activity takes place in the inner membrane and the space between the inner and outer membrane, called the intermembrane space. In addition to one molecule of ATP created during each Krebs cycle, three pairs of hydrogen are released and bound to $3NAD^+$ to create $3(NADH + H^+)$, and one pair of hydrogen is bound to FAD to form $FADH_2$ within the mitochondrial matrix. As described before, two cycles through the Krebs cycle are needed to fully oxidize one molecule of glucose, and thus $6(NADH + H^+)$ and $2FADH_2$ molecules are created.

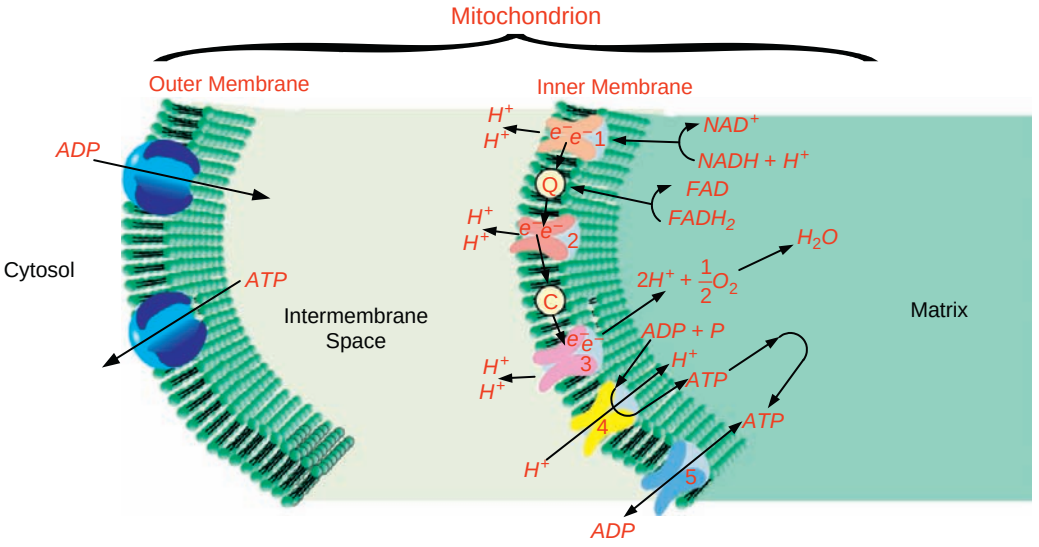


FIGURE 8.26 A simplified illustration of the mitochondrion electric transport chain. Hydrogen pumps are labeled 1 (NADH dehydrogenase), 2 (cytochrome bc_1 complex), and 3 (cytochrome c oxidase complex). Electron carriers are labeled Q (Coenzyme Q) and C (cytochrome c). The conversion of ADP+P to ATP is accomplished in the protein channel 4 (ATP synthetase), which also moves hydrogen ions back into the matrix, where they are used again in sites 1–3. Carrier mediated diffusion exchanges ATP and ADP between the matrix and the intermembrane space. Then ATP and ADP are exchanged between the intermembrane space and the cytosol by diffusion.

The energy stored in these molecules of $(NADH + H^+)$ and $FADH_2$ is used to create ATP by the release of hydrogen ions through the inner membrane and electrons within the inner membrane. The energy released by the transfer of each pair of electrons from $(NADH + H^+)$ and $FADH_2$ is used to pump a pair of hydrogen ions into the intermembrane space. The transfer of a pair of electrons is through a chain of acceptors from one to another, with each transfer providing the energy to move another pair of hydrogen ions through the membrane. At the end of the acceptor chain, the two electrons reduce an oxygen atom to form an oxygen ion, which is then combined with a pair of hydrogen ions to form H_2O. The movement of the hydrogen ions creates a large concentration of positively charged ions in the intermembrane space and a large concentration of negatively charged ions in the matrix, which sets up a large electrical potential. This potential is used by the enzyme ATP synthase to transfer hydrogen ions into the matrix and to create ATP. The ATP produced in this process is transported out of the mitochondrial matrix through the inner membrane using carrier facilitated diffusion and diffusion through the outer membrane. In the following description, we assume all of the hydrogen and electrons are available from these reactions. In reality, some are lost and not used to create ATP. Other descriptions of the electron transport chain have additional sites and are omitted here for simplicity.

We first consider the use of $(NADH + H^+)$ in the electron transport chain. During the first step, a pair of electrons from $NADH + H^+$ are transferred to the electron carrier coenzyme Q by NADH dehydrogenase (site 1 and Q in Figure 8.26), and using the energy released, a pair of hydrogen ions are pumped into the intermembrane space.

Next, the coenzyme Q carries the pair of electrons to the cytochrome bc_1 complex (site 2 in Figure 8.26). When the pair of electrons are transfered from the cytochrome bc_1 complex to cytochrome c (site C in Figure 8.26), the energy released is used to pump another pair of hydrogen ions into the intermembrane space through the cytochrome bc_1 complex.

In the third step, cytochrome c transfers electrons to the cytochrome c oxidase complex (site 3 in Figure 8.26), and another pair of hydrogen ions are pumped through the cytochrome c oxidase complex into the intermembrane space. A total of 6 hydrogen ions have now been pumped into the intermembrane space, which will allow the subseqent creation of 3 molecules of ATP.

Also occuring in this step, the cytochrome oxidase complex transfers the pair of electrons within the inner membrane from the cytochrome c to oxygen in the matrix. Oxygen then combines with a pair of hydrogen ions to form water.

As described previously, the transfer of hydrogen ions into the intermembrane space creates a large concentration of positive charges and a large concentration of negative charges in the matrix, creating a large electrical potential across the inner membrane. The energy from this potential is used in this step by the enzyme ATP synthase (site 4 in Figure 8.26) to move hydrogen ions in the intermembrane space into the matrix and to synthesize ATP from ADP and P.

The ATP in the matrix is then transported into the intermembrane space and ADP is transported into the matrix using a carrier-mediated transport process (site 5 in Figure 8.26). From the intermembrane space, ATP diffuses through the outer membrane into the cytosol, and ADP diffuses from the cytosol into the intermembrane space.

In parallel with $(NADH + H^+)$, $FADH_2$ goes through a similar process but starts at coenzyme Q, where it directly provides a pair of electrons. Thus, $FADH_2$ provides two fewer hydrogen ions than $(NADH + H^+)$.

The focus of this section has been the synthesis of ATP. Glycolysis and the Krebs cycle are also important in the synthesis of small molecules such as amino acids and nucleotides, and large molecules such as proteins, DNA, and RNA. There are other metabolic pathways to store and release energy that were not covered here. The interested reader can learn more about these pathways using the references at the end of this chapter and the website http://www.genome.jp.

8.6 ENZYME INHIBITION, ALLOSTERIC MODIFIERS, AND COOPERATIVE REACTIONS

Up to this point, we have considered the case of an enzyme binding with one substrate to make a product. Here, we examine the case in which the enzyme is free to bind with more than one molecule and form a product. As we will see, these reactions can regulate the amount of product synthesized and change the overall reaction rate in forming the product.

In this section, we begin with enzyme inhibitors that are either competitive or allosteric. A competitive enzyme inhibitor, referred to as the inhibitor, binds to the active site on the enzyme and prevents the substrate from binding with the enzyme. Thus, the inhibitor competes with the substrate to bind with the substrate and reduces the synthesis of the product and its overall reaction rate.

Some enzymes have more than one binding site and are called allosteric enzymes. The site that binds with the substrate is called the active site. The other site, called the allosteric or regulatory site, binds with another molecule called a modifier or effector. A modifier binds to the regulatory site on the enzyme and doesn't directly block the binding of the substrate with the enzyme at the active site. The effector role is to either increase (allosteric activator) or decrease (allosteric inhibitor) the activity of the enzyme. Some refer to allosteric inhibition as noncompetitive inhibition.

Finally, we examine reactions that are cooperative. These reactions are sequential and have a sigmoidal reaction velocity.

8.6.1 Competitive Enzyme Inhibitors

Consider an enzyme catalyzed reaction between a substrate, S, and enzyme, E, synthesizing product, P_1, as

$$S + E \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} C_1 \xrightarrow{K_2} P_1 + E \tag{8.114}$$

and another enzyme reaction, where enzyme inhibitor I reacts with E to form product P_2 as

$$I + E \underset{K_{-3}}{\overset{K_3}{\rightleftharpoons}} C_2 \xrightarrow{K_4} P_2 + E \tag{8.115}$$

where C_1 and C_2 are the intermediate complexes. Note that we have eliminated the reverse reaction from the product to the intermediate complex. Another form of a competitive enzyme inhibitor eliminates the formation of product P_2 (i.e., $K_4 = 0$).

The equations that describe this system in Eqs. (8.114) and (8.115) are

$$\begin{aligned}\dot{q}_S &= K_{-1}q_{C_1} - K_1 q_S q_E \\ \dot{q}_{C_1} &= K_1 q_S q_E - (K_{-1} + K_2)q_{C_1} \\ \dot{q}_I &= K_{-3}q_{C_2} - K_2 q_I q_E \\ \dot{q}_{C_2} &= K_3 q_I q_E - (K_{-3} + K_4)q_{C_2}\end{aligned} \tag{8.116}$$

We eliminate q_E from Eq. (8.116) by using $q_E = E_0 - q_{C_1} - q_{C_2}$, giving

$$\begin{aligned}\dot{q}_S &= (K_1 q_S + K_{-1})q_{C_1} + K_1 q_S q_{C_2} - K_1 E_0 q_S \\ \dot{q}_{C_1} &= K_1 E_0 q_S - (K_1 q_S + K_{-1} + K_2)q_{C_1} - K_1 q_S q_{C_2} \\ \dot{q}_I &= (K_3 q_I + K_{-3})q_{C_2} + K_3 q_I q_{C_1} - K_3 E_0 q_I \\ \dot{q}_{C_2} &= K_3 E_0 q_I - (K_3 q_I + K_{-3} + K_4)q_{C_2} - K_3 q_I q_{C_1}\end{aligned} \tag{8.117}$$

with nonzero initial conditions of $q_S(0) = S_0, q_I(0) = I_0$ and $q_E(0) = E_0$.

The quasi-steady-state approximation is found from Eq. (8.117) with $\dot{q}_{C_1} = \dot{q}_{C_2} = 0$, yielding

$$\begin{aligned}0 &= K_1 E_0 q_S - (K_1 q_S + K_{-1} + K_2)q_{C_1} - K_1 q_S q_{C_2} \\ 0 &= K_3 E_0 q_I - (K_3 q_I + K_{-3} + K_4)q_{C_2} - K_3 q_I q_{C_1}\end{aligned} \tag{8.118}$$

Next, we solve for q_{C_2}, which gives

$$q_{C_2} = \frac{K_3 q_I (E_0 - q_{C_1})}{K_3 q_I + K_{-3} + K_4} \tag{8.119}$$

and then we solve for q_{C_1}, giving

$$q_{C_1} = \frac{E_0 q_S}{q_S + K_M^s\left(1 + \frac{q_I}{K_M^i}\right)} \tag{8.120}$$

where $K_M^s = \frac{K_{-1}+K_2}{K_1}$ and $K_M^i = \frac{K_{-3}+K_4}{K_3}$. The velocity of the reaction is given by

$$V = K_2 q_{C_1} = \frac{K_2 E_0 q_S}{q_S + K_M^s\left(1 + \frac{q_I}{K_M^i}\right)} = \frac{V_{\max}}{1 + \frac{K_M^s}{q_S}\left(1 + \frac{q_I}{K_M^i}\right)} \tag{8.121}$$

where $V_{\max} = K_2 E_0$.

Comparing Eq. (8.121) with Eq. (8.47), we see that $V_{\max}$ does not change with the inclusion of an enzyme inhibitor. In this case the term K_M in Eq. (8.47) has been replaced by the term $K_M^s\left(1 + \frac{q_I}{K_M^i}\right)$ in Eq. (8.121), which reduces the reaction rate. The left side of Figure 8.27 shows a plot of the reaction rate versus the substrate with increasing quantities of the

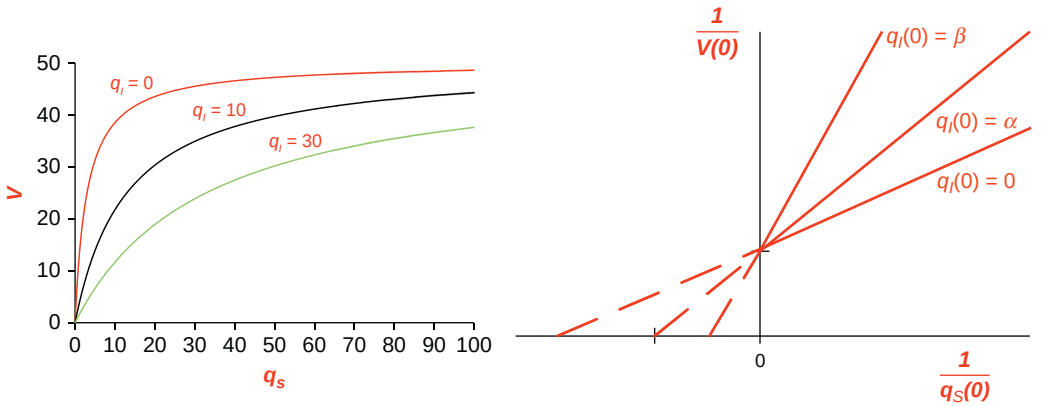


FIGURE 8.27 (Left) Velocity of product appearance for a substrate, enzyme, and enzyme inhibitor using Eq. (8.121). $V_{\max} = 50, K_M^s = 3$, and $K_M^i = 3$. (Right) Lineweaver-Burk plot, where $\beta > \alpha$.

inhibitor enzyme. As shown, increasing the quantity of the enzyme inhibitor shows a slower synthesis of the product, P_1. Keep in mind that all of the substrate will eventually be synthesized into the P_1 but it does so more slowly.

To obtain the Lineweaver-Burk equation for this system, we take the reciprocal of Eq. (8.121), giving

$$\frac{1}{V} = \frac{1}{V_{\max}}\left(1 + \left(\frac{1}{q_S}\right)K_M^s\left(1 + \frac{q_I}{K_M^i}\right)\right) \tag{8.122}$$

and at $t = 0$, we have

$$\frac{1}{V(0)} = \frac{1}{V_{\max}}\left(1 + \left(\frac{1}{q_S(0)}\right)K_M^s\left(1 + \frac{q_I(0)}{K_M^i}\right)\right) \tag{8.123}$$

The plot of the Lineweaver-Burk equation is shown on the right side of Figure 8.27 for three values of the quantity of the enzyme inhibitor. One again, the relationship between $\frac{1}{V(0)}$ vs. $\frac{1}{q_s(0)}$ is a straight line. Further, we note that the intercept of the $\frac{1}{V(0)}$ axis is a constant as the quantity of the enzyme inhibitor is increased. Additionally, the slope of the line increases as the quantity of the enzyme inhibitor is increased, which is indicative of a slowing reaction rate.

8.6.2 Allosteric Activators and Inhibitors

Next, consider the noncompetitive allosteric modifier that binds with an enzyme on a regulatory site and a substrate that binds with the enzyme on the active site. The effect of a modifier on the reaction is to either increase or decrease the activity of the enzyme.

Consider an enzyme catalyzed reaction between a substrate, S, modifier, M, and enzyme, E, synthesizing product, P, as

$$\begin{aligned}
&S + E \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} C_1 \xrightarrow{K_2} P + E \\
&M + E \underset{K_{-3}}{\overset{K_3}{\rightleftharpoons}} C_2 \\
&S + C_2 \underset{K_{-4}}{\overset{K_4}{\rightleftharpoons}} C_3 \xrightarrow{K_5} P + C_2 \\
&M + C_1 \underset{K_{-6}}{\overset{K_6}{\rightleftharpoons}} C_3 \xrightarrow{K_5} P + C_2
\end{aligned} \tag{8.124}$$

where $C_1 = SE, C_2 = ME$, and $C_3 = SME$ are the intermediate complexes.[6] Note that we have eliminated the reverse reaction from the product to the intermediate complex. Another variation of allosteric reaction eliminates the product from forming through C_3.

The equations that describe the system in Eq. (8.124) are

$$\begin{aligned}
\dot{q}_S &= K_{-1}q_{C_1} + K_{-4}q_{C_3} - K_1 q_S q_E - K_4 q_S q_{C_2} \\
\dot{q}_M &= K_{-3}q_{C_2} + K_{-6}q_{C_3} - K_3 q_M q_E - K_6 q_M q_{C_1} \\
\dot{q}_{C_1} &= K_1 q_S q_E + K_{-6}q_{C_3} - (K_{-1} + K_2)q_{C_1} - K_6 q_M q_{C_1} \\
\dot{q}_{C_2} &= K_3 q_M q_E + (K_{-4} + K_5)q_{C_3} - K_{-3}q_{C_2} - K_4 q_S q_{C_2} \\
\dot{q}_{C_3} &= K_4 q_S q_{C_2} + K_6 q_M q_{C_1} - (K_{-4} + K_5 + K_{-6})q_{C_3} \\
\dot{q}_P &= K_2 q_{C_1} + K_5 q_{C_3}
\end{aligned} \tag{8.125}$$

We eliminate q_E from Eq. (8.125) by using $q_E = E_0 - q_{C_1} - q_{C_2} - q_{C_3}$, giving

$$\begin{aligned}
\dot{q}_S &= (K_{-1} + K_1 q_S)q_{C_1} + (K_{-4} + K_1 q_S)q_{C_3} - K_1 q_S E_0 - (K_4 - K_1)q_S q_{C_2} \\
\dot{q}_M &= (K_{-3} + K_3 q_M)q_{C_2} + (K_{-6} + K_3 q_M)q_{C_3} - K_3 q_M E_0 - (K_6 - K_3)q_M q_{C_1} \\
\dot{q}_{C_1} &= K_1 q_S E_0 - K_1 q_S q_{C_2} + (K_{-6} - K_1 q_S)q_{C_3} - (K_{-1} + K_2 + K_6 q_M + K_1 q_S)q_{C_1} \\
\dot{q}_{C_2} &= K_3 q_M E_0 - K_3 q_M q_{C_1} + (K_{-4} + K_5 - K_3 q_M)q_{C_3} - (K_3 q_M + K_{-3} + K_4 q_S)q_{C_2} \\
\dot{q}_{C_3} &= K_4 q_S q_{C_2} + K_6 q_M q_{C_1} - (K_{-4} + K_5 + K_{-6})q_{C_3} \\
\dot{q}_P &= K_2 q_{C_1} + K_5 q_{C_3}
\end{aligned} \tag{8.126}$$

with nonzero initial conditions of $q_S(0) = S_0, q_M(0) = M_0$, and $q_E(0) = E_0$.

[6]This model is adapted from Rubinow, page 89.

Quasi-Steady-State Approximation to the Reaction Rate

The quasi-steady-state approximation is found from Eq. (8.126) with $\dot{q}_{C_1} = \dot{q}_{C_2} = \dot{q}_{C_3} = 0$, yielding

$$\begin{aligned} 0 &= K_1 q_S E_0 - K_1 q_S q_{C_2} + (K_{-6} - K_1 q_S) q_{C_3} - (K_{-1} + K_2 + K_6 q_M + K_1 q_S) q_{C_1} \\ 0 &= K_3 q_M E_0 - K_3 q_M q_{C_1} + (K_{-4} + K_5 - K_3 q_M) q_{C_3} - (K_3 q_M + K_{-3} + K_4 q_S) q_{C_2} \\ 0 &= K_4 q_S q_{C_2} + K_6 q_M q_{C_1} - (K_{-4} + K_5 + K_{-6}) q_{C_3} \end{aligned} \tag{8.127}$$

Equation (8.127) can be solved for q_{C_1} and q_{C_3} using the D-Operator method or the graph theory method from Rubinow, giving

$$\begin{aligned} q_{C_1} &= \frac{q_S E_0 \left\{ \frac{K_1}{K_3} K_6^i q_S + K_3^i \left(\frac{K_1}{K_4} K_6^i + \frac{K_1}{K_6} K_M^i \right) + K_6^i q_M \right\}}{D} \\ q_{C_3} &= \frac{E_0 q_M q_S \left\{ \frac{K_1}{K_3} q_S + \frac{K_1}{K_4} K_3^i + \frac{K_1}{K_6} K_M^s + q_M \right\}}{D} \end{aligned} \tag{8.128}$$

where

$$\begin{aligned} D = {} & q_S^2 \frac{K_1}{K_3} \left(K_6^i + q_M \right) \\ & + q_S \left\{ \begin{array}{l} \frac{K_1}{K_3} K_6^i K_M^s + K_3^i \left(\frac{K_1}{K_4} K_6^i + \frac{K_1}{K_6} K_M^i \right) \\ + q_M \left(K_6^i + \frac{K_1}{K_3} K_M^i + \frac{K_1}{K_4} K_3^i + \frac{K_1}{K_6} K_M^s \right) + q_M^2 \end{array} \right\} \\ & + K_M^s K_3^i \left(\frac{K_1}{K_4} K_6^i + \frac{K_1}{K_6} K_M^i \right) \\ & + q_M \left\{ K_M^s \left(\frac{K_1}{K_4} K_6^i + \frac{K_1}{K_6} K_3^i \right) + K_3^i K_M^i \right\} \\ & + K_M^i q_M^2 \end{aligned}$$

$$K_M^s = \frac{K_{-1} + K_2}{K_1}$$

$$K_M^i = \frac{K_{-4} + K_5}{K_4}$$

$$K_3^i = \frac{K_{-3}}{K_3}$$

$$K_6^i = \frac{K_{-6}}{K_6}$$

The reaction rate is

$$\begin{aligned} V = \dot{q}_P &= K_2 q_{C_1} + K_5 q_{C_3} \\ &= \frac{q_S \left[\begin{array}{l} K_2 E_0 \left\{ \frac{K_1}{K_3} K_6^i q_S + K_3^i \left(\frac{K_1}{K_4} K_6^i + \frac{K_1}{K_6} K_M^i \right) + K_6^i q_M \right\} \\ + K_5 E_0 q_M \left\{ \frac{K_1}{K_3} q_S + \frac{K_1}{K_4} K_3^i + \frac{K_1}{K_6} K_M^s + q_M \right\} \end{array} \right]}{D} \end{aligned} \tag{8.129}$$

This reaction rate is quite complex compared with the others derived previously. The reaction velocity from Eq. (8.129) is plotted in Figure 8.28. As observed, as the quantity of the allosteric modifier increases, it reduces $V_{\max}$ and the reaction rate.

Reaction Rate from the True Steady-State

We will consider a simpler allosteric modifier model[7] to further investigate the velocity of the reaction using the true steady-state rather than the quasi-steady-state approximation in the analysis. Here, the product is synthesized only from the intermediate complex C_1, and a single reaction rate is used for S and M as follows:

$$\begin{aligned} &S + E \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} C_1 \xrightarrow{K_2} P + E \\ &M + E \underset{K_{-3}}{\overset{K_3}{\rightleftarrows}} C_2 \\ &S + C_2 \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} C_3 \\ &M + C_1 \underset{K_{-3}}{\overset{K_3}{\rightleftarrows}} C_3 \end{aligned} \tag{8.130}$$

[7]This model is based on Keener and Sneyd, page 11.

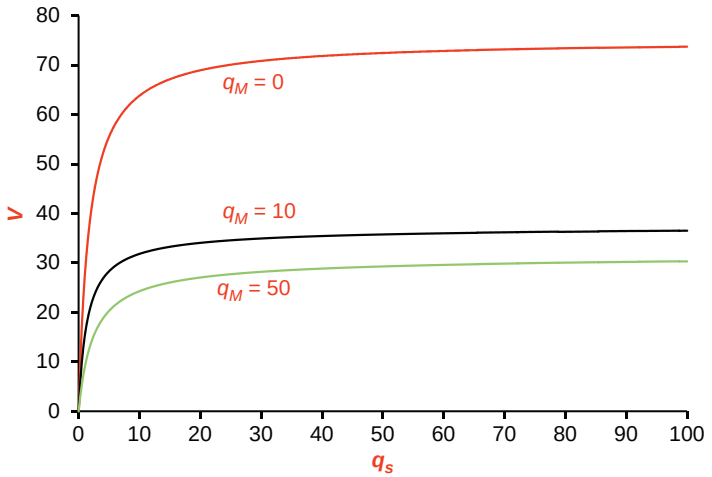


FIGURE 8.28 Velocity of product appearance for a substrate, enzyme, and allosteric modifier using Eq. (8.129). The parameters used are: $K_1 = 4, K_{-1} = 2, K_2 = 5, K_3 = 1, K_{-3} = 10, K_4 = 0.3,\ K_{-4} = 0.03, K_5 = 2, K_6 = 2, K_{-6} = 4$, and $E_0 = 15$.

The equations describing the system are given as

$$\begin{aligned}
\dot{q}_S &= K_{-1}(q_{C_1} + q_{C_3}) - K_1 q_S (q_E + q_{C_2}) \\
\dot{q}_M &= K_{-3}(q_{C_2} + q_{C_3}) - K_3 q_M (q_E + q_{C_1}) \\
\dot{q}_{C_1} &= K_{-3} q_{C_3} + K_1 q_S q_E - (K_{-1} + K_2 + K_3 q_M) q_{C_1} \\
\dot{q}_{C_2} &= K_3 q_M q_E + K_{-1} q_{C_3} - (K_{-3} + K_1 q_S) q_{C_2} \\
\dot{q}_{C_3} &= K_1 q_S q_{C_2} + K_3 q_M q_{C_1} - (K_{-1} + K_{-3}) q_{C_3} \\
\dot{q}_P &= K_2 q_{C_1}
\end{aligned} \tag{8.131}$$

where the variables and initial conditions are given as before. As before, q_E is eliminated from Eq. (8.131) by using $q_E = E_0 - q_{C_1} - q_{C_2} - q_{C_3}$, and a steady-state analysis is used to determine the reaction rate—that is, we let $\dot{q}_S = \dot{q}_M = \dot{q}_{C_1} = \dot{q}_{C_2} = \dot{q}_{C_3} = 0$. Keener and Sneyd give the reaction rate as

$$V = \frac{V_{\max}}{\left(1 + \dfrac{K_3 q_M}{K_{-3}}\right)\left(1 + \dfrac{K_{-1}}{K_1 q_S}\right)} \tag{8.132}$$

8.6.3 Cooperative Reactions

Our last topic in this chapter deals with cooperative reactions involving a sequence of biochemical reactions. The enzyme is able to bind with more than one substrate, and with each successive binding, the reaction of the next reaction is impacted. These reactions can have a sigmoidal reaction rate, allowing for a more precise control of the reaction, either

by enhancing or reducing the overall reaction. Hemoglobin is a prime example of a cooperative reaction involving the storing of oxygen in red blood cells.

A simple cooperative reaction is given by the following model:

$$S + E \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} C_1 \xrightarrow{K_2} E + P$$
$$S + C_1 \underset{K_{-3}}{\overset{K_3}{\rightleftarrows}} C_2 \xrightarrow{K_4} C_1 + P \tag{8.133}$$

using the variables previously defined. The equations that describe this system are given by

$$\begin{aligned}
\dot{q}_S &= K_{-1}q_{C_1} + K_{-3}q_{C_2} - K_1 q_S q_E - K_3 q_S q_{C_1} \\
\dot{q}_{C_1} &= K_1 q_S q_E + K_{-3}q_{C_2} + K_4 q_{C_2} - (K_{-1} + K_2)q_{c_1} - K_3 q_S q_{C_1} \\
\dot{q}_{C_2} &= K_3 q_S q_{C_1} - K_{-3}q_{C_2} \\
\dot{q}_P &= K_2 q_{C_1} + K_4 q_{C_2}
\end{aligned} \tag{8.134}$$

using the quantities previously defined. The quantity q_E is removed using $q_E = E_0 - q_{C_1} - q_{C_2}$, giving

$$\begin{aligned}
\dot{q}_S &= (K_{-1} + K_1 q_S)q_{C_1} + (K_{-3} + K_1 q_S)q_{C_2} - K_1 E_0 q_S - K_3 q_S q_{C_1} \\
\dot{q}_{C_1} &= K_1 E_0 q_S + K_{-3}q_{C_2} + (K_4 - K_1 q_S)q_{C_2} - (K_{-1} + K_2 - K_1 q_S)q_{c_1} - K_3 q_S q_{C_1} \\
\dot{q}_{C_2} &= K_3 q_S q_{C_1} - K_{-3}q_{C_2} \\
\dot{q}_P &= K_2 q_{C_1} + K_4 q_{C_2}
\end{aligned} \tag{8.135}$$

with nonzero initial conditions of $q_S(0) = S_0$, and $q_E(0) = E_0$. The quasi-steady-state approximation is found from Eq. (8.135) with $\dot{q}_{C_1} = \dot{q}_{C_2} = 0$, yielding

$$\begin{aligned}
0 &= K_1 E_0 q_S + K_{-3}q_{C_2} + (K_4 - K_1 q_S)q_{C_2} - (K_{-1} + K_2 - K_1 q_S)q_{c_1} - K_3 q_S q_{C_1} \\
0 &= K_3 q_S q_{C_1} - K_{-3}q_{C_2}
\end{aligned} \tag{8.136}$$

Equation (8.136) is solved for q_{C_1} and q_{C_2}, giving

$$\begin{aligned}
q_{C_1} &= \frac{K_M^i E_0 q_S}{K_M^i K_M^s + K_M^i q_S + q_S^2} \\
q_{C2} &= \frac{E_0 q_S^2}{K_M^i K_M^s + K_M^i q_S + q_S^2}
\end{aligned} \tag{8.137}$$

where $K_M^s = \frac{K_{-1}+K_2}{K_1}$ and $K_M^i = \frac{K_{-3}+K_4}{K_3}$. The velocity of the reaction is given by

$$\begin{aligned}
V = K_2 q_{C_1} + K_4 q_{C_2} &= \frac{K_2 K_M^i E_0 q_S}{K_M^i K_M^s + K_M^i q_S + q_S^2} + \frac{K_4 E_0 q_S^2}{K_M^i K_M^s + K_M^i q_S + q_S^2} \\
&= \frac{E_0 q_S \left(K_2 K_M^i + K_4 q_S\right)}{K_M^i K_M^s + K_M^i q_S + q_S^2}
\end{aligned} \tag{8.138}$$

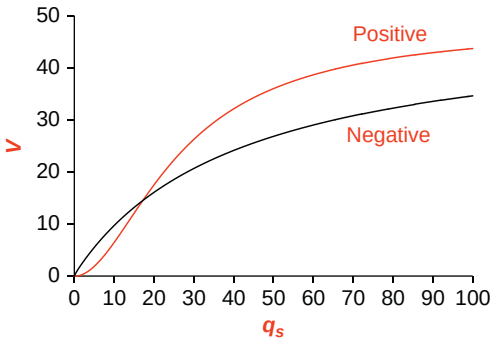


FIGURE 8.29 Velocity of product appearance for cooperative reactions using Eq. (8.138). Positive: The parameters used for the positive cooperativity are $K_1 = 0.001$, $K_{-1} = 0.00001$, $K_2 = 0.1$, $K_3 = 0.05$, $K_{-3} = 0.00002$, $K_4 = 0.3$, and $E_0 = 7$. The parameters used for the negative cooperativity are $K_1 = 1$, $K_{-1} = 0.2$, $K_2 = 0.1$, $K_3 = 0.01$, $K_{-3} = 0.4$, $K_4 = 0.02$, and $E_0 = 7$.

The overall reaction can exhibit positive or negative cooperativity depending on the parameter values. For positive cooperativity, one case is the slow binding of $S + E$ to form C_1 (K_1 is small), and the binding of $S + C_1$ to form C_2 is fast (K_3 is large), which increases the overall reaction. For negative cooperativity, one case is when the binding of $S + E$ to form C_1 (K_1 is large) slows the binding of $S + C_1$ to form C_2 (K_3 is small), which slows the overall reaction. Figure 8.29 illustrates positive and negative cooperativity. Notice the sigmoidal shape for the reaction rate for a positive cooperativity in Figure 8.29, where the negative cooperativity has a hyperbolic shape like the other reaction rates illustrated before.

8.7 EXERCISES

1. For the reaction given in Eq. (8.1) and with $q_A(0) = 15, q_B(0) = 8, q_P(0) = 0,$ and $K = 3,$ solve and simulate the solution for q_P.
2. For the reaction given in Eq. (8.1) and with $q_A(0) = 10, q_B(0) = 20, q_P(0) = 0,$ and $K = 5,$ solve and simulate the solution for q_A.
3. For the reaction given in Eq. (8.13) and with $q_A(0) = 10, q_B(0) = 20, q_P(0) = 0, K_1 = 7,$ and $K_{-1} = 5,$ solve and simulate the solution for q_A.
4. For the reaction given in Eq. (8.13) and with $q_A(0) = 18, q_B(0) = 5, q_P(0) = 0, K_1 = 10,$ and $K_{-1} = 4,$ solve and simulate the solution for q_P.
5. For the reaction given in Eq. (8.13) and with $q_A(0) = 30, q_B(0) = 10, q_P(0) = 0, K_1 = 1,$ and $K_{-1} = 0.1,$ solve and simulate the solution for q_B.
6. Show that Eq. (8.24) follows from Eq. (8.23).
7. For the reaction given in Eq. (8.25) and with $q_A(0) = 10, q_B(0) = 20, q_P(0) = 0, K_1 = 4$ $K_{-1} = 2, \alpha = 2,$ and $\beta = 1,$ simulate the solution for q_A.
8. For the reaction given in Eq. (8.25) and with $q_A(0) = 30, q_B(0) = 10, q_P(0) = 0, K_1 = 8$ $K_{-1} = 3, \alpha = 1,$ and $\beta = 3,$ simulate the solution for q_P.

Continued

9. For the reaction given in Eq. (8.25) and with $q_A(0) = 10, q_B(0) = 10, q_P(0) = 0,\ K_1 = 0.5, K_{-1} = 0.3, \alpha = 3,$ and $\beta = 2,$ simulate the solution for q_P.
10. For the reaction given in Eq. (8.27) and with $q_A(0) = 15, q_B(0) = 0, q_P(0) = 0, K_1 = 8,$ and $K_2 = 3,$ solve and simulate the solution for q_A, q_B and q_P. Compare these results with the quasi-steady-state solutions.
11. For the reaction given in Eq. (8.27) and with $q_A(0) = 25, q_B(0) = 0, q_P(0) = 0, K_1 = 2,$ and $K_2 = 10,$ solve and simulate the solution for q_A, q_B and q_P. Compare these results with the quasi-steady-state solutions.
12. For the reaction given in Eq. (8.27) and with $q_A(0) = 10, q_B(0) = 0, q_P(0) = 0, K_1 = 5,$ and $K_2 = 20,$ solve and simulate the solution for q_A, q_B and q_P. Compare these results with the quasi-steady-state solutions.
13. Generate the solutions for Figures 8.3 and 8.4.
14. Simulate the reaction given in Eq. (8.33) and compare with the quasi-steady-state approximation for q_S, q_E, q_{ES} and q_P. Assume that $K_1 = 5, K_{-1} = 0.3, K_2 = 1,\ q_S(0) = 9, q_E(0) = 0.01, q_{ES}(0) = 0,$ and $q_P(0) = 0$.
15. Simulate the reaction given in Eq. (8.33) and compare with the quasi-steady-state approximation for q_S, q_E, q_{ES} and q_P. Assume that $K_1 = 1, K_{-1} = 0.1, K_2 = 5,$ $q_S(0) = 20, q_E(0) = 0.008, q_{ES}(0) = 0,$ and $q_P(0) = 0$.
16. Simulate the reaction given in Eq. (8.33) and compare with the quasi-steady-state approximation for q_S, q_E, q_{ES} and q_P. Assume that $K_1 = 10, K_{-1} = 1, K_2 = 3,\ q_S(0) = 30, q_E(0) = 1, q_{ES}(0) = 0,$ and $q_P(0) = 0$.
17. Given the model in Eq. (8.51) and with $q_S(0) = 10, V_{\max} = 25, K_M = 5,$ and $f(t) = 5\delta(t-1),$ simulate the solution for q_S.
18. Given the model in Eq. (8.51) and with $q_S(0) = 50, V_{\max} = 5, K_M = 1,$ and $f(t) = u(t) - u(t-1),$ simulate the solution for q_S.
19. Given the model in Eq. (8.51) and with $q_S(0) = 25, V_{\max} = 10, K_M = 3,$ and $f(t) = 5e^{-t},$ simulate the solution for q_S.
20. Given the model in Eq. (8.51) and with $q_S(0) = 10, V_{\max} = 25, K_M = 5,$ and $f(t) = 5\delta(t-1),$ simulate the solution for q_S.
21. Simulate the model in Eq. (8.66) given that $K_{12} = 3, K_{21} = 1, V_{\max} = 10, K_M = 1,\ K_{10} = 0,$ and $K_{20} = 0.02$. The inputs are $f_1(t) = 4\delta(t)$ and $f_2(t) = 0$. The initial conditions are zero.
22. Simulate the model in Eq. (8.66) given that $K_{12} = 3, K_{21} = 2, V_{\max} = 20, K_M = 2,\ K_{10} = 0,$ and $K_{20}(0) = 0$. The inputs are $f_1(t) = 4(u(t) - u(t-10))$ and $f_2(t) = 3(u(t-1) - u(t-6))$. The initial conditions are zero.
23. Simulate the model in Eq. (8.67) given that $K_{12} = 0.25, K_{21} = 3, V_{\max} = 30, K_M = 5,\ K_{10} = 0,$ and $K_{20}(0) = 0.2$. The inputs are $f_1(t) = 2u(t)$ and $f_2(t) = 3u(t)$. The initial conditions are zero.
24. Simulate the model in Eq. (8.67) given that $K_{12} = 2, K_{21} = 3, V_{\max} = 40, K_M = 8,\ K_{10} = 0,$ and $K_{20}(0) = 0.2$. The inputs are $f_1(t) = 5u(t)$ and $f_2(t) = 3e^{-.03t}u(t)$. The initial conditions are zero.
25. Simulate the model in Eq. (8.67) given that $K_{12} = 3, K_{21} = 2, V_{\max} = 10, K_M = 0.8,\ K_{10} = 0.1,$ and $K_{20}(0) = 0.3$. The inputs are $f_1(t) = 4u(t)$ and $f_2(t) = 0$. The initial conditions are zero.

26. Given the model in Figure 8.15 with $q_{A_i}(0) = 30, q_{B_i}(0) = 15, q_{A_o}(0) = 20,\ C_{oi} = 3, C_{io} = 0.2,$ $D_{io} = 2, D_{oi} = 0.1, K_1 = 2,$ and $K_{-1} = 0.5,$ assume that there is flow of q_A into the exterior of the cell equal to $3u(t),$ and a production of q_{B_i} inside the cell equal to $u(t).$ All other initial conditions are zero. Simulate the solution for all quantities.

27. Given the model in Figure 8.15 with $q_{A_i}(0) = 10, q_{B_i}(0) = 25, q_{A_o}(0) = 10,\ C_{oi} = 5, C_{io} = 1,$ $D_{io} = 4,\ D_{oi} = 0.1, K_1 = 3,$ and $K_{-1} = 1,$ assume that there is flow of q_A into the exterior of the cell equal to $5u(t),$ and a production of q_{B_i} inside the cell equal to $2u(t).$ All other initial conditions are zero. Simulate the solution for all quantities.

28. Given the model in Figure 8.15 with $q_{A_i}(0) = 20, q_{B_i}(0) = 5, q_{A_o}(0) = 30,\ C_{oi} = 10, C_{io} = 0.2,$ $D_{io} = 6, D_{oi} = 0.1, K_1 = 5,$ and $K_{-1} = 0.2,$ assume that there is flow of q_A into the exterior of the cell equal to $5u(t),$ and a production of q_{B_i} inside the cell equal to $2u(t).$ All other initial conditions are zero. Simulate the solution for all quantities.

29. Given the model in Figure 8.19 with $q_{S_i}(0) = 10, q_E(0) = 0.25, q_{S_o}(0) = 10,\ B_{oi} = 4, B_{io} = 0.5,$ $D_{io} = 5, D_{oi} = 0.05, K_2 = 0.5, K_1 = 5,$ and $K_{-1} = 0.5,$ assume that there is flow of q_S into the exterior of the cell equal to $5u(t).$ All other initial quantities are zero. Simulate the solution for all quantities.

30. Given the model in Figure 8.19 with $q_{S_i}(0) = 50, q_E(0) = 2, q_{S_o}(0) = 70,\ B_{oi} = 2, B_{io} = 0.1, D_{io} = 4,\ D_{oi} = 0.1, K_2 = 5, K_1 = 8,$ and $K_{-1} = 0.1,$ assume that there is flow of q_S into the exterior of the cell equal to $8u(t).$ All other initial quantities are zero. Simulate the solution for all quantities.

31. Given the model in Figure 8.19 with $q_{S_i}(0) = 25, q_E(0) = 0.25, q_{S_o}(0) = 10,\ B_{oi} = 10, B_{io} = 0.5,$ $D_{io} = 10,\ D_{oi} = 1, K_2 = 10, K_1 = 3,$ and $K_{-1} = 0.5,$ assume that there is flow of q_S into the exterior of the cell equal to $10u(t).$ All other initial quantities are zero. Simulate the solution for all quantities.

32. Given the model in Figure 8.21 with $q_{S_i}(0) = 15, q_{C_o}(0) = q_{C_i}(0) = 5, q_{S_o}(0) = 10,\ K_2 = 4,$ $K_1 = 2,$ and $K_{-1} = 0.5,$ assume that there is flow of q_S into the exterior of the cell equal to $4u(t).$ All other initial quantities are zero. Simulate the solution for all quantities.

33. Given the model in Figure 8.21 with $q_{S_i}(0) = 25, q_{C_o}(0) = q_{C_i}(0) = 8, q_{S_o}(0) = 20,\ K_2 = 7,$ $K_1 = 5,$ and $K_{-1} = 0.1,$ assume that there is flow of q_S into the exterior of the cell equal to $10u(t).$ All other initial quantities are zero. Simulate the solution for all quantities.

34. Given the model in Figure 8.21 with $q_{S_i}(0) = 35, q_{C_o}(0) = q_{C_i}(0) = 4, q_{S_o}(0) = 50,\ q_{P_o}(0) = 0,$ $K_2 = 3, K_1 = 1,$ and $K_{-1} = 0.01,$ assume that there is flow of q_S into the exterior of the cell equal to $15u(t).$ Simulate the solution for all quantities.

35. Simulate the model in Eqs. (8.114) and (8.115) given that $q_S(0) = 10, q_E(0) = 0.1, K_1 = 4, K_{-1} = 0.2,\ K_2 = 3, K_3 = 2, K_{-3} = 0.01, K_4 = 1,$ and $q_I(0) = 3.$ All other initial quantities are zero.

36. Simulate the model in Eqs. (8.114) and (8.115) given that $q_S(0) = 20, q_E(0) = 0.01, K_1 = 6,$ $K_{-1} = 0.1, K_2 = 10, K_3 = 5, K_{-3} = 0.05, K_4 = 7,$ and $q_I(0) = 5.$ All other initial quantities are zero.

37. Simulate the model in Eqs. (8.114) and (8.115) given that $q_S(0) = 30,\ q_E(0) = 0.1,\ K_1 = 10,$ $K_{-1} = 0.1, K_2 = 3, K_3 = 7, K_{-3} = 0.09, K_4 = 9,$ and $q_I(0) = 8.$ All other initial quantities are zero.

Continued

38. Simulate the model in Eq. (8.124) given that $q_S(0) = 30$, $q_E(0) = 0.1$, $q_M = 10$, $K_1 = 10$, $K_{-1} = 0.1$, $K_2 = 3, K_3 = 3, K_{-3} = 0.1, K_4 = 2, K_{-4} = 0.1, K_5 = 5$, $K_6 = 1$, and $K_{-6} = 5$. All other initial quantities are zero.

39. Simulate the model in Eq. (8.124) given that $q_S(0) = 20$, $q_E(0) = 0.01$, $q_M = 20$, $K_1 = 5$, $K_{-1} = 1$, $K_2 = 7, K_3 = 3, K_{-3} = 10, K_4 = 0.2, K_{-4} = 0.001, K_5 = 1$, $K_6 = 6$, and $K_{-6} = 0.5$. All other initial quantities are zero.

40. Simulate the model in Eq. (8.124) given that $q_S(0) = 10$, $q_E(0) = 0.1$, $q_M = 40$, $K_1 = 7$, $K_{-1} = 1$, $K_2 = 5, K_3 = 1, K_{-3} = 5, K_4 = 2, K_{-4} = 0.1, K_5 = 2$, $K_6 = 2$, and $K_{-6} = 5$. All other initial quantities are zero.

41. Simulate the model in Eq. (8.133) given that $q_S(0) = 50$, $q_E(0) = 7$, $K_1 = 0.001$, $K_{-1} = 0.00001$, $K_2 = 0.1, K_3 = 0.05, K_{-3} = 0.00002$, and $K_4 = 0.3$. All other initial quantities are zero.

42. Simulate the model in Eq. (8.133) given that $q_S(0) = 50$, $q_E(0) = 7$, $K_1 = 1$, $K_{-1} = 0.2$, $K_2 = 0.1$, $K_3 = 0.01, K_{-3} = 0.4$, and $K_4 = 0.02$. All other initial quantities are zero.

Suggested Readings

E. Ackerman, L.C. Gatewood, Mathematical Models in the Health Sciences, University of Minnesota Press, Minneapolis, 1979.

E.S. Allman, J.A. Rhodes, Mathematical Models in Biology, An Introduction, Cambridge University Press, Cambridge, UK, 2004.

J.E. Bailey, D.F. Ollis, Biochemical Engineering Fundamentals, McGraw-Hill, Singapore, 1986.

G.E. Briggs, J.B.S. Haldane, A note on the kinetics of enzyme action, Biochem. J. 19 (1925) 338–339.

N.F. Britton, Essential Mathematical Biology, Springer, London, 2003.

M.B. Cutlip, M. Shacham, Problem Solving in Chemical and Biochemical Engineering with POLYMATH, Excel, and MATLAB, second ed., Prentice Hall, Upper Saddle River, NJ, 2008.

L. Edelstein-Keshet, Mathematical Models in Biology, SIAM, New York, 2005.

K. Godfrey, Compartmental Models and Their Applications, Academic Press, San Diego, California, 1983.

A.C. Guyton, Textbook of Medical Physiology, eighth ed., W.B. Saunders Company, Philadelphia, 1991.

V. Henri, Lois Générales de l'Action des Diastases, Hermann, Paris, 1903.

J. Keener, J. Sneyd, Mathematical Physiology, Springer, New York, 1998.

L. Michaelis, M. Menten, Die Kinetik der Invertinwirkung, Biochem. Z. 49 (1913) 333–369.

J.D. Murray, Mathematical Biology, third ed., Springer, New York, 2001.

R.B. Northrop, A.N. Connor, Introduction to Molecular Biology, Genomics and Proteomics for Biomedical Engineers, CRC Press, Boca Raton, FL, 2009.

R.B. Northrop, Endogenous and Exogenous Regulation and Control of Physiological Systems, CRC Press, Boca Raton, FL, 1999.

S.I. Rubinow, Introduction to Mathematical Biology, Dover Publications, Mineola, NY, 2002.

W.M. Saltzman, Biomedical Engineering, Bridging Medicine and Technology, Cambridge University Press, NY, 2009.

S. Schnell, C. Mendoza, Closed form solution for time-dependent enzyme kinetics, J. Theor. Biol. 187 (1997) 207–212.

M.L. Shuler, F. Kargi, Bioprocess Engineering, Basic Concepts, second ed., Prentice Hall, Upper Saddle River, NJ, 2008.

G.A. Truskey, F. Yuan, D.F. Katz, Transport Phenomena in Biological Systems, Pearson, Prentice Hall, Upper Saddle River, NJ, 2004.

CHAPTER

9

Bioinstrumentation

John D. Enderle, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Describe the components of a basic instrumentation system.
- Analyze linear circuits using the node-voltage method.
- Simplify complex circuits using Thévenin's equivalent circuits.
- Solve circuits involving resistors, capacitors, and inductors of any order.
- Analyze circuits that use operational amplifiers.
- Determine the steady-state response to sinusoidal inputs and work in the phasor domain.

- Understand the basic concepts of analog filter design and design basic filters.
- Design low-pass, high-pass, and band-pass filters.
- Explain the different types of noise in a biomedical instrument system.

9.1 INTRODUCTION

This chapter provides basic information about bioinstrumentation and electric circuit theory used in other chapters. Many biomedical instruments use a transducer or sensor to convert a signal created by the body into an electric signal. Our goal in this chapter is to develop expertise in electric circuit theory applied to bioinstrumentation. We begin with a description of variables used in circuit theory, charge, current, voltage, power, and energy. Next, Kirchhoff's current and voltage laws are introduced, followed by resistance, simplifications of resistive circuits, and voltage and current calculations. Circuit analysis techniques are then presented, followed by inductance and capacitance, and solutions of circuits using the differential equation method. Finally, the operational amplifier and time varying signals are introduced.

Before 1900, medicine had little to offer the typical citizen because its resources were mainly the education and little black bag of the physician. The origins of the changes that occurred within medical science are found in several developments that took place in the applied sciences. During the early nineteenth century, diagnosis was based on physical examination, and treatment was designed to heal the structural abnormality. By the late nineteenth century, diagnosis was based on laboratory tests, and treatment was designed to remove the cause of the disorder. The trend toward the use of technology accelerated throughout the twentieth century. During this period, hospitals became institutions of research and technology. Professionals in the areas of chemistry, physics, mechanical engineering, and electrical engineering began to work in conjunction with the medical field, and biomedical engineering became a recognized profession. As a result, medical technology advanced more in the twentieth century than it had in the rest of history combined (Figure 9.1).

During this period, the area of electronics had a significant impact on the development of new medical technology. Men such as Richard Caton and Augustus Desire proved that the human brain and heart depend on bioelectric events. In 1903, William Einthoven expanded on these ideas after he created the first string galvanometer. Einthoven placed two skin sensors on a man and attached them to the ends of a silvered wire that was suspended through holes drilled in both ends of a large permanent magnet. The suspended silvered wire moved rhythmically as the subject's heart beat. By projecting a tiny light beam across the silvered wire, Einthoven was able to record the movement of the wire as waves on a scroll of moving photographic paper. Thus, the invention of the string galvanometer led to the creation of the electrocardiogram (ECG), which is routinely used today to measure and record the electrical activity of abnormal hearts and to compare those signals to normal ones.

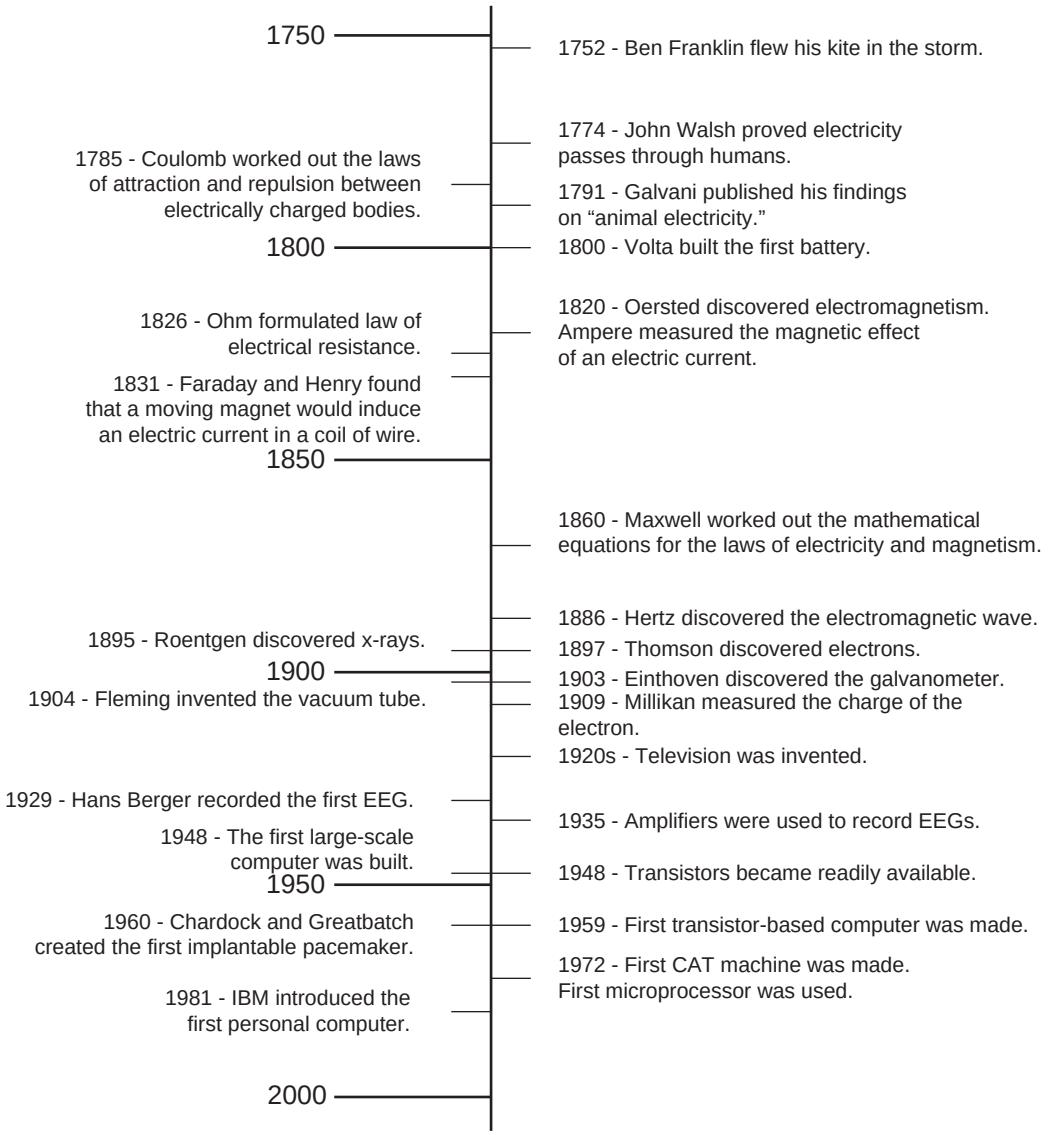


FIGURE 9.1 Timeline for major inventions and discoveries that led to modern medical instrumentation.

In 1929, Hans Berger created the first electroencephalogram (EEG), which is used to measure and record electrical activity of the brain. In 1935, electrical amplifiers were used to prove that the electrical activity of the cortex had a specific rhythm, and in 1960, electrical amplifiers were used in devices such as the first implantable pacemaker that was created by William Chardack and Wilson Greatbatch. These are just a small sample of the many

examples in which the field of electronics has been used to significantly advance medical technology.

Many other advancements that were made in medical technology originated from research in basic and applied physics. In 1895, the x-ray machine, one of the most important technological inventions in the medical field, was created when W. K. Roentgen found that x-rays could be used to give pictures of the internal structures of the body. Thus, the x-ray machine was the first imaging device to be created. (Radiation imaging is discussed in detail in Chapter 15.)

Another important addition to medical technology was provided by the invention of the computer, which allowed much faster and more complicated analyses and functions to be performed. One of the first computer-based instruments in the field of medicine, the sequential multiple analyzer plus computer, was used to store a vast amount of data pertaining to clinical laboratory information. The invention of the computer made it possible for laboratory tests to be performed and analyzed faster and more accurately.

The first large-scale computer-based medical instrument was created in 1972 when the computerized axial tomography (CAT) machine was invented. The CAT machine created an image that showed all of the internal structures that lie in a single plane of the body. This new type of image made it possible to have more accurate and easier diagnosis of tumors, hemorrhages, and other internal damage from information that was obtained noninvasively (for details, see Chapter 15).

Telemedicine, which uses computer technology to transmit information from one medical site to another, is being explored to permit access to health care for patients in remote locations. Telemedicine can be used to let a specialist in a major hospital receive information on a patient in a rural area and send back a plan of treatment specific to that patient.

Today, a wide variety of medical devices and instrumentation systems are available. Some are used to monitor patient conditions or acquire information for diagnostic purposes—for example, ECG and EEG machines—while others are used to control physiological functions—for example, pacemakers and ventilators. Some devices, like pacemakers, are implantable, while many others are used noninvasively. This chapter focuses on those features that are common to devices that are used to acquire and process physiological data.

9.2 BASIC BIOINSTRUMENTATION SYSTEM

The quantity, property, or condition that is measured by an instrumentation system is called the measurand (Figure 9.2). This can be a bioelectric signal, such as those generated by muscles or the brain, or a chemical or mechanical signal that is converted to an electrical signal. As explained in Chapter 10, sensors are used to convert physical measurands into electric outputs. The outputs from these biosensors are analog signals—that is, continuous signals—that are sent to the analog processing and digital conversion block. There, the signals are amplified, filtered, conditioned, and converted to digital form. Methods for modifying analog signals, such as amplifying and filtering an ECG signal, are discussed

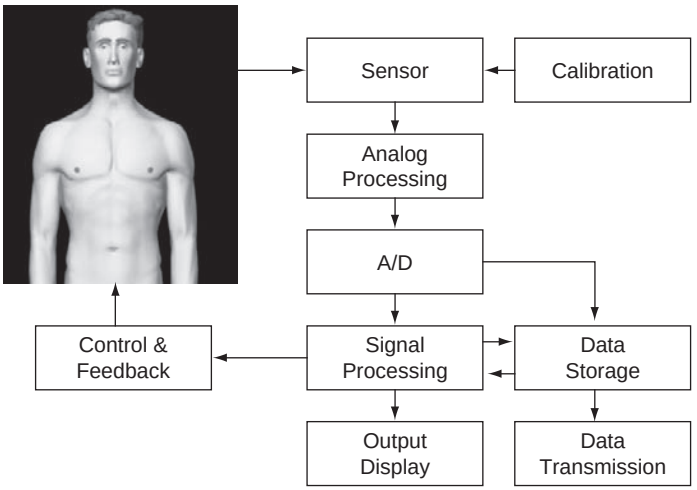


FIGURE 9.2 Basic instrumentation systems using sensors to measure a signal with data acquisition, storage, and display capabilities, along with control and feedback.

later in this chapter. Once the analog signals have been digitized and converted to a form that can be stored and processed by digital computers, many more methods of signal conditioning can be applied (for details, see Chapter 11).

Basic instrumentation systems also include output display devices that enable human operators to view the signal in a format that is easy to understand. These displays may be numerical or graphical, discrete or continuous, and permanent or temporary. Most output display devices are intended to be observed visually, but some also provide audible output—for example, a beeping sound with each heartbeat.

In addition to displaying data, many instrumentation systems have the capability of storing data. In some devices, the signal is stored briefly so further processing can take place or so an operator can examine the data. In other cases, the signals are stored permanently so different signal processing schemes can be applied at a later time. Holter monitors, for example, acquire 24 hours of ECG data that is later processed to determine arrhythmic activity and other important diagnostic characteristics.

With the invention of the telephone and now with the Internet, signals can be acquired with a device in one location, perhaps in a patient's home, and transmitted to another device for processing and/or storage. This has made it possible, for example, to provide quick diagnostic feedback if a patient has an unusual heart rhythm while at home. It has also allowed medical facilities in rural areas to transmit diagnostic images to tertiary care hospitals so that specialized physicians can help general practitioners arrive at more accurate diagnoses.

Two other components play important roles in instrumentation systems. The first is the calibration signal. A signal with known amplitude and frequency content is applied to the instrumentation system at the sensor's input. The calibration signal allows the

components of the system to be adjusted so that the output and input have a known, measured relationship. Without this information, it is impossible to convert the output of an instrument system into a meaningful representation of the measurand.

Another important component, a feedback element, is not a part of all instrumentation systems. These devices include pacemakers and ventilators that stimulate the heart or the lungs. Some feedback devices collect physiological data and stimulate a response—a heartbeat or breath—when needed or are part of biofeedback systems in which the patient is made aware of a physiological measurement, such as blood pressure, and uses conscious control to change the physiological response.

9.3 CHARGE, CURRENT, VOLTAGE, POWER, AND ENERGY

9.3.1 Charge

Two kinds of charge, positive and negative, are carried by protons and electrons, respectively. The negative charge carried by an electron, q_e, is the smallest amount of charge that exists and is measured in units called coulombs (C).

$$q_e = -1.602 \cdot 10^{-19}\ \text{C}$$

The symbol $q(t)$ is used to represent charge that changes with time, and Q is used for constant charge. The charge carried by a proton is the opposite of the electron.

9.3.2 Current

Electric current, $i(t)$, is defined as the change in the amount of charge that passes through a given point or area in a specified time period. Current is measured in amperes (A). By definition, one ampere equals one coulomb/second (C/s).

$$i(t) = \frac{dq}{dt} \tag{9.1}$$

and

$$q(t) = \int_{t_0}^{t} i(\lambda) d\lambda + q(t_0) \tag{9.2}$$

Current, defined by Eq. (9.1), also depends on the direction of flow, as illustrated in the circuit in Figure 9.3. Current is defined as positive if

a. A positive charge is moving in the direction of the arrow.
b. A negative charge is moving in the opposite direction of the arrow.

Since these two possibilities produce the same outcome, there is no need to be concerned as to which is responsible for the current. In electric circuits, current is carried by electrons in metallic conductors.

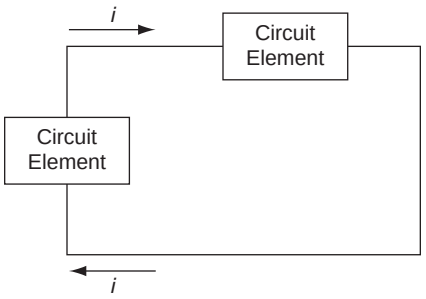


FIGURE 9.3 A simple electric circuit illustrating current flowing around a closed loop.

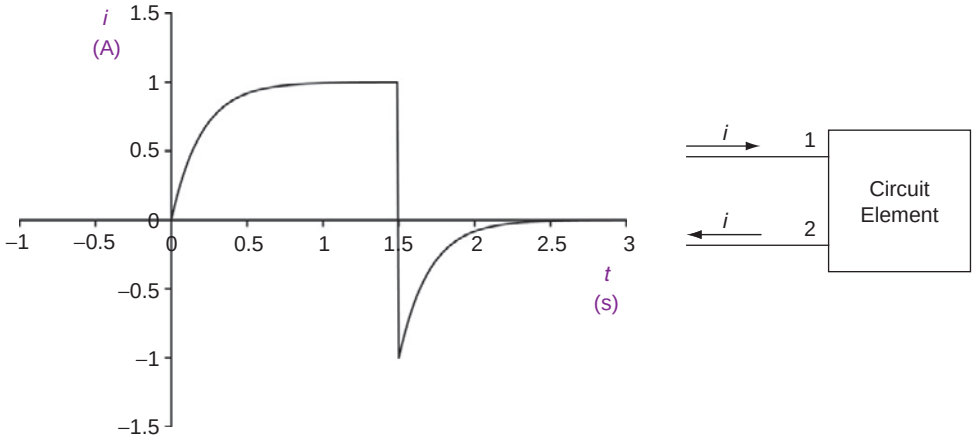


FIGURE 9.4 (Left) A sample current waveform. (Right) A circuit element with current entering terminal 1 and leaving terminal 2. Passive circuit elements have two terminals with a known voltage-current relationship. Examples of passive circuit elements include resistors, capacitors, and inductors.

Current is typically a function of time, as given by Eq. (9.1). Consider Figure 9.4, with the current entering terminal 1 in the circuit on the right. In the time interval 0 to 1.5 s, the current is positive and enters terminal 1. In the time interval 1.5 to 3 s, the current is negative and enters terminal 2 with a positive value. We typically refer to a constant current as a DC current and denote it with a capital letter such as I, indicating it doesn't change with time. We denote a time-varying current with a lowercase letter, such as $i(t)$, or just i.

Kirchhoff's Current Law

Current can flow only in a closed circuit, as shown in Figure 9.3. No current is lost as it flows around the circuit because net charge cannot accumulate within a circuit element and charge must be conserved. Whatever current enters one terminal must leave at the other terminal. Since charge cannot be created and must be conserved, the sum of the currents at any node—that is, a point at which two or more circuit elements have a common

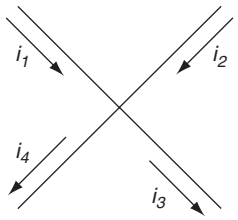


FIGURE 9.5 A node with four currents.

connection—must equal zero so no net charge accumulates. This principle is known as Kirchhoff's current law (KCL), given as

$$\sum_{n=1}^{N} i_n(t) = 0 \tag{9.3}$$

where there are N currents leaving the node. Consider the circuit in Figure 9.5. Using Eq. (9.3) and applying KCL for the currents *leaving* the node gives

$$-i_1 - i_2 + i_4 + i_3 = 0$$

The previous equation is equivalently written for the currents *entering* the node, since

$$i_1 + i_2 - i_4 - i_3 = 0$$

It should be clear that the application of KCL is for *all* currents whether they are all leaving or all entering the node.

In describing a circuit, we define its characteristics with the terms *node, branch, path, closed path,* and *mesh* as follows:

- **Node:** A point at which two or more circuit elements have a common connection.
- **Branch:** A circuit element or connected group of circuit elements. A connected group of circuit elements usually connect nodes together.
- **Path:** A connected group of circuit elements in which none is repeated.
- **Closed Path:** A path that starts and ends at the same node.
- **Mesh:** A closed path that does not contain any other closed paths within it.
- **Essential Node:** A point at which three or more circuit elements have a common connection.
- **Essential Branch:** A branch that connects two essential nodes.

Figure 9.6 shows five nodes—A, B, C, D, and E—that are all essential nodes. Kirchhoff's current law is applied to each of the nodes as follows:

$$\begin{aligned}
&\text{Node A: } -i_1 + i_2 - i_3 = 0\\
&\text{Node B: } i_3 + i_4 + i_5 - i_6 = 0\\
&\text{Node C: } i_1 - i_4 - i_8 = 0\\
&\text{Node D: } -i_7 - i_5 + i_8 = 0\\
&\text{Node E: } -i_2 + i_6 + i_7 = 0
\end{aligned}$$

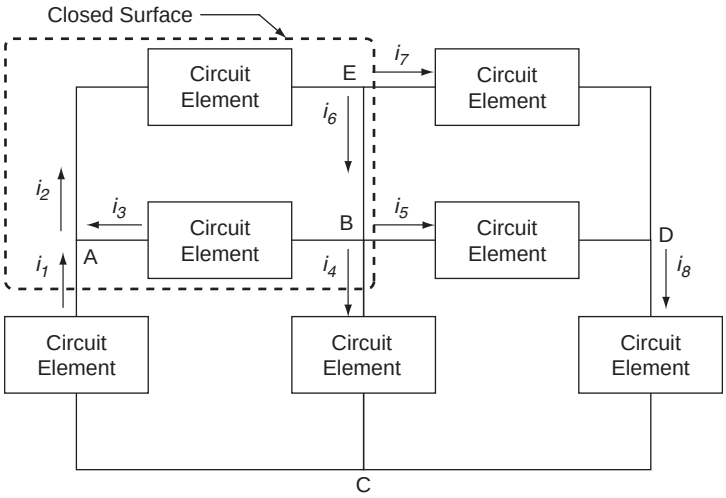


FIGURE 9.6 A circuit with a closed surface.

Kirchhoff's current law is also applicable to any closed surface surrounding a part of the circuit. It is understood that the closed surface does not intersect any of the circuit elements. Consider the closed surface drawn with dashed lines in Figure 9.6. Kirchhoff's current law applied to the closed surface gives

$$-i_1 + i_4 + i_5 + i_7 = 0$$

9.3.3 Voltage

Voltage represents the work per unit charge associated with moving a charge between two points (A and B in Figure 9.7) and is given as

$$v = \frac{dw}{dq} \tag{9.4}$$

The unit of measurement for voltage is the volt (V). A constant (DC) voltage source is denoted by V, while a time-varying voltage is denoted by $v(t)$, or just v. In Figure 9.7, the voltage, v, between two points (A and B) is the amount of energy required to move a charge from point A to point B.

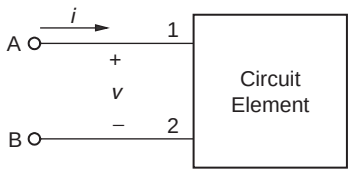


FIGURE 9.7 Voltage and current convention.

Kirchhoff's Voltage Law

Kirchhoff's voltage law (KVL) states that the sum of all voltages in a closed path is zero, or

$$\sum_{n=1}^{N} v_n(t) = 0 \tag{9.5}$$

where there are N voltage drops assigned around the closed path, with $v_n(t)$ denoting the individual voltage drops. The sign for each voltage drop in Eq. (9.5) is the first sign encountered while moving around the closed path.

Consider the circuit in Figure 9.8, with each circuit element assigned a voltage, v_n, with a given polarity and three closed paths, CP1, CP2, and CP3. Kirchhoff's voltage law for each closed path is given as

$$\text{CP1: } -v_3 + v_1 + v_4 = 0$$
$$\text{CP2: } -v_4 + v_2 + v_5 = 0$$
$$\text{CP3: } -v_3 + v_1 + v_2 + v_5 = 0$$

Kirchhoff's laws are applied in electric circuit analysis to determine unknown voltages and currents. Each unknown variable has its distinct equation. To solve for the unknowns using MATLAB, we create a matrix representation of the set of equations and solve using the techniques described in the appendix. This method is demonstrated in many examples in this chapter.

9.3.4 Power and Energy

Power is the rate of energy expenditure given as

$$p = \frac{dw}{dt} = \frac{dw}{dq}\frac{dq}{dt} = vi \tag{9.6}$$

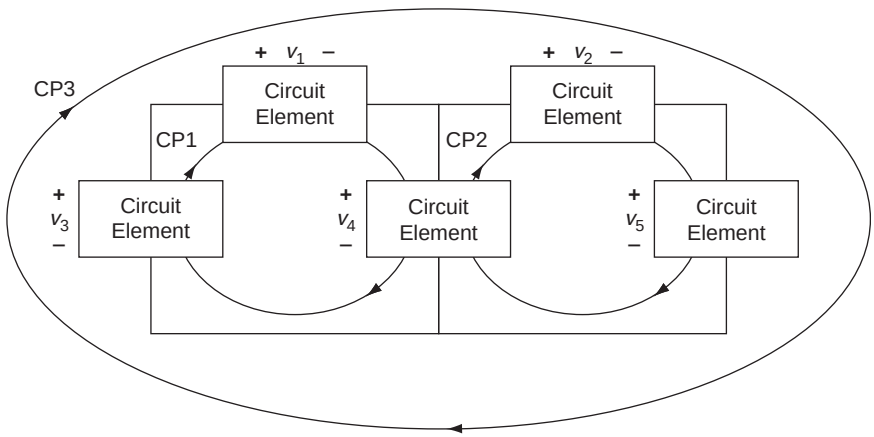


FIGURE 9.8 Circuit illustrating Kirchhoff's voltage law. Closed paths are identified as CP1, CP2, and CP3.

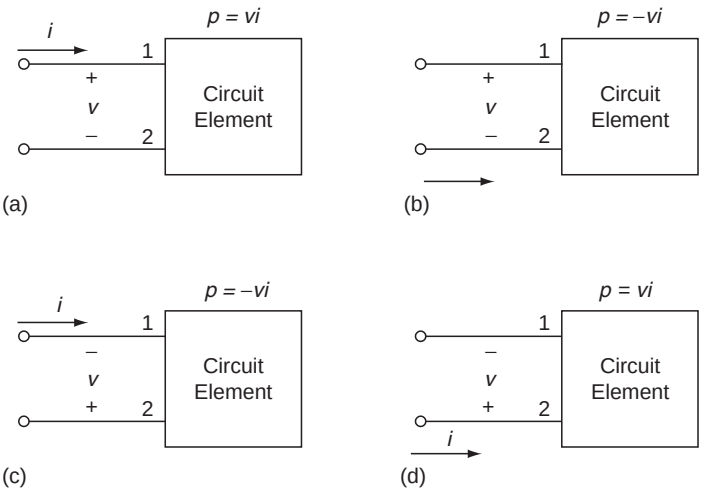


FIGURE 9.9 Polarity references for four cases of current and voltage. Cases (a) and (d) result in positive power being consumed by the circuit element. Cases (b) and (c) result in negative power being extracted from the circuit element.

where p is power measured in watts (W), and w is energy measured in joules (J). Power is usually determined by the product of voltage across a circuit element and the current through it. By convention, we assume that a positive value for power indicates that power is being delivered (or absorbed or consumed) by the circuit element. A negative value for power indicates that power is being extracted or generated by the circuit element—that is, a battery.

Figure 9.9 illustrates the four possible cases for a circuit element's current and voltage configuration. According to convention, if both i and v are positive, with the arrow and polarity shown in Figure 9.9(a), energy is absorbed (either lost by heat or stored). If either the current arrow or voltage polarity is reversed, as in (b) and (c), energy is supplied to the circuit. Note that if both the current direction and voltage polarity are reversed together, as in (d), energy is absorbed.

A passive circuit element is defined as an element whose power is always positive or zero, which may be dissipated as heat (resistance), stored in an electric field (capacitor), or stored in a magnetic field (inductor). We define an active circuit element as one whose power is negative and capable of generating energy. Energy is given by

$$w(t) = \int_{-\infty}^{t} p\,dt \tag{9.7}$$

9.3.5 Sources

Sources are two terminal devices that provide energy to a circuit. There is no direct voltage-current relationship for a source; when one of the two variables is given, the other

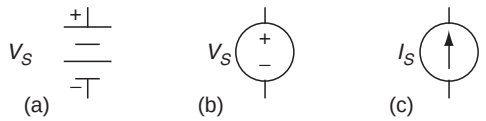


FIGURE 9.10 Basic symbols used for independent sources: (a) battery and (b) ideal voltage source. V_s can be a constant DC source (battery) or a time-varying source. (c) Ideal current source I_s.

FIGURE 9.11 Basic symbols used for dependent or controlled sources. (Left) Controlled voltage source. The voltage v_s is a known function of some other voltage or current in the circuit. (Right) Controlled current source. The current i_s is a known function of some other voltage or current in the circuit.

cannot be determined without knowledge of the rest of the circuit. Independent sources are devices for which the voltage or current is given and the device maintains its value regardless of the rest of the circuit. A device that generates a prescribed voltage at its terminals, regardless of the current flow, is called an ideal voltage source. Figures 9.10a and b show the general symbols for an ideal voltage source. Figure 9.10c shows an ideal current source that delivers a prescribed current to the attached circuit. The voltage generated by an ideal current source depends on the elements in the rest of the circuit.

Figure 9.11 shows a dependent voltage and current source. A dependent source takes on a value equaling a known function of some other voltage or current value in the circuit. We use a diamond-shaped symbol to represent a dependent source. Often, a dependent source is called a controlled source. The current generated for a dependent voltage source and the voltage for a dependent current source depend on circuit elements in the rest of the circuit. Dependent sources are very important in electronics. Later in this chapter, we will see that the operational amplifier uses a controlled voltage source for its operation.

9.4 RESISTANCE

9.4.1 Resistors

A resistor is a circuit element that limits the flow of current through it and is denoted with the symbol ^\/\/\. Resistors are made of different materials, and their ability to impede current is given with a value of resistance, denoted *R*. Resistance is measured in ohms (Ω), where $1\ \Omega = 1\ \text{V/A}$. A theoretical bare wire that connects circuit elements together has a resistance of zero. A gap between circuit elements has a resistance of infinity. An ideal resistor follows Ohm's law, which describes a linear relationship between voltage and current, with a slope equal to the resistance.

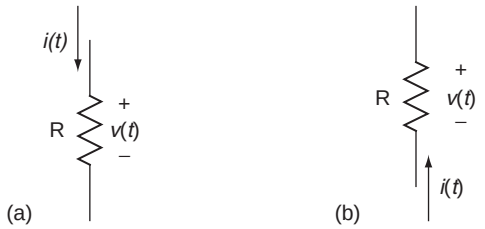


FIGURE 9.12 An ideal resister with resistance R in ohms (Ω).

There are two ways to write Ohm's law, depending on the current direction and voltage polarity. Ohm's law is written for Figure 9.12a as

$$v = iR \tag{9.8}$$

and for Figure 9.12b as

$$v = -iR \tag{9.9}$$

In this book, we use the convention shown in Figure 9.12a to write the voltage drop across a resistor. As described, the voltage across a resistor is equal to the product of the current flowing through the element and its resistance, R. This linear relationship does not apply at very high voltages and currents. Some electrically conducting materials have a very small range of currents and voltages in which they exhibit linear behavior. This is true of many physiological models as well: linearity is observed only within a range of values. Outside this range, the model is nonlinear. We define a short circuit as shown in Figure 9.13a, with $R = 0$ and having a 0 V voltage drop. We define an open circuit as shown in Figure 9.13b, with $R = \infty$ and having 0 A current pass through it.

Each material has a property called resistivity (ρ) that indicates the resistance of the material. Conductivity (σ) is the inverse of resistivity, and conductance (G) is the inverse of resistance. Conductance is measured in units called siemens (S) and has units of A/V. In terms of conductance, Ohm's law is written as

$$i = Gv \tag{9.10}$$

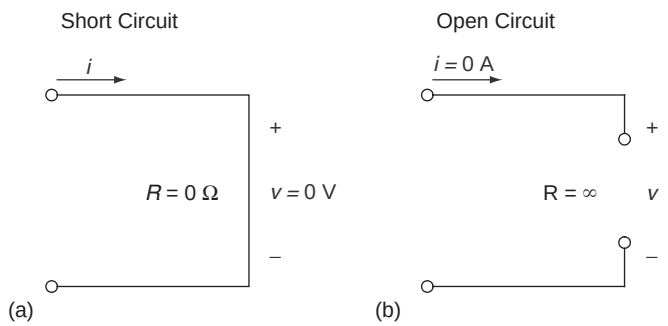
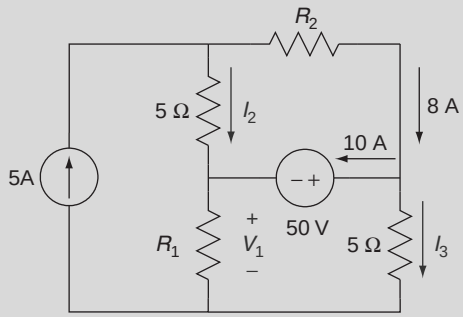


FIGURE 9.13 Short and open circuits.

EXAMPLE PROBLEM 9.1

From the following circuit, find I_2, I_3, and V_1.



Solution

Find I_2 first by applying KCL at the node in the upper left of the circuit.

$$-5 + I_2 + 8 = 0$$

and

$$I_2 = -3 \text{ A}$$

Current I_3 is determined by applying KCL at the node on the right of the circuit.

$$10 + I_3 - 8 = 0$$

and

$$I_3 = -2 \text{ A}$$

Voltage V_1 is determined by applying KVL around the lower right closed path and using Ohm's law.

$$-V_1 - 50 + 5I_3 = 0$$
$$V_1 = -50 + 5 \times (-2) = -60 \text{ V}$$

9.4.2 Power

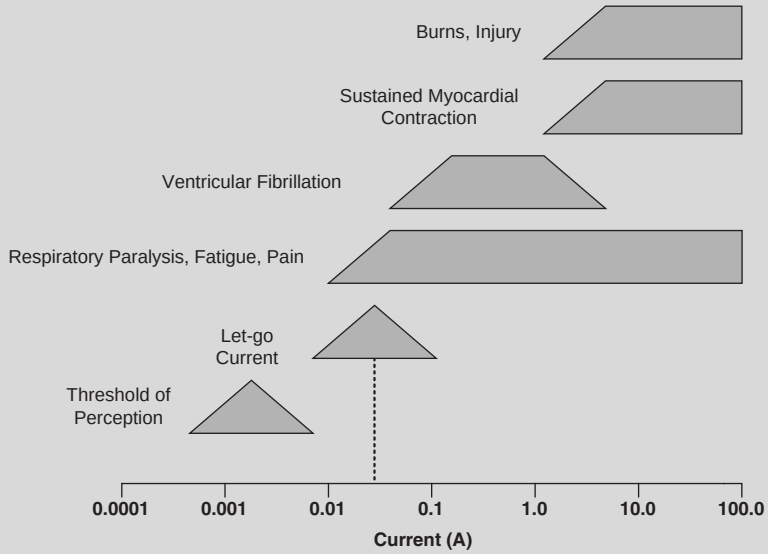
The power consumed by a resistor is given by the combination of Eq. (9.6) and either Eq. (9.8) or (9.9) as

$$p = vi = \frac{v^2}{R} = i^2 R \tag{9.11}$$

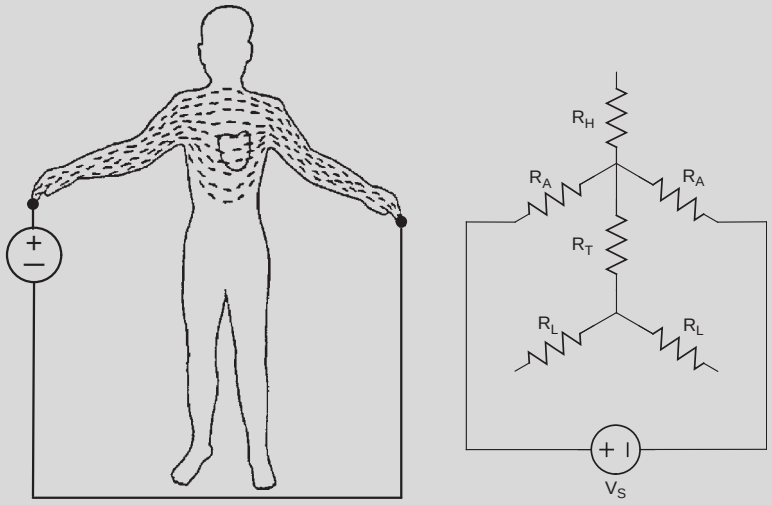
and given off as heat. Equation (9.11) demonstrates that regardless of the voltage polarity and current direction, power is consumed by a resistor. Power is always positive for a resistor, which is true for any passive element.

EXAMPLE PROBLEM 9.2

Electric safety is of paramount importance in a hospital or clinical environment. If sufficient current is allowed to flow through the body, significant damage can occur, as illustrated in the following figure.

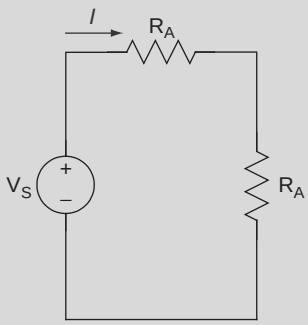


For instance, a current of magnitude 50 mA (dashed line) is enough to cause ventricular fibrillation, as well as other conditions. The figure on the left shows the current distribution from a macroshock from one arm to another. A crude electric circuit model of the body consisting of two arms (each with resistance R_A), two legs (each with resistance R_L), body trunk (with resistance R_T), and head (with resistance R_H) is shown in the following figure on the right.



Continued

Since the only elements that form a closed path through which the current can flow are given by the source in series with the two arms, we reduce the body electric circuits to



If $R_A = 400\ \Omega$ and $V_s = 120$ V, then find I.

Solution

Using Ohm's law, we get

$$I = \frac{V_s}{R_A + R_A} = \frac{120}{800} = 0.15 \text{ A}$$

The current I is the current passing through the heart, and at this level it would cause ventricular fibrillation.

9.4.3 Equivalent Resistance

It is sometimes possible to reduce complex circuits into simpler, equivalent circuits. Two circuits are considered equivalent if they cannot be distinguished from each other by voltage and current measurements—that is, the two circuits behave identically. Consider the two circuits A and B in Figure 9.14, consisting of combinations of resistors, each stimulated

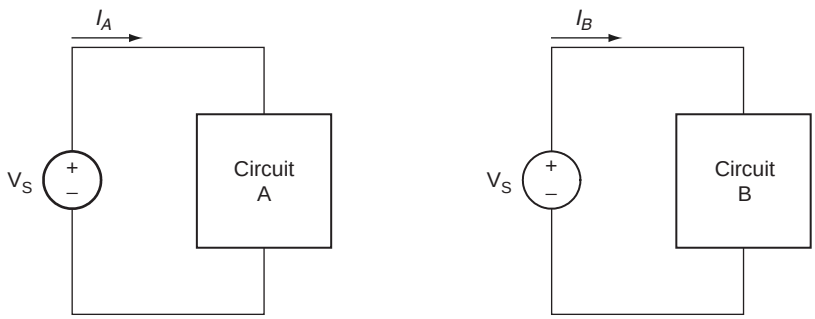


FIGURE 9.14 Two circuits.

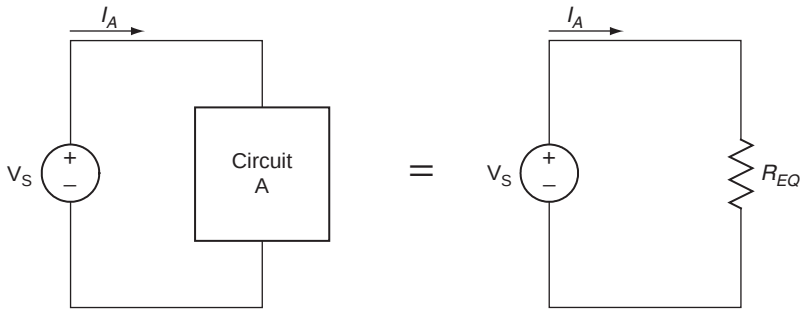


FIGURE 9.15 Equivalent circuits.

by a DC voltage V_s. These two circuits are equivalent if $I_A = I_B$. We represent the resistance of either circuit using Ohm's law as

$$R_{EQ} = \frac{V_s}{I_A} = \frac{V_s}{I_B} \tag{9.12}$$

Thus, it follows that any circuit consisting of resistances can be replaced by an equivalent circuit, as shown in Figure 9.15. When we discuss a Thévenin equivalent circuit later in this chapter, we will expand this remark to include any combination of sources and resistances.

9.4.4 Series and Parallel Combinations of Resistance

Resistors in Series

If the same current flows from one resistor to another, the two are said to be in series. If these two resistors are connected to a third and the same current flows through all of them, then the three resistors are in series. In general, if the same current flows through N resistors, then the N resistors are in series. Consider Figure 9.16 with three resistors in series. An equivalent circuit can be derived through KVL as

$$-V_s + IR_1 + IR_2 + IR_3 = 0$$

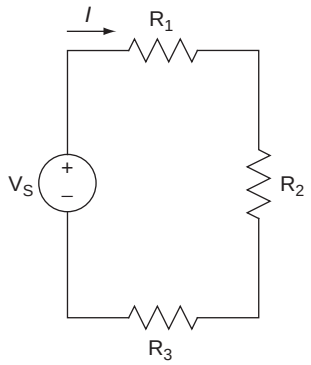


FIGURE 9.16 A series circuit.

or rewritten in terms of an equivalent resistance R_{EQ} as

$$R_{EQ} = \frac{V_s}{I} = R_1 + R_2 + R_3$$

In general, if we have N resistors in series,

$$R_{EQ} = \sum_{i=1}^{N} R_i \tag{9.13}$$

Resistors in Parallel

Two or more elements are said to be in parallel if the same voltage is across each of the resistors. Consider the three parallel resistors shown in Figure 9.17. We use a shorthand notation to represent resistors in parallel using the $\|$ symbol. Thus, in Figure 9.17, $R_{EQ} = R_1 \| R_2 \| R_3$. An equivalent circuit for Figure 9.17 is derived through KCL as

$$-I + \frac{V_s}{R_1} + \frac{V_s}{R_2} + \frac{V_s}{R_3} = 0$$

or rewritten in terms of an equivalent resistance R_{EQ} as

$$R_{EQ} = \frac{V_s}{I} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}}$$

In general, if we have N resistors in parallel,

$$R_{EQ} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_N}} \tag{9.14}$$

For just two resistors in parallel, Eq. (9.14) is written as

$$R_{EQ} = R_1 \| R_2 = \frac{R_1 R_2}{R_1 + R_2} \tag{9.15}$$

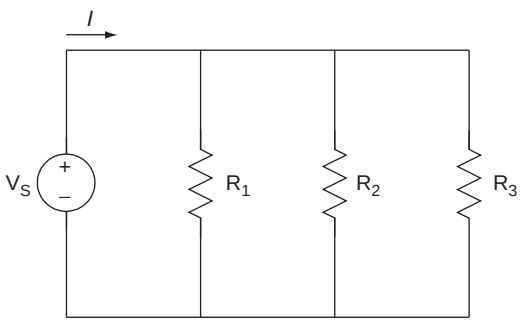
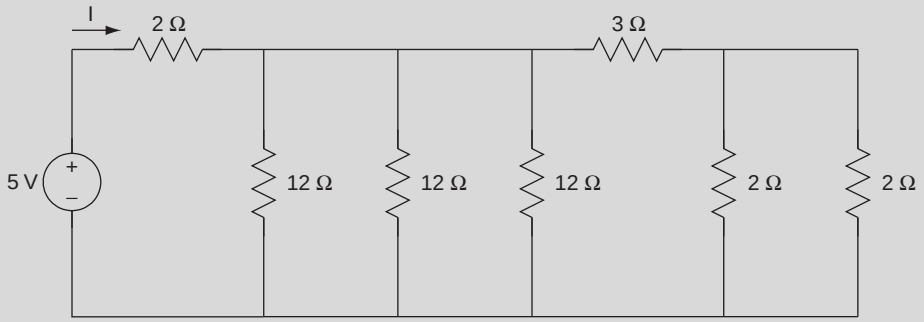


FIGURE 9.17 A parallel circuit.

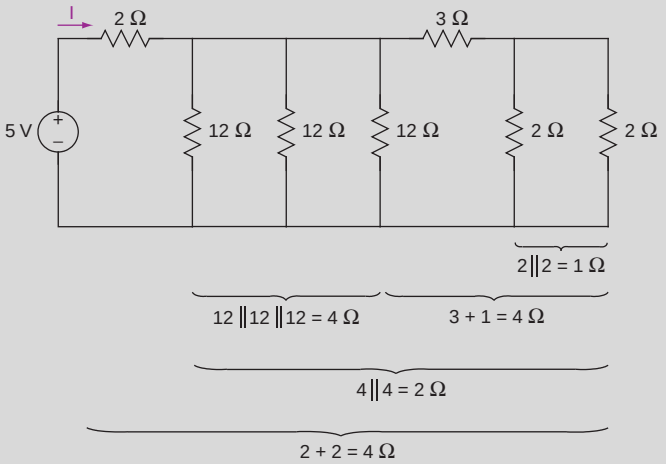
EXAMPLE PROBLEM 9.3

Find R_{EQ} and the power supplied by the source for the following circuit.



Solution

To solve for R_{EQ}, apply from right to left the parallel and series combinations. First, we have two 2 Ω resistors in parallel that are in series with the 3 Ω resistor. Next, this group is in parallel with the three 12 Ω resistors. Finally, this group is in series with the 2 Ω resistor. These combinations are shown in the following figure and calculation:



$$R_{EQ} = 2\,\Omega + ((12\,\Omega \parallel 12\,\Omega \parallel 12\,\Omega) \parallel (3\,\Omega + (2\,\Omega \| 2\,\Omega)))$$

$$= 2 + \left(\left(\frac{1}{\frac{1}{12}+\frac{1}{12}+\frac{1}{12}}\right) \parallel \left(3 + \frac{1}{\frac{1}{2}+\frac{1}{2}}\right)\right)$$

$$= 2 + ((4) \parallel (3+1)) = 2 + 2 = 4\,\Omega$$

Continued

Accordingly,

$$I = \frac{5}{R_{EQ}} = \frac{5}{4} = 1.25 \text{ A}$$

and

$$p = 5 \times I = 6.25 \text{ W}$$

9.4.5 Voltage and Current Divider Rules

Let us now extend the concept of equivalent resistance, $R_{EQ} = \frac{V}{I}$, to allow us to quickly calculate voltages in series resistor circuits and currents in parallel resistor circuits without digressing to the fundamentals.

Voltage Divider Rule

The voltage divider rule allows us to easily calculate the voltage across a given resistor in a series circuit. Consider finding V_2 in the series circuit shown in Figure 9.18, where $R_{EQ} = R_1 + R_2$. Accordingly,

$$I = \frac{V_s}{R_{EQ}} = \frac{V_s}{R_1 + R_2}$$

and therefore

$$V_2 = IR_2 = V_s \frac{R_2}{R_1 + R_2}$$

This same analysis can be used to find V_1 as

$$V_1 = V_s \frac{R_1}{R_1 + R_2}$$

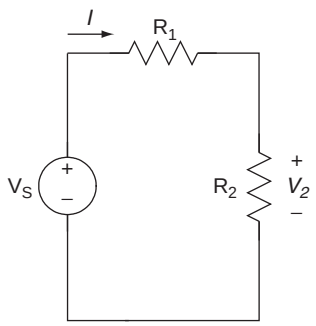


FIGURE 9.18 Voltage divider rule circuit.

In general, if a circuit contains N resistors in series, the voltage divider rule gives the voltage across any one of the resistors, R_i, as

$$V_i = V_s \frac{R_i}{R_1 + R_2 + \cdots R_N} \tag{9.16}$$

Current Divider Rule

The current divider rule allows us to easily calculate the current through any resistor in parallel resistor circuits. Consider finding I_2 in the parallel circuit shown in Figure 9.19, where $R_{EQ} = \dfrac{R_1 R_2}{R_1 + R_2}$. Accordingly,

$$I_2 = \frac{V_s}{R_2}$$

and

$$V_s = I \frac{R_1 R_2}{R_1 + R_2}$$

yielding after substituting V_s

$$I_2 = I \frac{\dfrac{1}{R_2}}{\dfrac{1}{R_1} + \dfrac{1}{R_2}}$$

In general, if a circuit contains N resistors in parallel, the current divider rule gives the current through any one of the resistors, R_i, as

$$I_i = I \frac{\dfrac{1}{R_i}}{\dfrac{1}{R_1} + \dfrac{1}{R_2} \cdots + \dfrac{1}{R_N}} \tag{9.17}$$

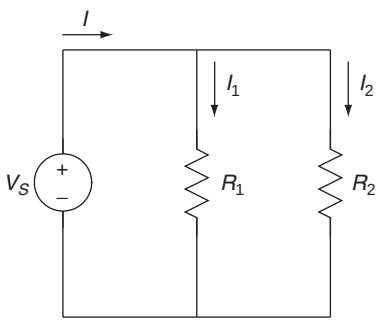
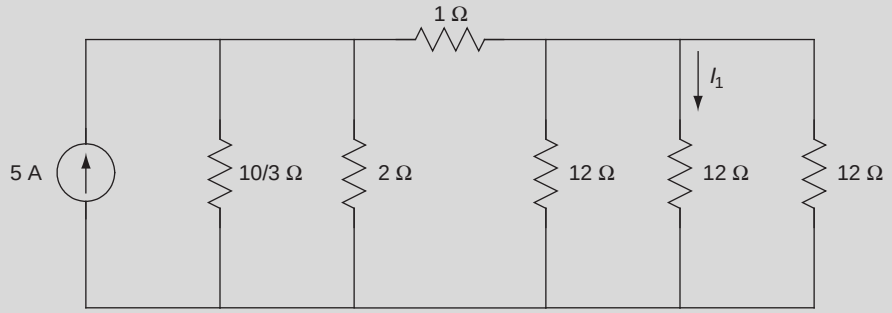


FIGURE 9.19 Current divider rule circuit.

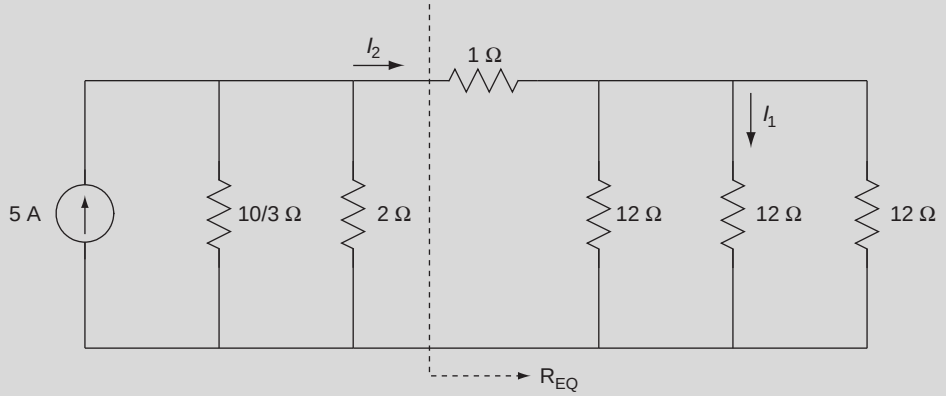
EXAMPLE PROBLEM 9.4

For the following circuit, find I_1.



Solution

This circuit problem is solved in two parts, as is evident from the redrawn circuit that follows, by first finding I_2 and then I_1.



To begin, first find R_{EQ}, which, when placed into the circuit, reduces to three parallel resistors from which I_2 is calculated. The equivalent resistance is found as

$$R_{EQ} = 1 + (12 \parallel 12 \parallel 12) = 1 + \frac{1}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 5\ \Omega$$

Applying the current divider rule on the three parallel resistors, $\frac{10}{3} \parallel 2 \parallel R_{EQ}$, we have

$$I_2 = 5\left(\frac{\frac{1}{5}}{\frac{3}{10} + \frac{1}{2} + \frac{1}{5}}\right) = 1\ \text{A}$$

I_2 flows through the 1 Ω resistor and then divides into three equal currents of $\frac{1}{3}$A through each 12 Ω resistor. The current I_1 can also be found by applying the current divider rule as

$$I_1 = I_2\left(\frac{\frac{1}{12}}{\frac{1}{12}+\frac{1}{12}+\frac{1}{12}}\right) = \frac{\frac{1}{12}}{\frac{1}{12}+\frac{1}{12}+\frac{1}{12}} = \frac{1}{3}\text{ A}$$

9.5 LINEAR NETWORK ANALYSIS

Our methods for solving circuit problems up to this point have included applying Ohm's law and Kirchhoff's laws, resistive circuit simplification, and the voltage and current divider rules. This approach works for all circuit problems, but as the circuit complexity increases, it becomes more difficult to solve problems. In this section, we introduce the node-voltage method to provide a systematic and easy solution of circuit problems. The application of the node-voltage method involves expressing the branch currents in terms of one or more node voltages and applying KCL at each of the nodes. This method provides a systematic approach that leads to a solution that is efficient and robust, resulting in a minimum number of simultaneous equations that save time and effort.

The use of node equations provides a systematic method for solving circuit analysis problems by the application of KCL at each essential node. The node-voltage method involves the following two steps:

1. Assign each node a voltage with respect to a reference node (ground). The reference node is usually the one with the most branches connected to it and is denoted with the symbol ⏚. All voltages are written with respect to the reference node.
2. Except for the reference node, we write KCL at each of the *N-1* nodes.

The current through a resistor is written using Ohm's law, with the voltage expressed as the difference between the potential on either end of the resistor with respect to the reference node, as shown in Figure 9.20. We express node-voltage equations as the currents leaving the node. Two adjacent nodes give rise to the current moving to the right (like Figure 9.20a) for one node and the current moving to the left (like Figure 9.20b) for the other node. The current is written for (a) as $I_A = \dfrac{V}{R} = \dfrac{V_1 - V_2}{R}$ and for (b) as $I_B = \dfrac{V}{R} = \dfrac{V_2 - V_1}{R}$. It is easy to verify in (a) that $V = V_1 - V_2$ by applying KVL.

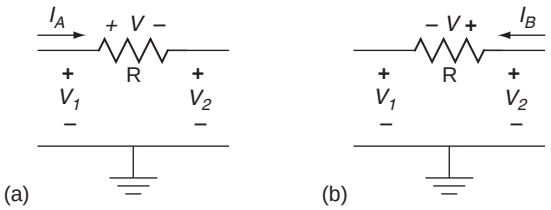
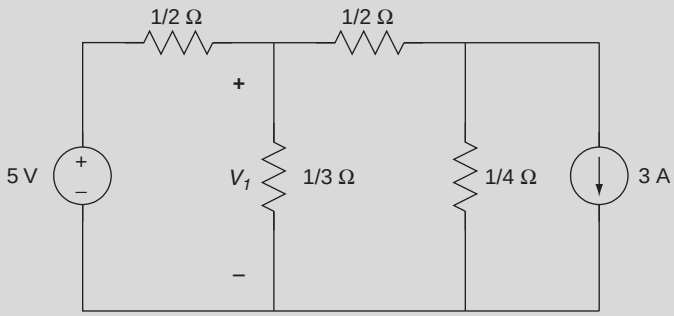


FIGURE 9.20 Ohm's law written in terms of node voltages.

If one of the branches located between an essential node and the reference node contains an independent or dependent voltage source, we do not write a node equation for this node because the node voltage is known. This reduces the number of independent node equations by one and the amount of work in solving for the node voltages. In writing the node equations for the other nodes, we write the value of the independent voltage source in those equations. Consider Figure 9.20a and assume the voltage V_2 results from an independent voltage source of 5 V. Since the node voltage is known, we do not write a node-voltage equation for node 2 in this case. When writing the node-voltage equation for node 1, the current I_A is written as $I_A = \dfrac{V_1 - 5}{R}$. Example Problem 9.5 further illustrates this case.

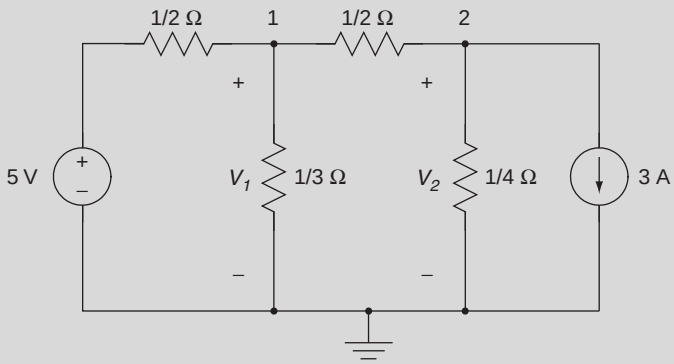
EXAMPLE PROBLEM 9.5

Find V_1 using the node-voltage method.



Solution

This circuit has two essential nodes, labeled 1 and 2 in the redrawn circuit that follows, with the reference node and two node voltages, V_1 and V_2, indicated. The node involving the 5 V voltage source has a known node voltage, and therefore we do not write a node equation for it.



Summing the currents leaving node 1 gives

$$2(V_1 - 5) + 3V_1 + 2(V_1 - V_2) = 0$$

which simplifies to

$$7V_1 - 2V_2 = 10$$

Summing the currents leaving node 2 gives

$$2(V_2 - V_1) + 4V_2 + 3 = 0$$

which simplifies to

$$-2V_1 + 6V_2 = -3$$

The two node equations are written in matrix format as

$$\begin{bmatrix} 7 & -2 \\ -2 & 6 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 10 \\ -3 \end{bmatrix}$$

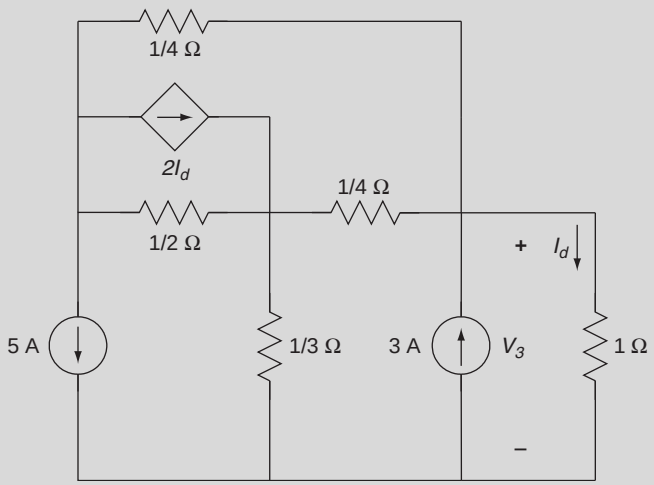
and solved with MATLAB as follows:

```
>> A = [7 -2;-2 6];
>> F = [10;-3];
>> V = A\F

V =

   1.4211
  -0.0263
```

Thus, $V_1 = 1.4211$ V.

EXAMPLE PROBLEM 9.6

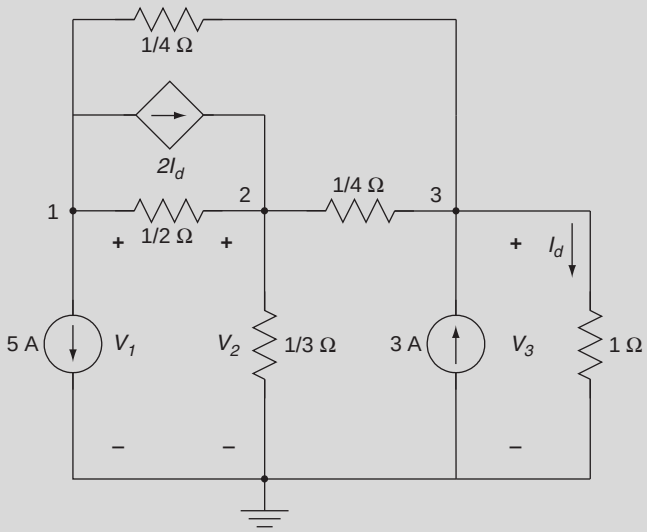
For the following circuit, find V_3 using the node-voltage method.



Continued

Solution

Notice that this circuit has three essential nodes and a dependent current source. We label the essential nodes 1, 2, and 3 in the redrawn circuit, with the reference node at the bottom of the circuit and three node voltages V_1, V_2, and V_3, as indicated.



Note that $I_d = V_3$ according to Ohm's law. Summing the currents leaving node 1 gives

$$5 + 2(V_1 - V_2) + 2I_d + 4(V_1 - V_3) = 0$$

which reduces to

$$6V_1 - 2V_2 - 2V_3 = -5$$

Summing the currents leaving node 2 gives

$$-2I_d + 2(V_2 - V_1) + 3V_2 + 4(V_2 - V_3) = 0$$

which simplifies to

$$-2V_1 + 9V_2 - 6V_3 = 0$$

Summing the currents leaving node 3 gives

$$4(V_3 - V_2) - 3 + V_3 + 4(V_3 - V_1) = 0$$

reducing to

$$-4V_1 - 4V_2 + 9V_3 = 3$$

The three node equations are written in matrix format as

$$\begin{bmatrix} 6 & -2 & -2 \\ -2 & 9 & -6 \\ -4 & -4 & 9 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ 3 \end{bmatrix}$$

Notice that the system matrix is no longer symmetrical because of the dependent current source, and two of the three nodes have a current source, giving rise to a nonzero term on the right-hand side of the matrix equation.

Solving with MATLAB gives

```
>> A = [6 -2 -2;-2 9 -6;-4 -4 9];
>> F = [-5; 0; 3];
>> V = A\F

V =

   -1.1471
   -0.5294
   -0.4118
```

Thus, $V_3 = -0.4118$ V.

If one of the branches has an independent or controlled voltage source located between two essential nodes, as shown in Figure 9.21, the current through the source is not easily expressed in terms of node voltages. In this situation, we form a supernode by combining the two nodes. The supernode technique requires only one node equation in which the current, I_A, is passed through the source and written in terms of currents leaving node 2. Specifically, we replace I_A with $I_B + I_C + I_D$ in terms of node voltages. Because we have two unknowns and one supernode equation, we write a second equation by applying KVL for the two node voltages 1 and 2 and the source as

$$-V_1 - V_\Delta + V_2 = 0$$

or

$$V_\Delta = V_1 - V_2$$

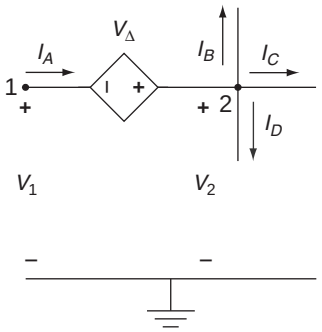
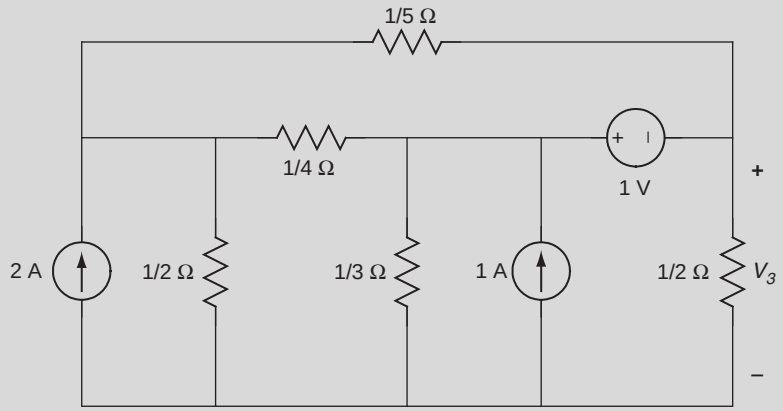


FIGURE 9.21 A dependent voltage source is located between nodes 1 and 2.

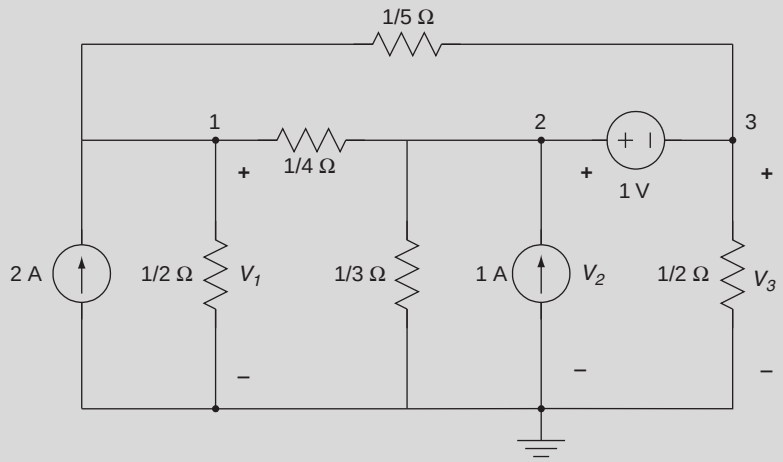
EXAMPLE PROBLEM 9.7

For the following circuit, find V_3.



Solution

The circuit has three essential nodes, two of which are connected to an independent voltage source and form a supernode. We label the essential nodes as 1, 2, and 3 in the redrawn circuit, with the reference node at the bottom of the circuit and three node voltages, V_1, V_2, and V_3 as indicated.



Summing the currents leaving node 1 gives

$$-2 + 2V_1 + 5(V_1 - V_3) + 4(V_1 - V_2) = 0$$

Simplifying gives

$$11V_1 - 4V_2 - 5V_3 = 2$$

Nodes 2 and 3 are connected by an independent voltage source, so we form a supernode 2+3. Summing the currents leaving the supernode 2+3 gives

$$4(V_2 - V_1) + 3V_2 - 1 + 2V_3 + 5(V_3 - V_1) = 0$$

Simplifying yields

$$-9V_1 + 7V_2 + 7V_3 = 1$$

The second supernode equation is KVL through the node voltages and the independent source, giving

$$-V_2 + 1 + V_3 = 0$$

or

$$-V_2 + V_3 = -1$$

The two node and KVL equations are written in matrix format as

$$\begin{bmatrix} 11 & -4 & -5 \\ -9 & 7 & 7 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

Solving with MATLAB gives

```
≫ A = [11 -4 -5;-9 7 7; 0 -1 1];
≫ F = [2;1;-1];
≫ V = A\F

V =

   0.4110
   0.8356
   -0.1644
```

Thus, $V_3 = -0.1644$.

9.6 LINEARITY AND SUPERPOSITION

If a linear system is excited by two or more independent sources, then the total response is the sum of the separate individual responses to each input. This property is called the principle of superposition. Specifically for circuits, the response to several independent sources is the sum of responses to each independent source with the other independent sources dead, where

- A dead voltage source is a short circuit.
- A dead current source is an open circuit.

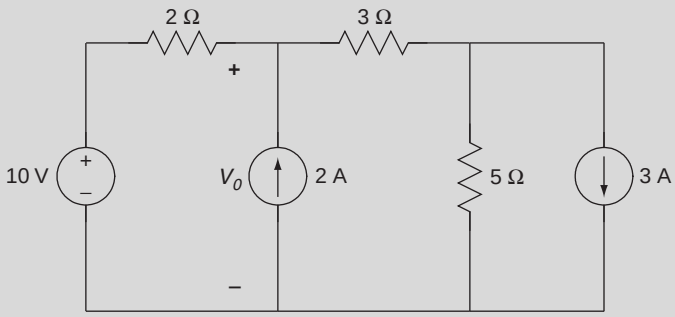
In linear circuits with multiple independent sources, the total response is the sum of each independent source taken one at a time. This analysis is carried out by removing all of

the sources except one and assuming the other sources are dead. After the circuit is analyzed with the first source, it is set equal to a dead source, and the next source is applied with the remaining sources dead. When each of the sources has been analyzed, the total response is obtained by summing the individual responses. Note carefully that this principle holds true solely for independent sources. Dependent sources must remain in the circuit when applying this technique, and they must be analyzed based on the current or voltage for which it is defined. It should be apparent that voltages and currents in one circuit differ among circuits and that we cannot mix and match voltages and currents from one circuit with another.

Generally, superposition provides a simpler solution than is obtained by evaluating the total response with all of the applied sources. This property is especially valuable when dealing with an input consisting of a pulse or delays. These are considered in future sections.

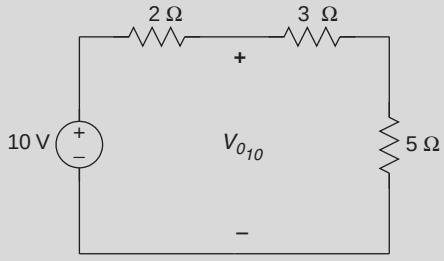
EXAMPLE PROBLEM 9.8

Using superposition, find V_0 as shown in the following figure.



Solution

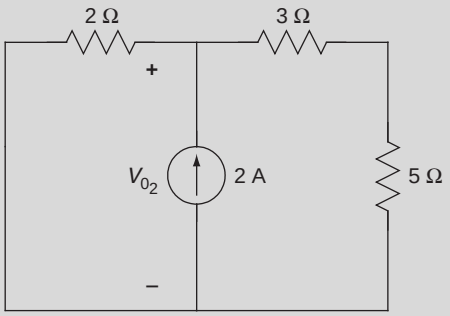
Start by analyzing the circuit with just the 10 V source active and the two current sources dead, as shown in the following figure.



The voltage divider rule easily gives the response, $V_{0_{10}}$, due to the 10 V source

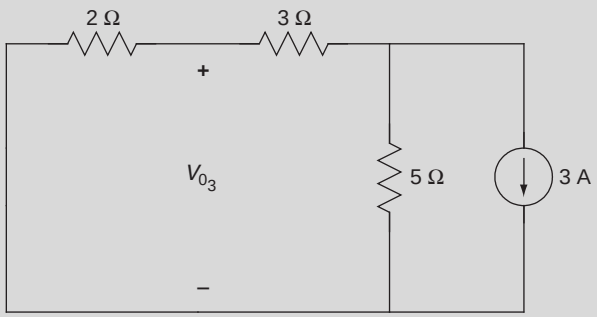
$$V_{0_{10}} = 10\left(\frac{8}{2+8}\right) = 8 \text{ V}$$

Next consider the 2 A source active and the other two sources dead, as shown in the following circuit.



Combining the resistors in an equivalent resistance, $R_{EQ} = 2 \parallel (3+5) = \dfrac{2 \times 8}{2+8} = 1.6\ \Omega$, and then applying Ohm's law gives $V_{0_2} = 2 \times 1.6 = 3.2$ V.

Finally, consider the response, V_{0_3}, to the 3 A source as shown in the following figure.



To find V_{0_3}, note that the 3 A current splits into 1.5 A through each branch ($2 + 3\ \Omega$ and $5\ \Omega$), and $V_{0_3} = -1.5 \times 2 = -3$ V.

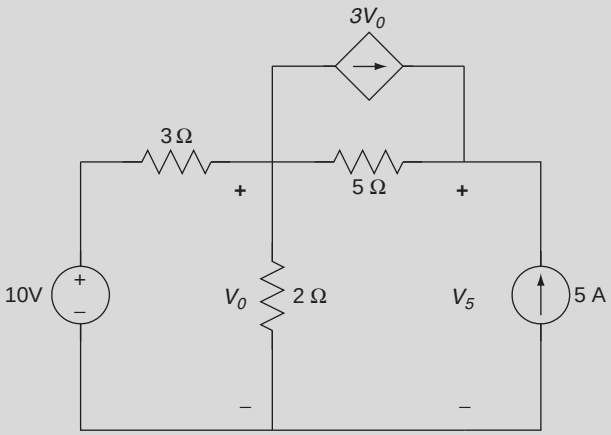
The total response is given by the sum of the individual responses as

$$V_0 = V_{0_{10}} + V_{0_2} + V_{0_3} = 8 + 3.2 - 3 = 8.2 \text{ V}$$

This is the same result we would have found if we analyzed the original circuit directly using the node-voltage method.

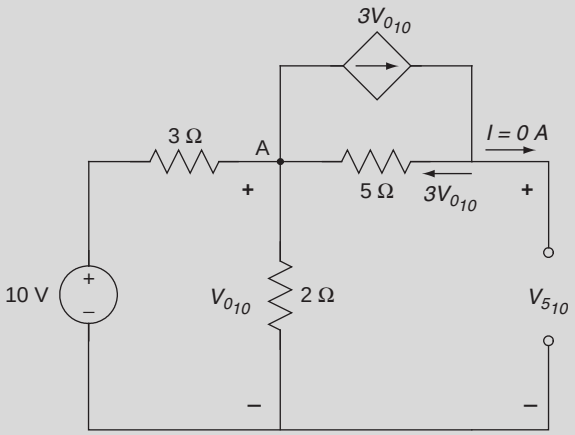
EXAMPLE PROBLEM 9.9

Find the voltage across the 5 A current source, V_5, in the following figure using superposition.



Solution

First consider finding the response, $V_{0_{10}}$, due to the 10 V source only, with the 5 A source dead, as shown in the following figure. As required during the analysis, the dependent current source is kept in the modified circuit and should not be set dead.

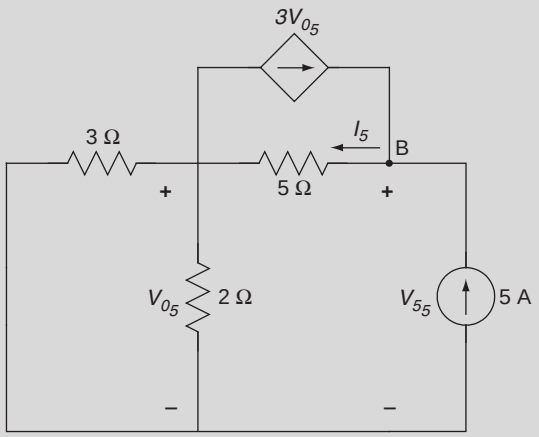


Notice that no current flows through the open circuit created by the dead current source and that the current flowing through the 5 Ω resistor is $3V_0$. Therefore, applying KCL at node A gives

$$\frac{V_{0_{10}} - 10}{3} + \frac{V_{0_{10}}}{2} + 3V_{0_{10}} - 3V_{0_{10}} = 0$$

which gives $V_{0_{10}} = 4$V. KVL gives $-V_{0_{10}} - 5 \cdot 3V_{0_{10}} + V_{5_{10}} = 0$, and therefore $V_{5_{10}} = 64$ V.

Next, consider finding the response, V_{0_5}, due to the 5 A source, with the 10 V source dead.



First combine the two resistors in parallel ($3\Omega \parallel 2\Omega$), giving 1.2 Ω. V_{0_5} is easily calculated by Ohm's law as $V_{0_5} = 5 \cdot 1.2 = 6\text{V}$. KCL is then applied at node B to find I_5, giving

$$-3V_{0_5} + I_5 - 5 = 0$$

With $V_{0_5} = 6$ V, $I_5 = 3 \cdot 6 + 5 = 23$ A. Finally, apply KVL around the closed path

$$-V_{0_5} - 5I_5 + V_{5_5} = 0$$

or

$$V_{5_5} = V_{0_5} + 5I_5 = 6 + 5 \cdot 23 = 121 \text{ V}.$$

The total response is given by the sum of the individual responses as

$$V_5 = V_{5_{10}} + V_{5_5} = 64 + 121 = 185 \text{ V}$$

9.7 THÉVENIN'S THEOREM

Any combination of resistances, controlled sources, and independent sources with two external terminals (A and B, denoted A,B) can be replaced by a single resistance and an independent source, as shown in Figure 9.22. A Thévenin equivalent circuit reduces the original circuit into a voltage source in series with a resistor. This theorem helps reduce complex circuits into simpler circuits. We refer to the circuit elements connected across the terminals A,B (that are not shown) as the *load*. The Thévenin equivalent circuit is

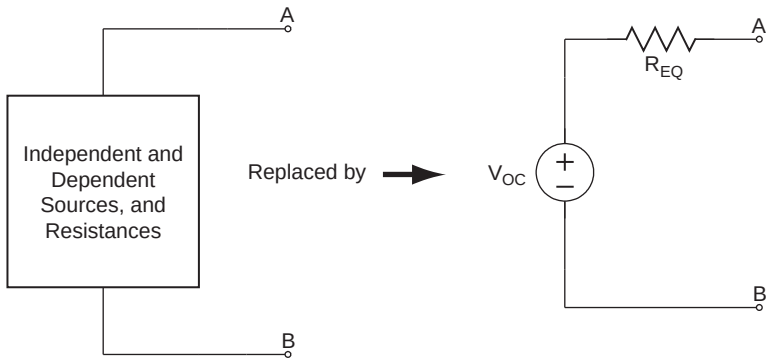


FIGURE 9.22 A general circuit consisting of independent and dependent sources can be replaced by a voltage source (V_{oc}) in series with a resistor (R_{EQ}).

equivalent to the original circuit in that the same voltage and current are observed across any load. Usually the load is not included in the simplification because it is important for other analysis, such as maximum power expended by the load. Although we focus here on sources and resistors, this theorem can be extended to any circuit composed of linear elements with two terminals.

Thévenin's Theorem states that an equivalent circuit consisting of an ideal voltage source, V_{OC}, in series with an equivalent resistance, R_{EQ}, can be used to replace any circuit that consists of independent and dependent voltage and current sources and resistors. V_{OC} is equal to the open circuit voltage across terminals A,B, as shown in Figure 9.23, and calculated using standard techniques such as the node-voltage method.

The resistor R_{EQ} is the resistance seen across the terminals A,B when all sources are *dead*. Recall that a dead voltage source is a short circuit, and a dead current source is an open circuit.

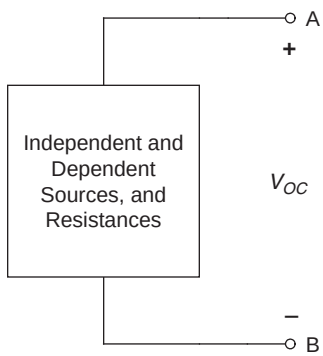
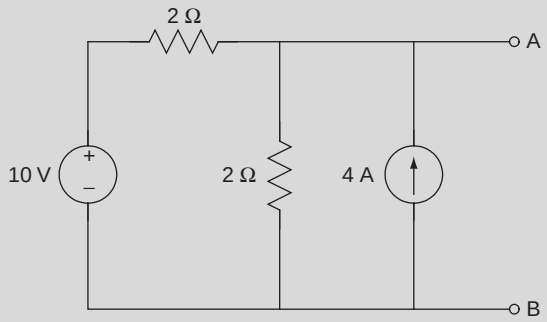


FIGURE 9.23 The open circuit voltage, V_{oc}, is calculated across the terminals A,B using standard techniques such as the node-voltage method.

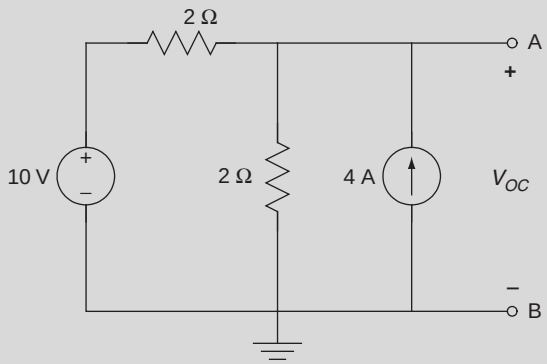
EXAMPLE PROBLEM 9.10

Find the Thévenin equivalent circuit with respect to terminals A,B for the following circuit.



Solution

The solution to finding the Thévenin equivalent circuit is done in two parts: first finding V_{OC} and then solving for R_{EQ}. The open circuit voltage, V_{OC}, is easily found using the node-voltage method, as shown in the following circuit.



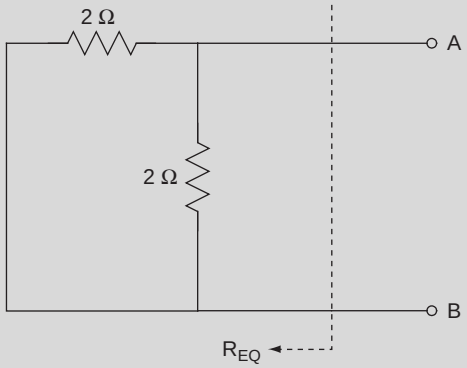
The sum of currents leaving the node is

$$\frac{V_{OC} - 10}{2} + \frac{V_{OC}}{2} - 4 = 0$$

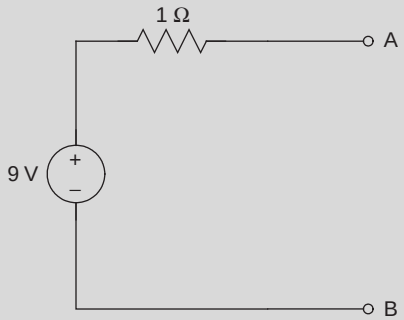
and $V_{OC} = 9$ V.

Next, R_{EQ} is found by first setting all sources dead (the current source is an open circuit and the voltage source is a short circuit) and then finding the resistance seen from the terminals A,B, as shown in the following figure.

Continued



From the previous circuit, it is clear that R_{EQ} is equal to 1Ω (that is, 2Ω || 2Ω). Thus, the Thévenin equivalent circuit is



It is important to note that the circuit used in finding V_{OC} is not to be used in finding R_{EQ} as not all voltages and currents are relevant in the other circuit and one cannot simply mix and match.

If the terminals A,B are shorted as shown in Figure 9.24, the current that flows is denoted I_{SC}, and the following relationship holds:

$$R_{EQ} = \frac{V_{OC}}{I_{SC}} \tag{9.18}$$

9.8 INDUCTORS

In the previous sections of this chapter, we considered circuits involving sources and resistors that are described with algebraic equations. Any changes in the source are instantaneously observed in the response. In this section we examine the inductor, a passive element that relates the voltage-current relationship with a differential equation. Circuits that

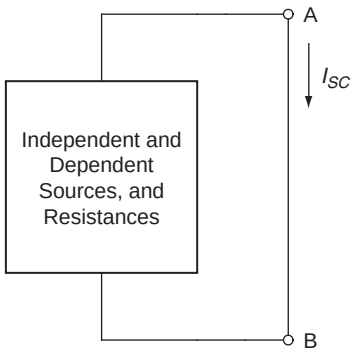


FIGURE 9.24 The short circuit current, I_{sc}, is calculated by placing a short across the terminals A,B and finding the current through the short using standard techniques such as the node-voltage method.

contain inductors are written in terms of derivatives and integrals. Any changes in the source with circuits that contain inductors—that is, a step input—have a response that is not instantaneous but have a natural response that changes exponentially and a forced response that is the same form as the source.

An inductor is a passive element that is able to store energy in a magnetic field and is made by winding a coil of wire around a core that is an insulator or a ferromagnetic material. A magnetic field is established when current flows through the coil. We use the symbol ⌒⌒⌒ to represent the inductor in a circuit; the unit of measure for inductance is the henry or henries (H), where 1 H = 1 V – s/A. The relationship between voltage and current for an inductor is given by

$$v = L\frac{di}{dt} \tag{9.19}$$

The convention for writing the voltage drop across an inductor is similar to that of a resistor, as shown in Figure 9.25.

Physically, current cannot change instantaneously through an inductor, since an infinite voltage is required, according to Eq. (9.19) (i.e., the derivative of current at the time of the instantaneous change is infinity). Mathematically, a step change in current through an inductor is possible by applying a voltage that is a Dirac delta function. For convenience, when a circuit has just DC currents (or voltages), the inductors can be replaced by short circuits, since the voltage drops across the inductors are zero.

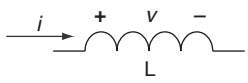
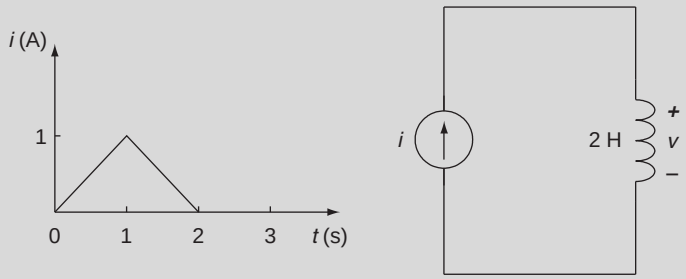


FIGURE 9.25 An inductor.

EXAMPLE PROBLEM 9.11

Find v in the following circuit.



Solution

The solution to this problem is best approached by breaking it up into time intervals consistent with the changes in input current. Clearly, for $t < 0$ and $t > 2$, the current is zero and therefore $v = 0$. We use Eq. (9.19) to determine the voltage in the other two intervals as follows.

For 0 < t < 1

In this interval, the input is $i = t$, and

$$v = L\frac{di}{dt} = 2\frac{d(t)}{dt} = 2\ \text{V}$$

For 1 ≤ t ≤ 2

In this interval, the input is $i = -(t-2)$, and

$$v = L\frac{di}{dt} = 2\frac{d(-(t-2))}{dt} = -2\ \text{V}$$

Equation (9.19) defines the voltage across an inductor for a given current. Suppose one is given a voltage across an inductor and asked to find the current. We start from Eq. (9.19) by multiplying both sides by dt, giving

$$v(t)dt = Ldi$$

Integrating both sides yields

$$\int_{t_0}^{t} v(\lambda)d\lambda = L\int_{i(t_0)}^{i(t)} d\alpha \tag{9.20}$$

or

$$i(t) = \frac{1}{L}\int_{t_0}^{t} v(\lambda)d\lambda + i(t_0)$$

For $t_0 = 0$, as is often the case in solving circuit problems, Eq. (9.20) reduces to

$$i(t) = \frac{1}{L}\int_0^t v(\lambda)d\lambda + i(0) \tag{9.21}$$

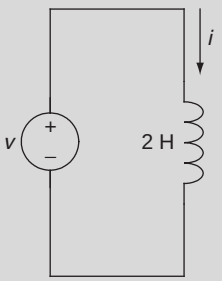
and for $t_0 = -\infty$, the initial current is by definition equal to zero, and therefore Eq. (9.20) reduces to

$$i(t) = \frac{1}{L}\int_{-\infty}^t v(\lambda)d\lambda \tag{9.22}$$

The initial current in Eq. (9.20), $i(t_0)$, is usually defined in the same direction as i, which means $i(t_0)$ is a positive quantity. If the direction of $i(t_0)$ is in the opposite direction of i (as will happen when we write node equations), then $i(t_0)$ is negative.

EXAMPLE PROBLEM 9.12

Find i for $t \geq 0$ if $i(0) = 2$ A and $v(t) = 4e^{-3t}u(t)$ in the following circuit.



Solution

From Eq. (9.20), we have

$$i(t) = \frac{1}{L}\int_{t_0}^t v d\lambda + i(t_0) = \frac{1}{2}\int_0^t 4e^{-3\lambda}d\lambda + 2$$

$$= 2\frac{e^{-3\lambda}}{-3}\bigg|_{\lambda=0}^{t} + 2$$

$$= \frac{2}{3}\left(4 - e^{-3t}\right)u(t)\ \text{V}$$

9.9 CAPACITORS

A capacitor is a device that stores energy in an electric field by charge separation when appropriately polarized by a voltage. Simple capacitors consist of parallel plates of conducting material that are separated by a gap filled with a dielectric material. Dielectric materials—that is, air, mica, or Teflon—contain a large number of electric dipoles that become polarized in the presence of an electric field. The charge separation caused by the polarization of the dielectric is proportional to the external voltage and given by

$$q(t) = C\,v(t) \tag{9.23}$$

where C represents the capacitance of the element. The unit of measure for capacitance is the farad or farads (F), where 1 F = 1 C/V. We use the symbol $\dashv\!\vdash$ C to denote a capacitor; most capacitors are measured in terms of microfarads (1 μF = 10^{-6} F) or picofarads (1 pF = 10^{-12} F). Figure 9.26 illustrates a capacitor in a circuit.

Using the relationship between current and charge, Eq. (9.23) is written in a more useful form for circuit analysis problems as

$$i = \frac{dq}{dt} = C\frac{dv}{dt} \tag{9.24}$$

The capacitance of a capacitor is determined by the permittivity of the dielectric ($\varepsilon = 8.854 \times 10^{-12}\,\frac{\text{F}}{\text{m}}$ for air) that fills the gap between the parallel plates, the size of the gap between the plates, d, and the cross-sectional area of the plates, A, as

$$C = \frac{\varepsilon A}{d} \tag{9.25}$$

As described, the capacitor physically consists of two conducting surfaces that store charge, separated by a thin insulating material that has a very large resistance. In actuality, current does not flow through the capacitor plates. Rather, as James Clerk Maxwell hypothesized when he described the unified electromagnetic theory, a displacement current flows internally between capacitor plates, and this current equals the current flowing into the capacitor and out of the capacitor. Thus, KCL is maintained. It should be clear from Eq. (9.29) that dielectric materials do not conduct DC currents; capacitors act as open circuits when DC currents are present.

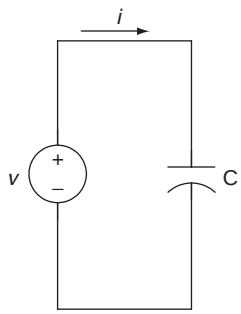
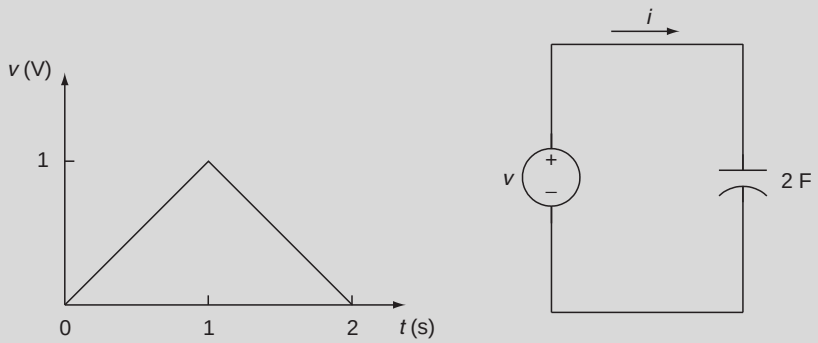


FIGURE 9.26 Circuit with a capacitor.

EXAMPLE PROBLEM 9.13

Find i for the following circuit.



Solution

For $t < 0$ and $t > 2$, $v = 0$ V, and therefore $i = 0$ in this interval. For nonzero values, the voltage waveform is described with two different functions: $v = t$ V for $0 \leq t \leq 1$, and $v = -(t-2)$V for $1 < t \leq 2$. Equation (9.24) is used to determine the current for each interval as follows.

For 0 < t < 1

$$i = C\frac{dv}{dt} = 2 \times \frac{d}{dt}(t) = 2 \text{ A}$$

For 1 ≤ t ≤ 2

$$i = C\frac{dv}{dt} = 2 \times \frac{d}{dt}(-(t-2)) = -2 \text{ A}$$

Voltage cannot change instantaneously across a capacitor. To have a step change in voltage across a capacitor, an infinite current must flow through the capacitor, and that is not physically possible. Of course, this is mathematically possible using a Dirac delta function.

Equation (9.24) defines the current through a capacitor for a given voltage. Suppose one is given a current through a capacitor and asked to find the voltage. To find the voltage, we start from Eq. (9.24) by multiplying both sides by dt, giving

$$i(t)dt = C\,dv$$

Integrating both sides yields

$$\int_{t_o}^{t} i(\lambda)d\lambda = C\int_{v(t_o)}^{v(t)} dv$$

or

$$v(t) = \frac{1}{C}\int_{t_o}^{t} idt + v(t_o) \tag{9.26}$$

For $t_0 = 0$, Eq. (9.26) reduces to

$$v(t) = \frac{1}{C}\int_{0}^{t} idt + v(0) \tag{9.27}$$

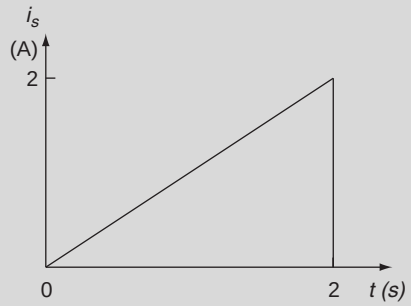
and for $t_0 = -\infty$, Eq. (9.27) reduces to

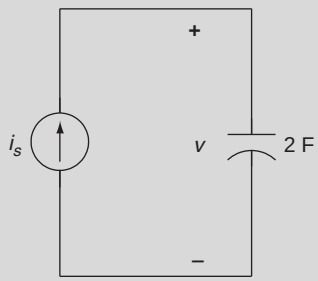
$$v(t) = \frac{1}{C}\int_{-\infty}^{t} i(\lambda)d\lambda \tag{9.28}$$

The initial voltage in Eq. (9.26), $v(t_0)$, is usually defined with the same polarity as v, which means $v(t_0)$ is a positive quantity. If the polarity of $v(t_0)$ is in the opposite direction of v, then $v(t_0)$ is negative.

EXAMPLE PROBLEM 9.14

Find v for the circuit that follows.





Solution

The current waveform is described with three different functions: for the interval $t \leq 0$, for the interval $0 < t \leq 2$, and for $t > 2$. To find the voltage, we apply Eq. (9.28) for each interval as follows.

For t < 0

$$v(t) = \frac{1}{C}\int_{-\infty}^{t} idt = \frac{1}{2}\int_{-\infty}^{0} 0dt = 0\ \text{V}$$

For $0 \leq t \leq 2$

$$v(t) = \frac{1}{C}\int_0^t i\,dt + v(0)$$

and with $v(0) = 0$, we have

$$v(t) = \frac{1}{2}\int_0^t \lambda\,d\lambda = \frac{1}{2}\left(\frac{\lambda^2}{2}\right)\Big|_0^t = \frac{t^2}{4}\ \text{V}$$

The voltage at $t = 2$ needed for the initial condition in the next part is

$$v(2) = \frac{t^2}{4}\Big|_{t=2} = 1\ \text{V}$$

For $t > 2$

$$v(t) = \frac{1}{C}\int_2^t i\,dt + v(2) = \frac{1}{2}\int_2^t 0\,dt + v(2) = 1\ \text{V}$$

9.10 A GENERAL APPROACH TO SOLVING CIRCUITS INVOLVING RESISTORS, CAPACITORS, AND INDUCTORS

Sometimes a circuit consisting of resistors, inductors, and capacitors cannot be simplified by bringing together like elements in series and parallel combinations. Consider the circuit shown in Figure 9.27. In this case, the absence of parallel or series combinations of resistors, inductors, or capacitors prevents us from simplifying the circuit for ease in solution. In this section, the node-voltage method is applied to write equations involving integrals and differentials using element relationships for resistors, inductors, and capacitors. From these equations, any unknown currents and voltages of interest can be solved using the standard differential equation approach.

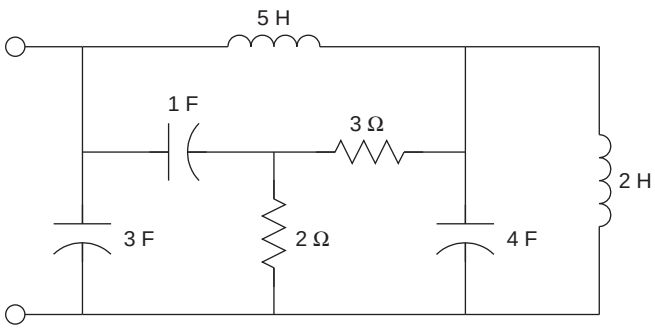
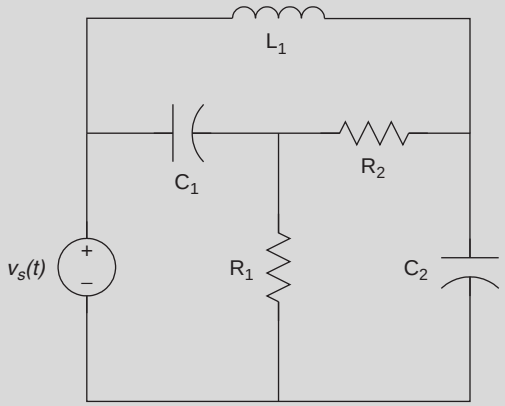


FIGURE 9.27 A circuit that cannot be simplified.

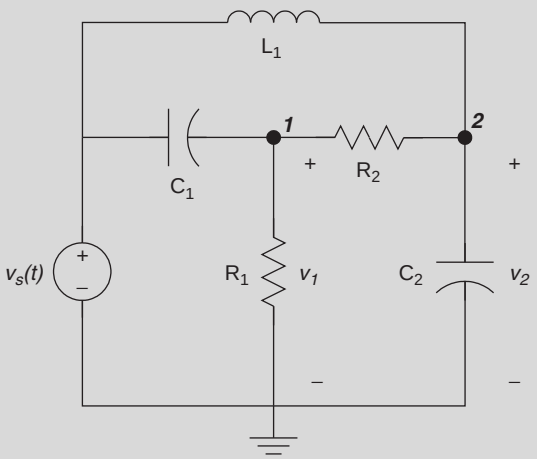
EXAMPLE PROBLEM 9.15

Write the node equations for the following circuit for $t \geq 0$ if the initial conditions are zero.



Solution

With the reference node at the bottom of the circuit, we have two essential nodes, as shown in the following redrawn circuit. Recall that the node involving the voltage source is a known voltage and that we do not write a node equation for it. When writing the node-voltage equations, the current through a capacitor is $i_c = C\,\Delta\dot{v}$, where $\Delta\dot{v}$ is the derivative of the voltage across the capacitor, and the current through an inductor is $i_L = \dfrac{1}{L}\displaystyle\int_0^t \Delta v\, d\lambda + i_L(0)$, where Δv is the voltage across the inductor. Since the initial conditions are zero, the term $i_L(0) = 0$.



Summing the currents leaving node 1 gives

$$C_1(\dot{v}_1 - \dot{v}_s) + \frac{v_1}{R_1} + \frac{v_1 - v_2}{R_2} = 0$$

which simplifies to

$$C_1\dot{v}_1 + \left(\frac{1}{R_1} + \frac{1}{R_2}\right)v_1 - \frac{1}{R_2}v_2 = C_1\dot{v}_s$$

Summing the currents leaving node 2 gives

$$\frac{v_2 - v_1}{R_2} + C_2\dot{v}_2 + \frac{1}{L_1}\int_0^t (v_2 - v_s)\, d\lambda = 0$$

Typically we eliminate integrals in the node equations by differentiating. When applied to the previous expression, this gives

$$\frac{1}{R_2}\dot{v}_2 - \frac{1}{R_2}\dot{v}_1 + C_2\ddot{v}_2 + \frac{1}{L_1}v_2 - \frac{1}{L_1}v_s = 0$$

and after rearranging yields

$$\ddot{v}_2 + \frac{1}{C_2R_2}\dot{v}_2 + \frac{1}{C_2L_1}v_2 - \frac{1}{C_2R_2}\dot{v}_1 = \frac{1}{C_2L_1}v_s$$

When applying the node-voltage method, we generate one equation for each essential node. To write a single differential equation involving just one node voltage and the inputs, we use the other node equations and substitute into the node equation of the desired node voltage. Sometimes this involves differentiation as well as substitution. The easiest case involves a node equation containing an undesired node voltage without its derivatives. Another method for creating a single differential equation is to use the D operator or the Laplace transform.

Consider the node equations for Example Problem 9.15, and assume that we are interested in obtaining a single differential equation involving node voltage v_1 and its derivatives, and the input. For ease in analysis, let us assume that the values for the circuit elements are $R_1 = R_2 = 1\ \Omega, C_1 = C_2 = 1$ F, and $L_1 = 1$ H, giving us

$$\dot{v}_1 + 2v_1 - v_2 = \dot{v}_s$$

and

$$\ddot{v}_2 + \dot{v}_2 + v_2 - \dot{v}_1 = v_s$$

Using the first equation, we solve for v_2, calculate $\dot{v}_2$ and $\ddot{v}_2$, and then substitute into the second equation as follows.

$$\begin{aligned} v_2 &= \dot{v}_1 + 2v_1 - \dot{v}_s \\ \dot{v}_2 &= \ddot{v}_1 + 2\dot{v}_1 - \ddot{v}_s \\ \ddot{v}_2 &= \dddot{v}_1 + 2\ddot{v}_1 - \dddot{v}_s \end{aligned}$$

After substituting into the second node equation, we have

$$\dddot{v}_1 + 2\ddot{v}_1 - \dddot{v}_s + \ddot{v}_1 + 2\dot{v}_1 - \ddot{v}_s + \dot{v}_1 + 2v_1 - \dot{v}_s - \dot{v}_1 = v_s$$

and after simplifying

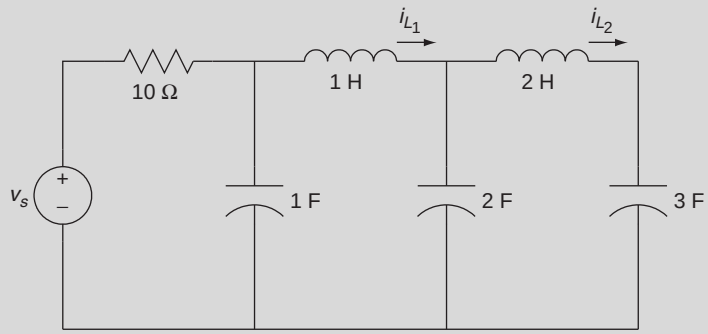
$$\dddot{v}_1 + 3\ddot{v}_1 + 2\dot{v}_1 + 2v_1 = \dddot{v}_s + \ddot{v}_s + \dot{v}_s - v_s$$

In general, the order of the differential equation relating a single output variable and the inputs is equal to the number of energy storing elements in the circuit (capacitors and inductors). In some circuits, the order of the differential equation is less than the number of capacitors and inductors in the circuit. This occurs when capacitor voltages and inductor currents are not independent; that is, there is an algebraic relationship between the capacitor—specifically, voltages and the inputs, or the inductor currents and the inputs. This occurs when capacitors are connected directly to a voltage source or when inductors are connected directly to a current source.

Example Problem 9.15 involved a circuit with zero initial conditions. When circuits involve nonzero initial conditions, our approach remains the same as before except that the initial inductor currents are included when writing the node-voltage equations.

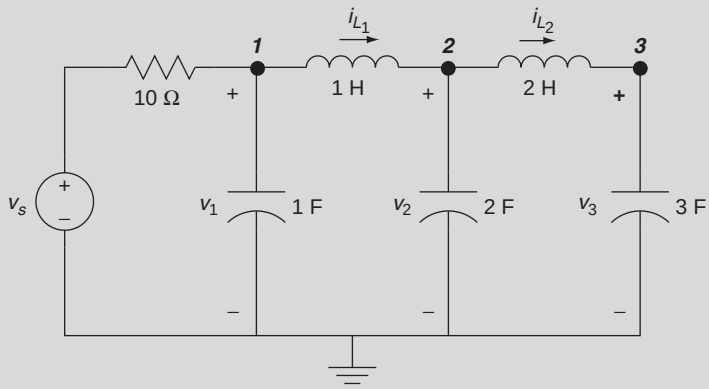
EXAMPLE PROBLEM 9.16

Write the node equations for the following circuit for $t \geq 0$ assuming the initial conditions are $i_{L_1}(0) = 8$ A and $i_{L_2}(0) = -4$ A.



Solution

With the reference node at the bottom of the circuit, there are three essential nodes, as shown in the redrawn circuit that follows.



Summing the currents leaving node 1 gives

$$\frac{(v_1 - v_s)}{10} + \dot{v}_1 + \int_0^t (v_1 - v_2)\, d\lambda + 8 = 0$$

where $i_{L_1}(0) = 8$ A.

Summing the currents leaving node 2 gives

$$\int_0^t (v_2 - v_1)\, d\lambda - 8 + 2\dot{v}_2 + \frac{1}{2}\int_0^t (v_2 - v_3)\, d\lambda - 4 = 0$$

where $i_{L_2}(0) = -4$ A. Notice that the sign for the initial inductor current is negative because the direction is from right to left and the current is defined on the circuit diagram in the opposite direction for the node 2 equation.

Summing the currents leaving node 3 gives

$$\frac{1}{2}\int_0^t (v_3 - v_2)\, d\lambda + 4 + 3\dot{v}_3 = 0$$

In this example, the node equations were not simplified by differentiating to remove the integral, which would have eliminated the initial inductor currents from the node equations. If we were to write a single differential equation involving just one node voltage and the input, a fifth-order differential equation would result because there are five energy storing elements in the circuit. To solve the differential equation, we would need five initial conditions, the initial node voltage for the variable selected, and the first through fourth derivatives at time zero.

9.10.1 Discontinuities and Initial Conditions in a Circuit

Discontinuities in voltage and current occur when an input such as a unit step is applied or a switch is thrown in a circuit. As we have seen, when solving an nth order differential equation, one must know n initial conditions, typically the output variable and its $(n-1)$ derivatives at the time the input is applied or the switch thrown. As we will see, if the inputs to a circuit are known for all time, we can solve for initial conditions directly based on energy considerations and not have to depend on being provided with them in the problem statement. Almost all of our problems involve the input applied at time zero, so our discussion here is focused on time zero, but it may be easily extended to any time an input is applied.

Energy cannot change instantaneously for elements that store energy. Thus, there are no discontinuities allowed in current through an inductor or voltage across a capacitor at any time—specifically, the value of the variable remains the same at $t = 0^-$ and $t = 0^+$. In the previous problem when we were given initial conditions for the inductors and capacitors, this implied, $i_{L_1}(0^-) = i_{L_1}(0^+)$ and $i_{L_2}(0^-) = i_{L_2}(0^+)$, and $v_1(0^-) = v_1(0^+), v_2(0^-) = v_2(0^+)$, and $v_3(0^-) = v_3(0^+)$. With the exception of variables associated with current through an inductor and voltage across a capacitor, other variables can have discontinuities, especially at a time when a unit step is applied or when a switch is thrown; however, these variables must obey KVL and KCL.

While it may not seem obvious at first, a discontinuity is allowed for the derivative of the current through an inductor and voltage across a capacitor at $t = 0^-$ and $t = 0^+$, since

$$\frac{di_L(0+)}{dt} = \frac{v_L(0+)}{L} \text{ and } \frac{dv_C(0+)}{dt} = \frac{i_C(0+)}{L}$$

as discontinuities are allowed in $v_L(0+)$ and $i_C(0+)$. Keep in mind that the derivatives in the previous expression are evaluated at zero after differentiation—that is,

$$\frac{di_L(0+)}{dt} = \left.\frac{di_L(t)}{dt}\right|_{t=0^+} \text{ and } \frac{dv_C(0+)}{dt} = \left.\frac{dv_C(t)}{dt}\right|_{t=0^+}$$

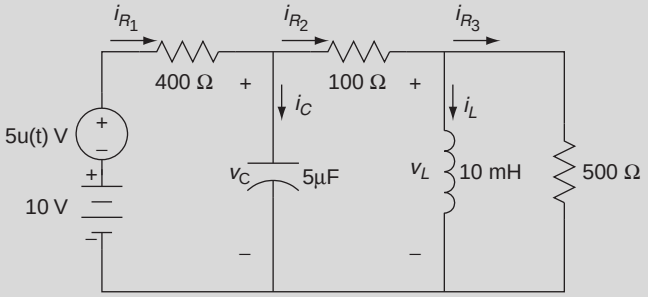
In calculations to determine the derivatives of variables not associated with current through an inductor and voltage across a capacitor, the derivative of a unit step input may be needed. Here we assume the derivative of a unit step input is zero at $t = 0^+$.

The initial conditions for variables not associated with current through an inductor and voltage across a capacitor at times of a discontinuity are determined only from the initial conditions from variables associated with current through an inductor and voltage across a capacitor and any applicable sources. The analysis is done in two steps involving KCL and KVL or using the node-voltage method.

1. First, we analyze the circuit at $t = 0^-$. Recall that when a circuit is at steady state, an inductor acts as a short circuit and a capacitor acts as an open circuit. Thus, at steady-state at $t = 0^-$, we replace all inductors by short circuits and capacitors by open circuits in the circuit. We then solve for the appropriate currents and voltages in the circuit to find the currents through the inductors (actually the shorts connecting the sources and resistors) and voltages across the capacitors (actually the open circuits among the sources and resistors).
2. Second, we analyze the circuit at $t = 0^+$. Since the inductor current cannot change in going from $t = 0^-$ to $t = 0^+$, we replace the inductors with current sources whose values are the currents at $t = 0^-$. Moreover, since the capacitor voltage cannot change in going from $t = 0^-$ to $t = 0^+$, we replace the capacitors with voltage sources whose values are the voltages at $t = 0^-$. From this circuit we solve for all desired initial conditions necessary to solve the differential equation.

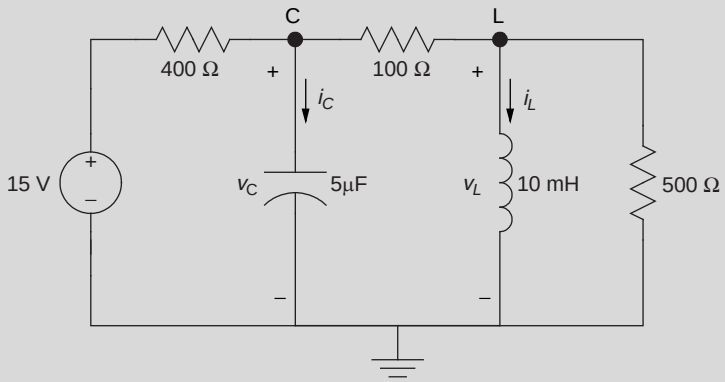
EXAMPLE PROBLEM 9.17

Use the node-voltage method to find v_c for the following circuit for $t \geq 0$.



Solution

For $t \geq 0$, the circuit is redrawn for analysis in the following figure.



Summing the currents leaving node C gives

$$\frac{v_C - 15}{400} + 5 \times 10^{-6}\dot{v}_C + \frac{v_C - v_L}{100} = 0$$

which simplifies to

$$\dot{v}_C + 2500v_C - 2000v_L = 7500$$

Summing the currents leaving node L gives

$$\frac{v_L - v_C}{100} + \frac{1}{10 \times 10^{-3}}\int_0^t v_L d\lambda + i_L(0^+) + \frac{v_L}{500} = 0$$

which, after multiplying by 500 and differentiating, simplifies to

$$6\dot{v}_L + 50 \times 10^3 v_L - 5\dot{v}_C = 0$$

Using the D operator method, the two differential equations are written as

$$Dv_C + 2500v_C - 2000v_L = 7500 \text{ or } (D + 2500)v_C - 2000v_L = 7500$$

$$6Dv_L + 50 \times 10^3 v_L - 5Dv_C = 0 \text{ or } \left(6D + 50 \times 10^3\right)v_L - 5Dv_C = 0$$

We then solve for v_L from the first equation,

$$v_L = \left(0.5 \times 10^{-3}D + 1.25\right)v_C - 3.75$$

and then substitute v_L into the second equation, giving

$$\left(6D + 50 \times 10^3\right)v_L - 5Dv_C = \left(6D + 50 \times 10^3\right)\left(\left(0.5 \times 10^{-3}D + 1.25\right)v_C - 3.750\right) - 5Dv_C = 0$$

Reducing this expression yields

$$D^2v_C + 10.417 \times 10^3 Dv_C + 20.83 \times 10^6 v_C = 62.5 \times 10^6$$

Continued

Returning to the time domain gives

$$\ddot{v}_C + 10.417 \times 10^3 \dot{v}_C + 20.83 \times 10^6 v_C = 62.5 \times 10^6$$

The characteristic equation for the previous differential equation is

$$s^2 + 10.417 \times 10^3 s + 20.833 \times 10^6 = 0$$

with roots -7.718×10^3 and -2.7×10^3 and the natural solution

$$v_{C_n}(t) = K_1 e^{-7.718 \times 10^3 t} + K_2 e^{-2.7 \times 10^3 t} \text{ V}$$

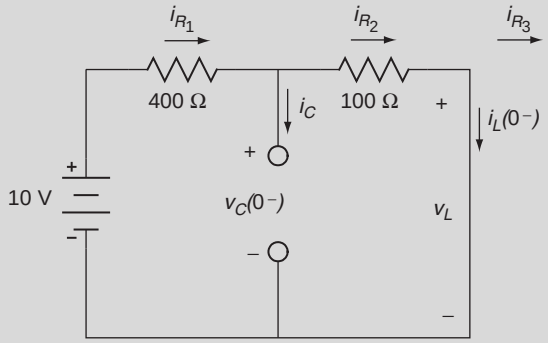
Next, we solve for the forced response, assuming that $v_{C_f}(t) = K_3$. After substituting into the differential equation, this gives

$$20.833 \times 10^6 K_3 = 62.5 \times 10^6$$

or $K_3 = 3$. Thus, our solution is now

$$v_C(t) = v_{C_n}(t) + v_{C_f}(t) = K_1 e^{-7.718 \times 10^3 t} + K_2 e^{-2.7 \times 10^3 t} + 3 \text{ V}$$

Initial conditions for $v_C(0^+)$ and $\dot{v}_C(0^+)$ are necessary to solve for K_1 and K_2. For $t = 0^-$, the capacitor is replaced by an open circuit and the inductor by a short circuit as shown in the following circuit.



Notice $v_L(0^-) = 0$ V because the inductor is a short circuit. Also note that the 500 Ω resistor is not shown in the circuit, since it is shorted out by the inductor, and so $i_{R_3}(0^-) = 0$A. Using the voltage divider rule, we have

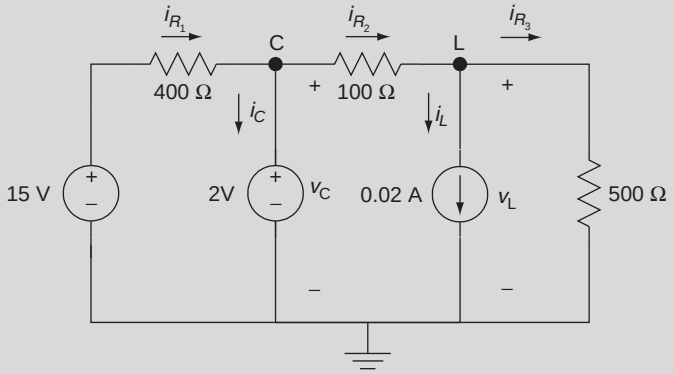
$$v_C(0^-) = 10 \times \frac{100}{400 + 100} = 2 \text{ V}$$

and by Ohm's law

$$i_L(0^-) = \frac{10}{100 + 400} = 0.02 \text{ A}$$

It follows that $i_{R_1}(0^-) = i_{R_2}(0^-) = i_L(0^-) = 0.02$ A. Because voltage across a capacitor and current through an inductor are not allowed to change from $t = 0^-$ to $t = 0^+$ we have $v_C(0^+) = v_C(0^-) = 2$V and $i_L(0^+) = i_L(0^-) = 0.02$ A.

The circuit for $t = 0^+$ is drawn by replacing the inductors in the original circuit with current sources whose values equal the inductor currents at $t = 0^-$ and the capacitors with voltage sources whose values equal the capacitor voltages at $t = 0^-$, as shown in the following figure with nodes C and L and reference. Note also that the input is now $10 + 5u(t) = 15$ V.



To find $v_L(0^+)$, we sum the currents leaving node L, yielding

$$\frac{v_L - 2}{100} + 0.02 + \frac{v_L}{500} = 0$$

which gives $v_L(0^+) = 0$ V. Now $i_{R_3}(0^+) = \dfrac{v_L(0^+)}{500} = 0$ A, $i_{R_2}(0^+) = 0.02 + i_{R_3}(0^+) = 0.02$ A, and $i_{R_1}(0^+) = \dfrac{15 - 2}{400} = 0.0325$ A.

To find $i_C(0^+)$, we write KCL at node C, giving

$$-i_{R_1}(0^+) + i_C(0^+) + i_{R_2}(0^+) = 0$$

or

$$i_C(0^+) = i_{R_1}(0^+) - i_{R_2}(0^+) = 0.0325 - 0.02 = 0.125 \text{ A}$$

To find $\dot{v}_C(0^+)$, note that $i_C(0^+) = C\dot{v}_C(0^+)$ or

$$\dot{v}_C(0^+) = \frac{i_C(0^+)}{C} = \frac{0.0125}{5 \times 10^{-6}} = 2.5 \times 10^3 \frac{\text{V}}{\text{s}}.$$

With the initial conditions, the constants K_1 and K_2 are solved as

$$v_C(0) = 2 = K_1 + K_2 + 3$$

Next,

$$\dot{v}_C(t) = -7.718 \times 10^3 K_1 e^{-7.718 \times 10^3 t} - 2.7 \times 10^3 K_2 e^{-2.7 \times 10^3 t}$$

Continued

and at $t = 0$,

$$\dot{v}_C(0) = 2.5 \times 10^3 = -7.718 \times 10^3 K_1 - 2.7 \times 10^3 K_2$$

Solving gives $K_1 = 0.04$ and $K_2 = -1.04$. Substituting these values into the solution gives

$$v_C(t) = 0.04e^{-7.718\times10^3 t} - 1.04e^{-2.7\times10^3 t} + 3 \text{ V}$$

for $t \geq 0$.

9.11 OPERATIONAL AMPLIFIERS

Section 9.3 introduced controlled voltage and current sources that are dependent on a voltage or current elsewhere in a circuit. These devices were modeled as a two-terminal device. In this section, we look at the operational amplifier, also known as an op amp, which is a multiterminal device. An operational amplifier is an electronic device that consists of large numbers of transistors, resistors, and capacitors. Fully understanding its operation requires knowledge of diodes and transistors—topics that are not covered in this book. However, fully understanding how an operational amplifier operates in a circuit involves a topic already covered: the controlled voltage source. Circuits involving operational amplifiers form the cornerstone for any bioinstrumentation, from amplifiers to filters. Amplifiers used in biomedical applications have very high-input impedance to keep the current drawn from the system being measured low. Most body signals have very small magnitudes. For example, an ECG has a magnitude in the millivolts, and the EEG has a magnitude in the microvolts. Analog filters are often used to remove noise from a signal, typically through frequency domain analysis to design the filter.

As the name implies, the operation amplifier is an amplifier, but as we will see, when it is combined with other circuit elements, it integrates, differentiates, sums, and subtracts. One of the first operational amplifiers appeared as an eight-lead dual-in-line package (DIP), shown in Figure 9.28.

Differing from previous circuit elements, this device has two input and one output terminals. Rather than draw the operational amplifier using Figure 9.28, the operational amplifier is drawn with the symbol in Figure 9.29. The input terminals are labeled the noninverting input (+) and the inverting input (–). The power supply terminals are labeled V+ and V–, which are frequently omitted, since they do not affect the circuit behavior except in saturation conditions, as will be described. Most people shorten the name of the operational amplifier to the "op amp."

Figure 9.30 shows a model of the op amp, focusing on the internal behavior of the input and output terminals. The input-output relationship is

$$v_o = A(v_p - v_n) \tag{9.29}$$

Since the internal resistance is very large, we will replace it with an open circuit to simplify analysis, leaving us with the op amp model show in Figure 9.31.

With the replacement of the internal resistance with an open circuit, the currents $i_n = i_p = 0$ A. In addition, current i_A, the current flowing out of the op amp, is not zero. Because i_A is unknown, seldom is KCL applied at the output junction. In solving op amp problems, KCL is almost always applied at input terminals.

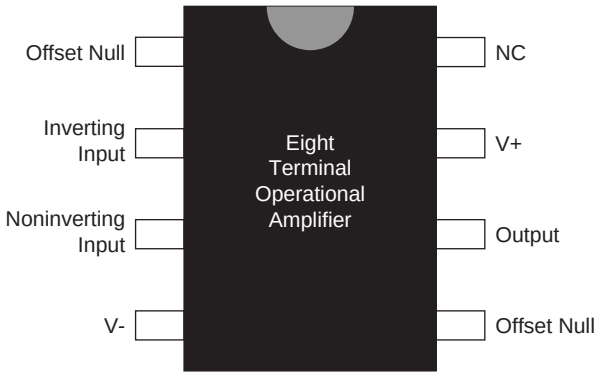


FIGURE 9.28 An eight-terminal operational amplifier. The terminal NC is not connected, and the two terminal offset nulls are used to correct imperfections (typically not connected). V+ and V– are terminal power to provide energy to the circuit. Keep in mind that a ground exists for both V+ and V–, a ground that is shared by other elements in the circuit. Modern operational amplifiers have ten or more terminals.

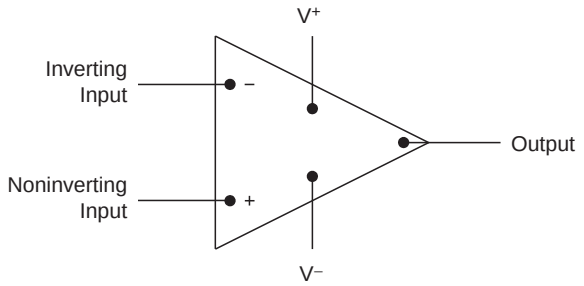


FIGURE 9.29 Circuit element symbol for the operational amplifier.

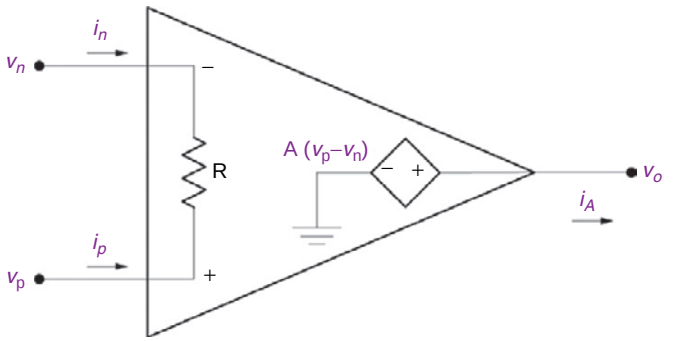


FIGURE 9.30 An internal model of the op amp. The internal resistance between the input terminals, R, is very large, exceeding 1 MΩ. The gain of the amplifier, A, is also large, exceeding 10^4. Power supply terminals are omitted for simplicity.

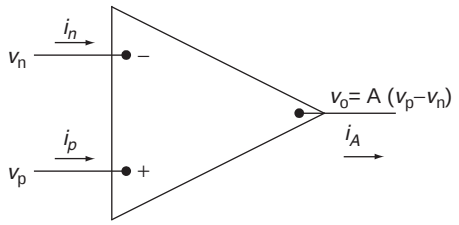
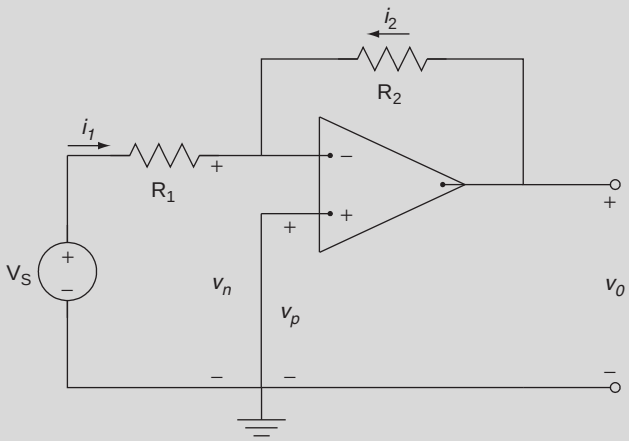


FIGURE 9.31 Idealized model of the op amp with the internal resistance, R, replaced by an open circuit.

EXAMPLE PROBLEM 9.18

Find v_0 for the following circuit.



Solution

Using the op amp model of Figure 9.31, we apply KCL at the inverting terminal giving

$$-i_1 - i_2 = 0$$

since no current flows into the op amp's input terminals. Replacing the current using Ohm's law gives

$$\frac{v_s - v_n}{R_1} + \frac{v_o - v_n}{R_2} = 0$$

Multiplying by R_1R_2 and collecting like terms, we have

$$R_2 v_s = (R_1 + R_2)v_n - R_1 v_o$$

Now $v_o = A(v_p - v_n)$, and since the noninverting terminal is connected to ground, $v_p = 0$,

$$v_o = -Av_n$$

or

$$v_n = -\frac{v_o}{A}$$

Substituting v_n into the KCL inverting input equation gives

$$R_s v_s = (R_1 + R_2)\left(-\frac{v_o}{A}\right) - R_1 v_o$$

$$= \left(\frac{R_1 + R_2}{A} + R_1\right) v_o$$

or

$$v_o = \frac{-R_2 v_s}{\left(R_1 + \dfrac{R_1 + R_2}{A}\right)}$$

As A goes to infinity, the previous equation goes to

$$v_o = -\frac{R_2}{R_1} v_s$$

Interestingly, with A going to infinity, v_0 remains finite due to the resistor R_2. This happens because a negative feedback path exists between the output and the inverting input terminal through R_2. This circuit is called an inverting amplifier with an overall gain of $-\frac{R_2}{R_1}$.

An operational amplifier with a gain of infinity is known as an ideal op amp. Because of the infinite gain, there must be a feedback path between the output and input, and we cannot connect a voltage source directly between the inverting and noninverting input terminals. When analyzing an ideal op amp circuit, we simplify the analysis by letting

$$v_n = v_p$$

Consider the previous example. Because $v_p = 0$, $v_n = 0$. Applying KCL at the inverting input gives

$$-\frac{v_s}{R_1} + \frac{-v_o}{R_2} = 0$$

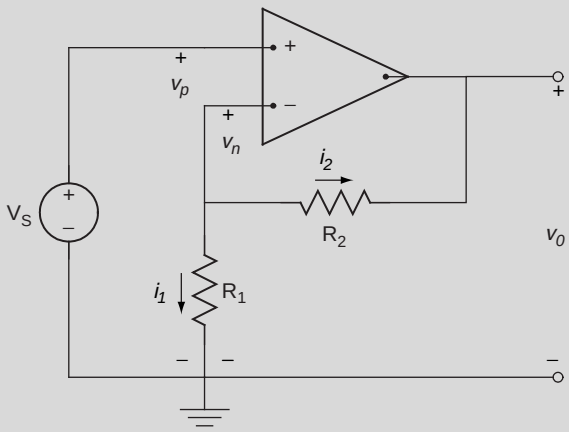
or

$$v_o = -\frac{R_2}{R_1} v_s$$

Notice how simple the analysis becomes when we assume $v_n = v_p$. Keep in mind that this approximation is valid as long as A is very large (infinity) and a feedback is included.

EXAMPLE PROBLEM 9.19

Find the overall gain for the following circuit.



Solution

Assuming the op amp is ideal, we start with $v_n = v_p$. Then, since the op amp's noninverting terminal is connected to the source, $v_n = v_p = v_s$. Because no current flows into the op amp, by KCL we have

$$i_1 + i_2 = 0$$

and

$$\frac{v_s}{R_1} + \frac{v_s - v_o}{R_2} = 0$$

or

$$v_o = \left(\frac{R_1 + R_2}{R_1}\right) v_s$$

The overall gain is

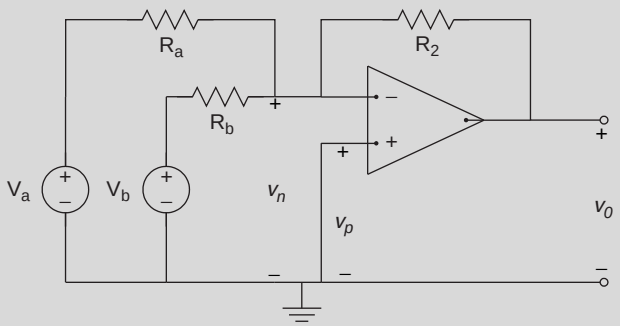
$$\frac{v_o}{v_s} = \frac{R_1 + R_2}{R_1}$$

This circuit is a noninverting op amp circuit used to amplify the source input. Amplifiers are used in almost all clinical instrumentation for ECG, EEG, EOG, and so on.

Example Problem 9.20 describes a summing op amp circuit.

EXAMPLE PROBLEM 9.20

Find the overall gain for the following circuit.



Solution

As before, we start the solution with $v_n = v_p$ and note that the noninverting input is connected to ground, yielding $v_n = v_p = 0$ V. Applying KCL at the inverting input node gives

$$-\frac{V_a}{R_a} - \frac{V_b}{R_b} - \frac{v_o}{R_2} = 0$$

or

$$v_o = -\left(\frac{R_2}{R_a}V_a + \frac{R_2}{R_b}V_b\right)$$

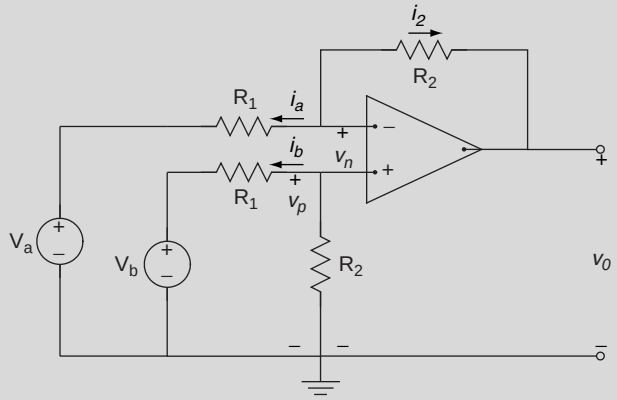
This circuit is a weighted summation of the input voltages. We can add additional source resistor inputs so that in general

$$v_o = -\left(\frac{R_2}{R_a}V_a + \frac{R_2}{R_b}V_b + \ldots + \frac{R_2}{R_m}V_m\right)$$

The op amp circuit in Example Problem 9.21 provides an output proportional to the difference of two input voltages. This op amp is often referred to as a differential amplifier.

EXAMPLE PROBLEM 9.21

Find the overall gain for the following circuit.



Solution

Assuming an ideal op amp, we note no current flows into the input terminals and that $v_n = v_p$. Apply KCL at the inverting input terminal gives

$$i_a = -i_2$$

or

$$\frac{v_n - V_a}{R_1} + \frac{v_n - v_o}{R_2} = 0$$

and

$$(R_1 + R_2)v_n - R_2 V_a = R_1 v_o$$

The previous equation involves two unknowns, so we need another equation easily found by applying the voltage divider at the noninverting input.

$$v_p = \frac{R_2}{R_1 + R_2} v_b = v_n$$

Substituting this result for v_n into the KCL equation at the inverting terminal gives

$$R_2 V_b - R_2 V_a = R_1 v_o$$

or

$$v_o = \frac{R_2}{R_1}(V_b - V_a)$$

As shown, this op amp circuit, also known as the differential amplifier, subtracts the weighted input signals. This amplifier is used for bipolar measurements involving ECG and EEG, since the typical recording is obtained between two bipolar input terminals. Ideally, the measurement

contains only the signal of interest uncontaminated by noise from the environment. The noise is typically called a *common-mode signal*. A common-mode signal comes from lighting, 60-Hz power line signals, inadequate grounding, and power supply leakage. A differential amplifier with appropriate filtering can reduce the impact of a common-mode signal.

The response of a differential amplifier can be decomposed into differential-mode and common-mode components:

$$v_{dm} = v_b - v_a$$

and

$$v_{cm} = \frac{(v_a + v_b)}{2}$$

As described, the common-mode signal is the average of the input voltages. Using the two previous equations, one can solve v_a and v_b in terms of v_{dm} and v_{cm} as

$$v_a = v_{cm} - \frac{v_{dm}}{2}$$

and

$$v_b = v_{cm} + \frac{v_{dm}}{2}$$

When substituted into the response in Example Problem 9.21, we get

$$v_0 = \left(\frac{R_1R_2 - R_1R_2}{R_1(R_1 + R_2)}\right)v_{cm} + \left(\frac{R_2(R_1 + R_2) + R_2(R_1 + R_2)}{2R_1(R_1 + R_2)}\right)v_{dm} = A_{cm}v_{cm} + A_{dm}v_{dm}$$

Notice the term multiplying v_{cm}, A_{cm}, is zero, characteristic of the ideal op amp that amplifies only the differential-mode of the signal. Since real amplifiers are not ideal and resistors are not truly exact, the common-mode gain is not zero. So when one designs a differential amplifier, the goal is to keep A_{cm} as small as possible and A_{dm} as large as possible.

The rejection of the common-mode signal is called *common-mode rejection*, and the measure of how ideal the differential amplifier is called the *common-mode rejection ratio*, given as

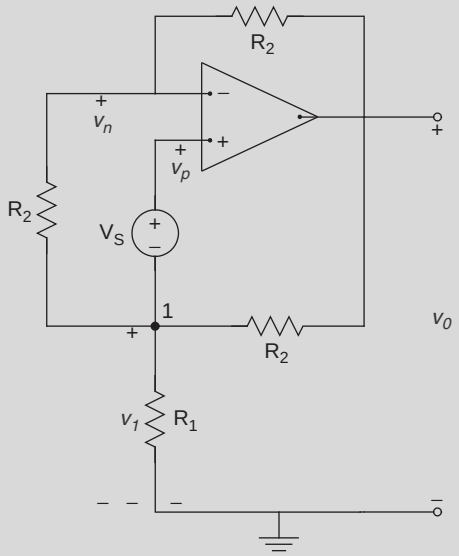
$$CMRR = 20\log_{10}\left|\frac{A_{dm}}{A_{cm}}\right|$$

where the larger the value of *CMRR*, the better. Values of *CMRR* for a differential amplifier for EEG, ECG, and EMG are 100 to 120 db.

The general approach to solving op amp circuits is to first assume that the op amp is ideal and $v_p = v_n$. Next, we apply KCL or KVL at the two input terminals. In more complex circuits, we continue to apply our circuit analysis tools to solve the problem, as Example Problem 9.22 illustrates.

EXAMPLE PROBLEM 9.22

Find v_0 for the following circuit.



Solution

With $v_n = v_p$, we apply KCL at the inverting input

$$\frac{v_n - v_1}{R_2} + \frac{v_n - v_o}{R_2} = 0$$

and

$$2v_n - v_1 - v_o = 0$$

Next, we apply KVL from ground to node 1 to the noninverting input and back to ground, giving

$$-v_1 - V_s + v_p = 0$$

and with $v_n = v_p$, we have $v_n - v_1 = V_s$.

Now we apply KCL at node 1, noting no current flows into the noninverting input terminal:

$$\frac{v_1}{R_1} + \frac{v_1 - v_o}{R_2} + \frac{v_1 - v_n}{R_2} = 0$$

Combining like terms in the previous equation gives

$$-R_1 v_n + (2R_1 + R_2)v_1 - R_1 v_o = 0$$

With three equations and three unknowns, we first eliminate v_1 by subtracting the inverting input KCL equation by the KVL equation, giving

$$v_1 = v_o - 2V_s$$

Next, we eliminate v_n by substituting v_1 into the inverting input KCL equation, as follows:

$$v_n = \frac{1}{2}(v_1 + v_o)$$

$$= \frac{1}{2}(v_o - 2V_s + v_o)$$

$$= v_o - V_s$$

Finally, we substitute the solutions for v_1 and v_n into the node 1 KCL equation, giving

$$-R_1 v_n + (2R_1 + R_2)v_1 - R_1 v_o = 0$$
$$-R_1(v_o - V_s) + (2R_1 + R_2)(v_o - 2V_s) - R_1 v_o = 0$$

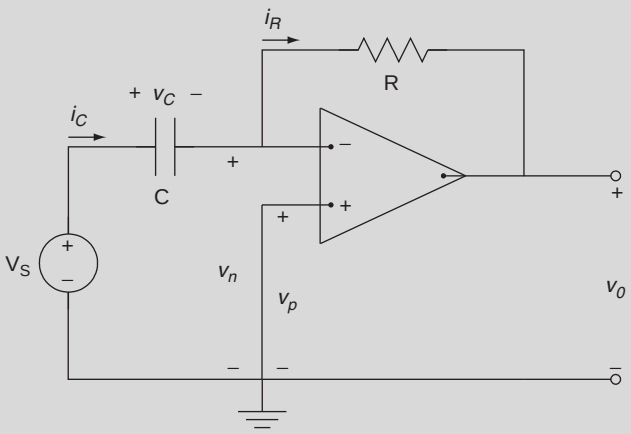
After simplification, we have

$$v_o = \frac{(3R_1 + 2R_2)}{R_2} V_s$$

Example Problems 9.23 and 9.24 illustrate an op amp circuit that differentiates and integrates by using a capacitor.

EXAMPLE PROBLEM 9.23

Find v_0 for the following circuit.



Solution

With the noninverting input connected to ground, we have $v_p = 0 = v_n$. From KVL

$$v_C = V_s$$

Continued

and it follows that

$$i_C = C\frac{dv_C}{dt} = C\frac{dV_s}{dt}$$

Since no current flows into the op amp, $i_C = i_R$. With

$$i_R = \frac{v_n - v_o}{R} = -\frac{v_o}{R}$$

and

$$i_C = C\frac{dV_s}{dt} = i_R = -\frac{v_o}{R}$$

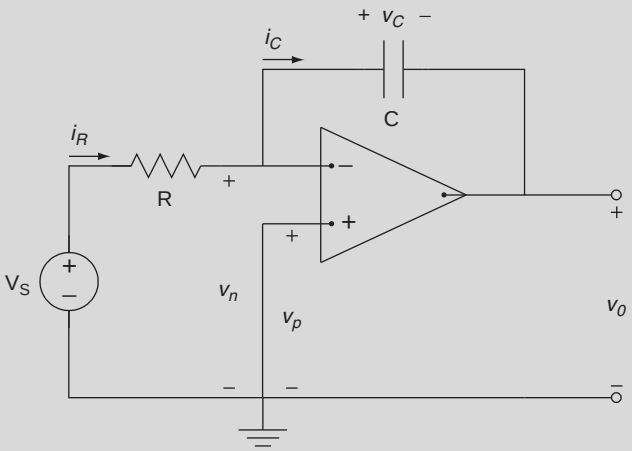
we have

$$v_o = -RC\frac{dV_s}{dt}$$

If $R = \frac{1}{C}$, the circuit in this example differentiates the input, $v_o = -\frac{dV_s}{dt}$.

EXAMPLE PROBLEM 9.24

Find v_0 for the following circuit.



Solution

It follows that

$$v_n = v_p = 0$$

and

$$i_C = i_R = \frac{V_s}{R}$$

Therefore,

$$v_C = \frac{1}{C}\int_{-\infty}^{t} i_C d\lambda = \frac{1}{C}\int_{-\infty}^{t} \frac{V_s}{R} d\lambda$$

From KVL, we have

$$v_C + v_o = 0$$

and

$$v_o = -\frac{1}{RC}\int_{-\infty}^{t} V_s d\lambda$$

With $R = \frac{1}{C}$, the circuit operates as an integrator

$$v_o = -\int_{-\infty}^{t} V_s d\lambda$$

9.11.1 Voltage Characteristics of the Op Amp

In the preceding examples involving the op amp, we did not consider the supply voltage (shown in Figure 9.29) and that the output voltage of an ideal op amp is constrained to operate between the supply voltages V^+ and V^-. If analysis determines v_0 is greater than V^+, v_0 saturates at V^+. If analysis determines v_0 is less than V^-, v_0 saturates at V^-. The output voltage characteristics are shown in Figure 9.32.

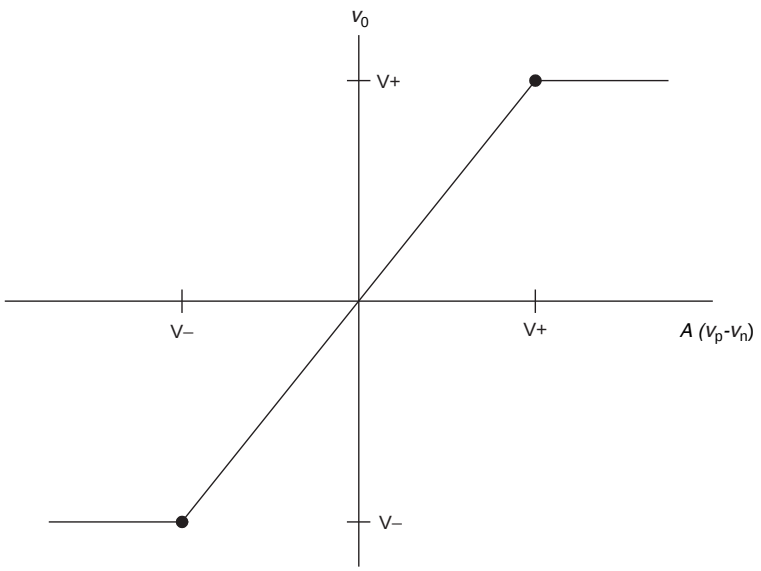


FIGURE 9.32 Voltage characteristics of an op amp.

EXAMPLE PROBLEM 9.25

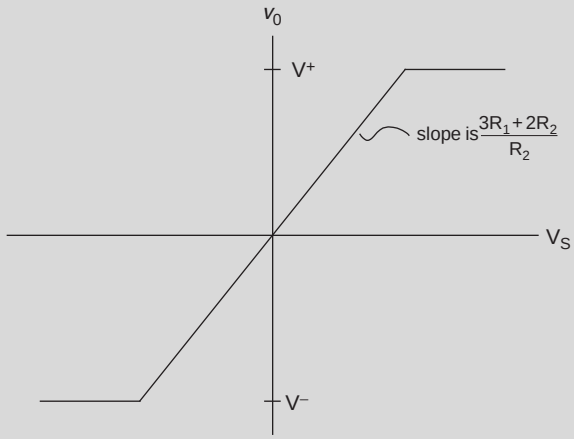
For the circuit shown in Example Problem 9.22, let $V^+ = +10$ V and $V^- = -10$ V. Graph the output voltage characteristics of the circuit.

Solution

The solution for Example Problem 9.22 is

$$v_o = \left(\frac{3R_1 + 2R_2}{R_2}\right) V_s$$

which saturates whenever v_0 is less than V^- and greater than V^+, as shown in the following graph.



9.12 TIME-VARYING SIGNALS

An alternating current (*a-c*) or sinusoidal source of 50 or 60 Hz is common throughout the world as a power source supplying energy for most equipment and other devices. While most of this chapter has focused on the transient response, when dealing with sinusoidal sources, attention is now focused on the steady-state or forced response. In bioinstrumentation, analysis in the steady-state simplifies the design by focusing only on the steady-state response, which is where the device actually operates. A sinusoidal voltage source is a time-varying signal given by

$$v_s = V_m \cos(\omega t + \phi) \tag{9.30}$$

where the voltage is defined by angular frequency (ω in radians/s), phase angle (ϕ in radians or degrees), and peak magnitude (V_m). The period of the sinusoid T is related to frequency f (Hz or cycles/s) and angular frequency by

$$\omega = 2\Pi f = \frac{2\Pi}{T} \tag{9.31}$$

An important metric of a sinusoid is its *rms value* (square *r*oot of the *m*ean value of the *s*quared function), given by

$$V_{rms} = \sqrt{\frac{1}{T}\int_0^T V_m^2 \cos^2(\omega t + \phi)dt} \tag{9.32}$$

which reduces to $V_{rms} = \dfrac{V_m}{\sqrt{2}}$.

To appreciate the response to a time-varying input, $v_s = V_m \cos(\omega t + \phi)$, consider the circuit shown in Figure 9.33, in which the switch is closed at $t = 0$ and there is no initial energy stored in the inductor. Applying KVL to the circuit gives

$$L\frac{di}{dt} + iR = V_m \cos(\omega t + \phi)$$

and after some work, the solution is

$$\begin{aligned} i &= i_n + i_f \\ &= \frac{-V_m}{\sqrt{R^2 + \omega^2 L^2}} \cos\left(\phi - \frac{\omega L}{R}\right) e^{-\frac{R}{L}t} + \frac{V_m}{\sqrt{R^2 + \omega^2 L^2}} \cos\left(\omega t + \phi - \frac{\omega L}{R}\right) \end{aligned}$$

The first term is the natural response that goes to zero as t goes to infinity. The second term is the forced response that has the same form as the input (i.e., a sinusoid with the same frequency ω, but a different phase angle and maximum amplitude). If all you are interested in is the steady-state response, as in most bioinstrumentation applications, then the only unknowns are the response amplitude and phase angle. The remainder of this section deals with techniques involving the *phasor* to efficiently find these unknowns.

9.12.1 Phasors

The phasor is a complex number that contains amplitude and phase angle information of a sinusoid and for the signal in Eq. (9.30) is expressed as

$$\mathbf{V} = V_m e^{j\phi} = V_m \angle \phi \tag{9.33}$$

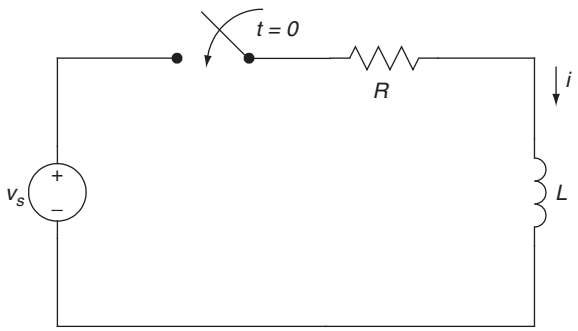


FIGURE 9.33 An RL circuit with sinusoidal input.

In Eq. (9.33), by practice, the angle in the exponential is written in radians, and in the $\angle$ notation is written in degrees. Work in the phasor domain involves the use of complex algebra in moving between the time and phasor domain, so the rectangular form of the phasor is also used, given as

$$\mathbf{V} = V_m(\cos\phi + j\sin\phi) \tag{9.34}$$

9.12.2 Passive Circuit Elements in the Phasor Domain

To use phasors with passive circuit elements for steady-state solutions, the relationship between voltage and current is needed for the resistor, inductor, and capacitor. Assume that

$$i = I_m\cos(\omega t + \theta)$$

$$I = I_m\angle\theta = I_m e^{j\theta}$$

For a resistor,

$$v = IR = RI_m\cos(\omega t + \theta)$$

and the phasor of v is

$$\mathbf{V} = RI_m\angle\theta = R\mathbf{I} \tag{9.35}$$

where $\mathbf{I} = I_m\angle\theta$. Note that there is no phase shift for the relationship between the phasor current and voltage for a resistor.

For an inductor,

$$v = L\frac{di}{dt} = -\omega LI_m\sin(\omega t + \theta) = -\omega LI_m\cos(\omega t + \theta - 90^\circ)$$

and the phasor of v is

$$\begin{aligned}\mathbf{V} &= -\omega LI_m\angle\theta - 90^\circ = -\omega LI_m e^{j(\theta-90^\circ)}\\ &= -\omega LI_m e^{j\theta}e^{-j90^\circ} = -\omega LI_m e^{j\theta}(-j)\\ &= j\omega LI_m e^{j\theta}\\ &= j\omega L\mathbf{I}\end{aligned} \tag{9.36}$$

Note that inductor current and voltage are out of phase by 90°—that is, current lags behind voltage by 90°.

For a capacitor, define $v = V_m\cos(\omega t + \theta)$ and $\mathbf{V} = V_m\angle\theta$. Now

$$\begin{aligned}i &= C\frac{dv}{dt} = C\frac{d}{dt}(V_m\cos(\omega t + \theta))\\ &= -CV_m\omega\sin(\omega t + \theta) = -CV_m\omega\cos(\omega t + \theta - 90^\circ)\end{aligned}$$

and the phasor for i is

$$\mathbf{I} = -\omega C V_m \underline{|\theta - 90^\circ} = -\omega C V_m e^{j\theta} e^{\ j90^\circ}$$

$$= -\omega C V_m e^{j\theta} (\cos(90^\circ) - j\sin(90^\circ))$$

$$= j\omega C V_m e^{j\theta}$$

$$= j\omega C \mathbf{V}$$

or

$$\mathbf{V} = \frac{1}{j\omega C}\mathbf{I} = \frac{-j}{\omega C}\mathbf{I} \tag{9.37}$$

Note that capacitor current and voltage are out of phase by 90°—that is, voltage lags behind current by 90°.

Equations (9.35)–(9.37) all have the form of $\mathbf{V}=Z\mathbf{I}$, where Z represents the impedance of the circuit element and is, in general, a complex number, with units of ohms. The impedance for the resistor is R, the inductor, $j\omega L$, and the capacitor, $\frac{-j}{\omega C}$. The impedance is a complex number and not a phasor even though it may look like one. The imaginary part of the impendence is called reactance.

The final part to working in the phasor domain is to transform a circuit diagram from the time to phasor domain. For example, the circuit shown in Figure 9.34 is transformed into the phasor domain shown in Figure 9.35 by replacing each circuit element with their impedance equivalent and sources by their phasor. For the voltage source, we have

$$v_s = 100\sin 500t = 100\cos(500t - 90^\circ)\ \text{mV} \qquad \leftrightarrow \qquad 500\underline{|-90^\circ}\ \text{mV}$$

For the capacitor, we have

$$0.5\ \mu\text{F} \qquad \leftrightarrow \qquad \frac{-j}{\omega C} = -j4000\ \Omega$$

For the resistor, we have

$$1000\ \Omega \qquad \leftrightarrow \qquad 1000\ \Omega$$

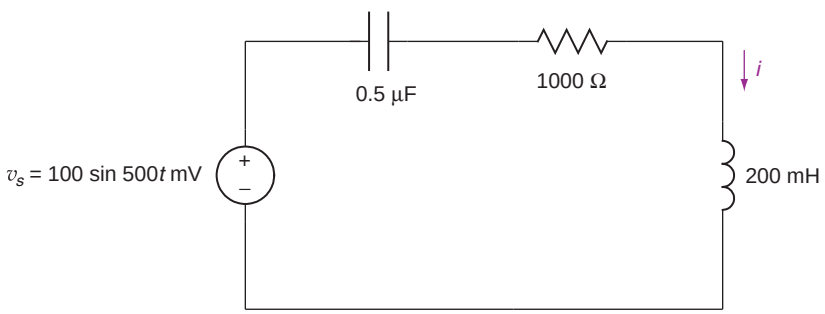


FIGURE 9.34 A circuit diagram.

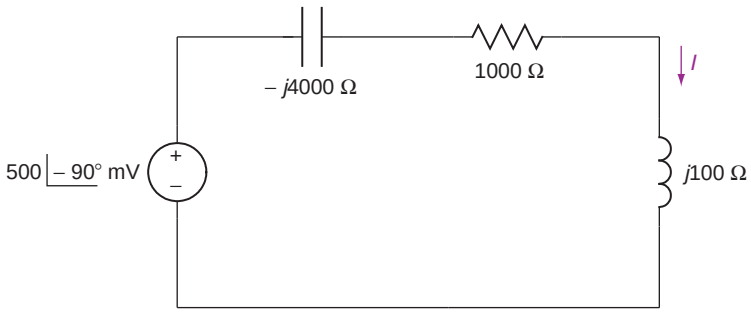


FIGURE 9.35 Phasor and impedance equivalent circuit for Figure 9.34.

For the inductor, we have

$$200 \text{ mH} \quad \leftrightarrow \quad j\omega L = j100\ \Omega$$

Each of the elements is replaced by its phasor and impedance equivalents, as shown in Figure 9.35.

9.12.3 Kirchhoff's Laws and Other Techniques in the Phasor Domain

It is fortunate that all of the material presented before in this chapter involves Kirchhoff's current and voltage laws, and all of the other techniques apply to phasors. That is, for KVL, the sum of phasor voltages around any closed path is zero

$$\sum \mathbf{V}_i = 0 \tag{9.38}$$

and for KCL, the sum of phasor currents leaving any node is zero

$$\sum \mathbf{I}_i = 0 \tag{9.39}$$

Impedances in series are given by

$$Z = Z_1 + \cdots + Z_n \tag{9.40}$$

Impedances in parallel are given by

$$Z = \frac{1}{\frac{1}{Z_1} + \cdots + \frac{1}{Z_n}} \tag{9.41}$$

The node-voltage method, superposition and Thévenin equivalent circuits are also applicable in the phasor domain. Example Problems 9.26 and 9.27 illustrate the process, with the most difficult aspect involving complex algebra.

EXAMPLE PROBLEM 9.26

For the circuit shown in Figure 9.35, find the steady-state response i.

Solution

The impedance for the circuit is

$$Z = -j4000 + 1000 + j100 = 1000 - j3900\ \Omega$$

Using Ohm's law,

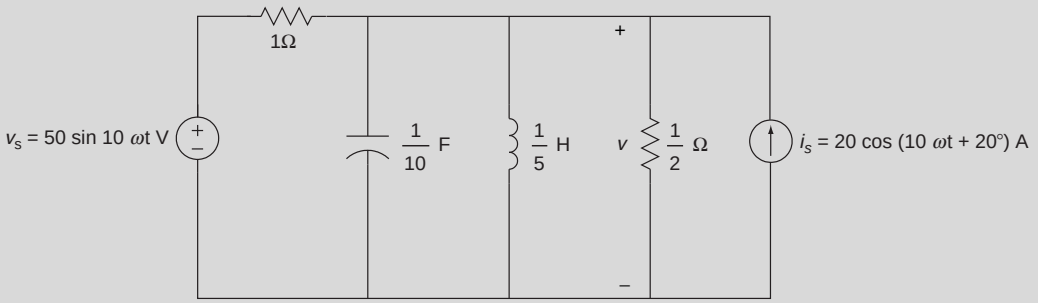
$$\mathbf{I} = \frac{\mathbf{V}}{Z} = \frac{0.5\angle -90^\circ}{1000 - j3900} = \frac{0.5\angle -90^\circ}{4026\angle -76^\circ} = 124\angle -14^\circ\ \mu\text{A}$$

Returning to the time domain, the steady-state current is

$$i = 124\cos(500t - 14^\circ)\ \mu\text{A}$$

EXAMPLE PROBLEM 9.27

Find the steady-state response v using the node-voltage method for the following circuit.



Solution

The first step is to transform the circuit elements into their impedances, which for the capacitor and inductor are

$$\frac{1}{10}\ \text{F} \quad \leftrightarrow \quad \frac{-j}{\omega C} = -j\ \Omega$$

$$\frac{1}{5}\ \text{H} \quad \leftrightarrow \quad j\omega L = j2\ \Omega$$

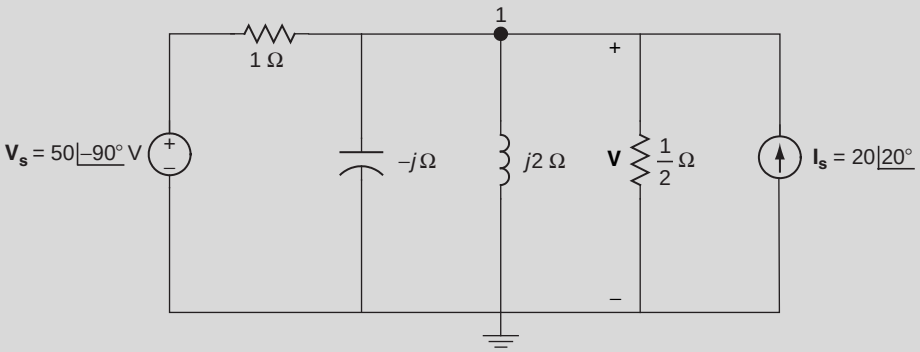
The phasors for the two sources are

$$v_s = 50\sin\omega t\ \text{V} \quad \leftrightarrow \quad \mathbf{V}_s = 50\angle -90^\circ\ \text{V}$$

$$i_s = 20\cos(\omega t + 20^\circ)\ \text{A} \quad \leftrightarrow \quad \mathbf{I}_s = 20\angle 20^\circ$$

Continued

Since the two resistors retain their values, the phasor drawing of the circuit is shown in the following figure with the ground at the lower node.



Writing the node-voltage equation for node 1 gives

$$\mathbf{V} - 50\underline{|-90^\circ} + \frac{\mathbf{V}}{j} + \frac{\mathbf{V}}{j2} + 2\mathbf{V} - 20\underline{|20^\circ} = 0$$

Collecting like terms, converting to rectangular form, and converting to polar form gives

$$\mathbf{V}\left(3 + \frac{j}{2}\right) = 50\underline{|-90^\circ} + 20\underline{|20^\circ}$$

$$\mathbf{V}\left(3 + \frac{j}{2}\right) = -50j + 18.8 + j6.8 = 18.8 - j43.2$$

$$\mathbf{V} \times 3.04\underline{|9.5^\circ} = 47.1\underline{|-66.5^\circ}$$

$$\mathbf{V} = \frac{47.1\underline{|-66.5^\circ}}{3.04\underline{|9.5^\circ}} = 15.5\underline{|-76^\circ}$$

The steady-state solution is

$$v = 15.6\cos(10t - 76^\circ)\ \text{V}$$

9.13 ACTIVE ANALOG FILTERS

This section presents several active analog filters involving the op amp. Passive analog filters use passive circuit elements: resistors, capacitors, and inductors. To improve performance in a passive analog filter, the resistive load at the output of the filter is usually increased. By using the op amp, fine control of the performance is achieved without increasing the load at the output of the filter. Filters are used to modify the measured signal by removing noise. A filter is designed in the frequency domain so the measured signal to be retained is passed through and noise is rejected.

Figure 9.36 shows the frequency characteristics of four filters: low-pass, high-pass, band-pass, and notch filters. The signal that is passed through the filter is indicated by the frequency interval called the passband. The signal that is removed by the filter is indicated by the frequency interval called the stopband. The magnitude of the filter, $|H(j\omega)|$, is one in the passband and zero in the stopband. The low-pass filter allows slowly changing signals with frequency less than ω_1 to pass through the filter and eliminates any signal or noise above ω_1. The high-pass filter allows quickly changing signals with frequency greater than ω_2 to pass through the filter and eliminates any signal or noise with frequency less than ω_2. The band-pass filter allows signals in the frequency band greater than ω_1 and less than ω_2 to pass through the filter and eliminates any signal or noise outside this interval. The notch filter allows signals in the frequency band less than ω_1 and greater than ω_2 to pass through

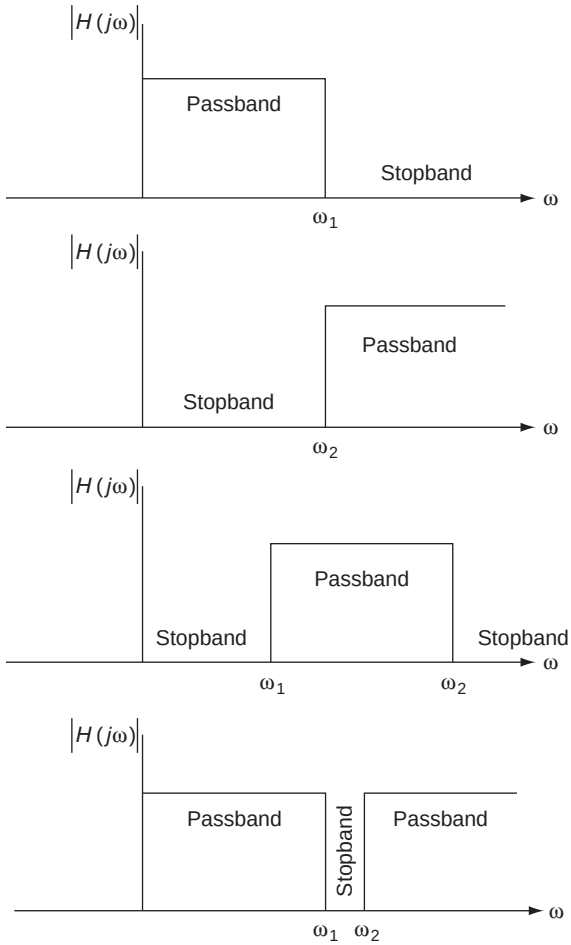


FIGURE 9.36 Ideal magnitude-frequency response for four filters, from top to bottom: low-pass, high-pass, band-pass, and notch.

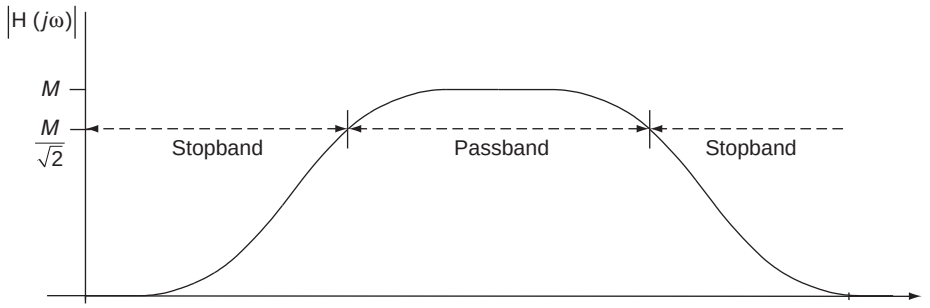


FIGURE 9.37 A realistic magnitude-frequency response for a band-pass filter. Note that the magnitude M does not necessarily need to be one. The passband is defined as the frequency interval when the magnitude is greater than $\frac{M}{\sqrt{2}}$.

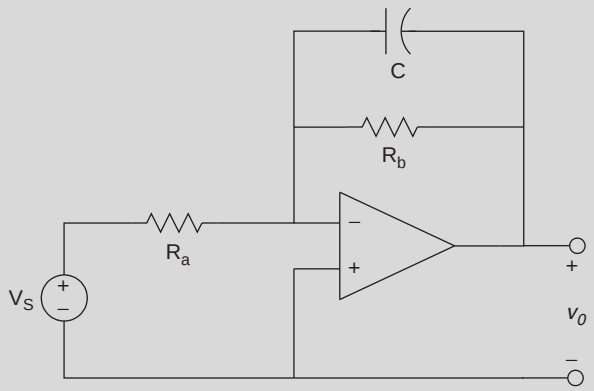
the filter and eliminates any signal or noise outside this interval. The frequencies ω_1 and ω_2 are typically called cutoff frequencies for the low-pass and high-pass filters.

In reality, any real filter cannot possibly have these ideal characteristics but instead has a smooth transition from the passband to the stopband, as shown, for example, in Figure 9.37 (the reason for this behavior is discussed in Chapter 11). Further, it is sometimes convenient to include both amplification and filtering in the same circuit, so the maximum of the magnitude does not need to be one, but it can be a value of M specified by the needs of the application.

To determine the filter's performance, the filter is driven by a sinusoidal input. One varies the input over the entire spectrum of interest (at discrete frequencies) and records the output magnitude. The critical frequencies are when $|H(j\omega)| = \dfrac{M}{\sqrt{2}}$.

EXAMPLE PROBLEM 9.28

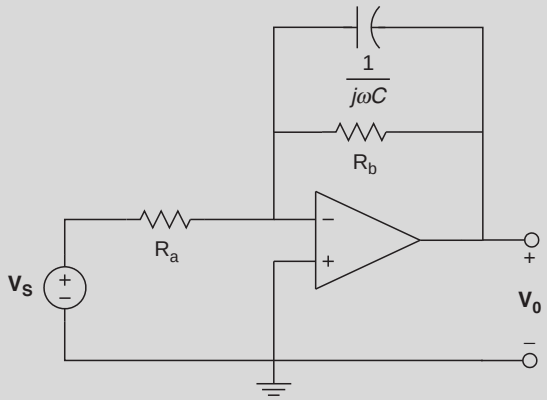
Using the low-pass filter in the following circuit, design the filter to have a gain of 5 and a cutoff frequency of 500 $\frac{\text{rad}}{\text{s}}$.



Solution

By treating the op amp as ideal, note that the noninverting input is connected to ground and, therefore, the inverting input is also connected to ground. The operation of this filter is readily apparent because at low frequencies, the capacitor acts like an open circuit, reducing the circuit to an inverting amplifier that passes low-frequency signals. At high frequencies, the capacitor acts like a short circuit, which connects the output terminal to the inverting input and ground.

The phasor method will be used to solve this problem by first transforming the circuit into the phasor domain, as shown in the following figure.



Summing the currents leaving the inverting input gives

$$-\frac{\mathbf{V}_s}{R_a}-\frac{\mathbf{V}_0}{\dfrac{1}{j\omega C}}-\frac{\mathbf{V}_0}{R_b}=0$$

Collecting like terms and rearranging yields

$$-\mathbf{V}_0\left(\frac{1}{\dfrac{1}{j\omega C}}+\frac{1}{R_b}\right)=\frac{\mathbf{V}_s}{R_a}$$

After further manipulation,

$$\frac{\mathbf{V}_0}{\mathbf{V}_s}=-\frac{1}{R_a}\left(\frac{1}{\dfrac{1}{\dfrac{1}{j\omega C}}+\dfrac{1}{R_b}}\right)=-\frac{1}{R_a}\left(\frac{1}{j\omega C+\dfrac{1}{R_b}}\right)$$

$$\frac{\mathbf{V}_0}{\mathbf{V}_s}=-\frac{1}{R_aC}\left(\frac{1}{j\omega+\dfrac{1}{R_bC}}\right)$$

Continued

Similar to the reasoning for the characteristic equation for a differential equation, the cutoff frequency is defined as $\omega_c = \frac{1}{R_b C}$, (i.e., the denominator term, $j\omega + \frac{1}{R_b C}$ is set equal to zero). Thus, with the cutoff frequency set at $\omega_c = 500\frac{\text{rad}}{\text{s}}$, then $\frac{1}{R_b C} = 500$. The cutoff frequency is also defined as when $|H(j\omega)| = \frac{M}{\sqrt{2}}$, where $M = 5$. The magnitude of $\frac{\mathbf{V}_0}{\mathbf{V}_s}$ is given by

$$\left|\frac{\mathbf{V}_0}{\mathbf{V}_s}\right| = \frac{\frac{1}{R_a C}}{\sqrt{\omega^2 + \left(\frac{1}{R_b C}\right)^2}}$$

and at the cutoff frequency, $\omega_c = 500\frac{\text{rad}}{\text{s}}$,

$$\frac{5}{\sqrt{2}} = \frac{\frac{1}{R_a C}}{\sqrt{\omega_c^2 + \left(\frac{1}{R_b C}\right)^2}}$$

With $\frac{1}{R_b C} = 500$, the magnitude is

$$\frac{5}{\sqrt{2}} = \frac{\frac{1}{R_a C}}{\sqrt{\omega_c^2 + \left(\frac{1}{R_b C}\right)^2}} = \frac{\frac{1}{R_a C}}{\sqrt{500^2 + 500^2}} = \frac{\frac{1}{R_a C}}{500\sqrt{2}}$$

which gives

$$R_a C = \frac{1}{2500}$$

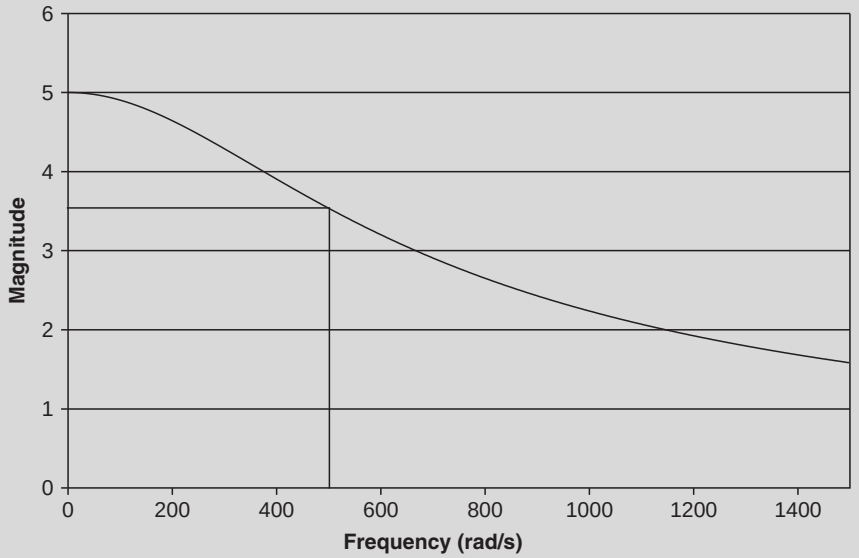
Since we have three unknowns and two equations ($R_a C = \frac{1}{2500}$ and $\frac{1}{R_b C} = 500$), there are an infinite number of solutions. Therefore, one can select a convenient value for one of the elements—say, $R_a = 20\ \text{k}\Omega$—and the other two elements are determined as

$$C = \frac{1}{2500 \times R_a} = \frac{1}{2500 \times 20000} = 20\ \text{nF}$$

and

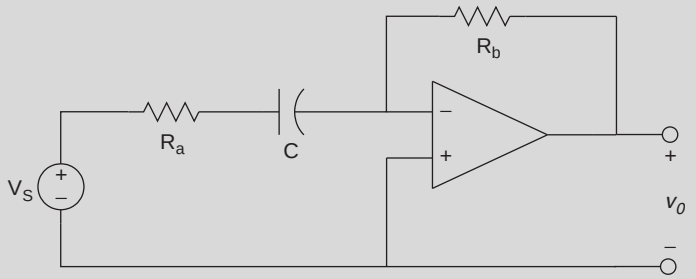
$$R_b = \frac{1}{500 \times C} = \frac{1}{500 \times 20 \times 10^{-9}} = 100\ \text{k}\Omega$$

A plot of the magnitude versus frequency is shown in the following figure. As can be seen, the cutoff frequency gives a value of magnitude equal to 3.53 at 100 Hz, which is the design goal.



EXAMPLE PROBLEM 9.29

Using the high-pass filter in the following circuit, design the filter to have a gain of 5 and a cutoff frequency of 100 $\frac{\text{rad}}{\text{s}}$.

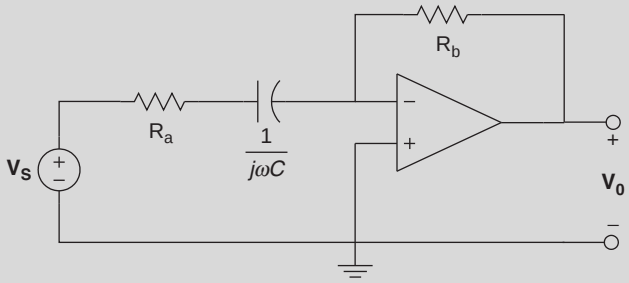


Solution

Since the op amp is assumed ideal and the noninverting input is connected to ground, the inverting input is also connected to ground. The operation of this filter is readily apparent because at low frequencies, the capacitor acts like an open circuit, so no input voltage is seen at the noninverting input. Since there is no input, then the output is zero. At high frequencies, the capacitor acts like a short circuit, which reduces the circuit to an inverting amplifier that passes through high-frequency signals.

Continued

As before, the phasor method will be used to solve this problem by first transforming the circuit into the phasor domain, as shown in the following figure.



Summing the currents leaving the inverting input gives

$$-\frac{\mathbf{V}_s}{R_a+\dfrac{1}{j\omega C}}-\frac{\mathbf{V}_0}{R_b}=0$$

Rearranging yields

$$\frac{\mathbf{V}_0}{\mathbf{V}_s}=-\frac{R_b}{R_a+\dfrac{1}{j\omega C}}=-\frac{R_b}{R_a}\frac{j\omega}{j\omega+\dfrac{1}{R_aC}}$$

At cutoff frequency $\omega_c=100\dfrac{\text{rad}}{\text{s}}=\dfrac{1}{R_aC}$. The magnitude of $\dfrac{\mathbf{V}_0}{\mathbf{V}_s}$ is given by

$$\left|\frac{\mathbf{V}_0}{\mathbf{V}_s}\right|=\frac{R_b}{R_a}\frac{\omega}{\sqrt{\omega^2+\left(\dfrac{1}{R_aC}\right)^2}}$$

and at the cutoff frequency,

$$\frac{5}{\sqrt{2}}=\frac{R_b}{R_a}\frac{\omega_c}{\sqrt{\omega_c^2+\left(\dfrac{1}{R_aC}\right)^2}}$$

With $\dfrac{1}{R_aC}=100$ and $\omega_c=100\dfrac{\text{rad}}{\text{s}}$, gives

$$\frac{5}{\sqrt{2}}=\frac{R_b}{R_a}\frac{\omega_c}{\sqrt{\omega_c^2+\left(\dfrac{1}{R_aC}\right)^2}}=\frac{R_b}{R_a}\frac{\dfrac{1}{R_aC}}{\sqrt{100^2+100^2}}=\frac{R_b}{R_a}\frac{100}{100}=\frac{R_b}{\sqrt{2}R_a}$$

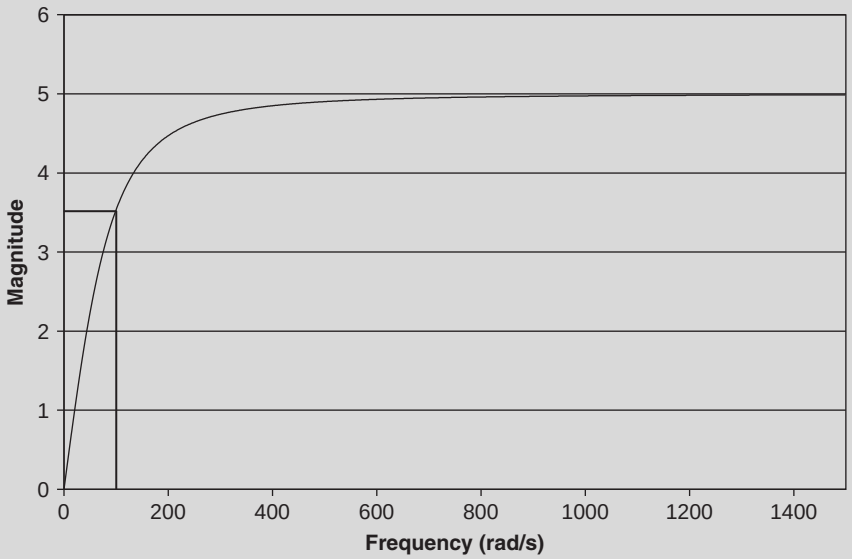
Thus, $\dfrac{R_b}{R_a}=5$. Since we have three unknowns and two equations, one can select a convenient value for one of the elements—say, $R_b=20\ \text{k}\Omega$—and the other two elements are determined as

$$R_a=\frac{R_b}{5}=\frac{20000}{5}=4\ \text{k}\Omega$$

and

$$C = \frac{1}{100R_a} = \frac{1}{100 \times 4000} = 2.5\ \mu\text{F}$$

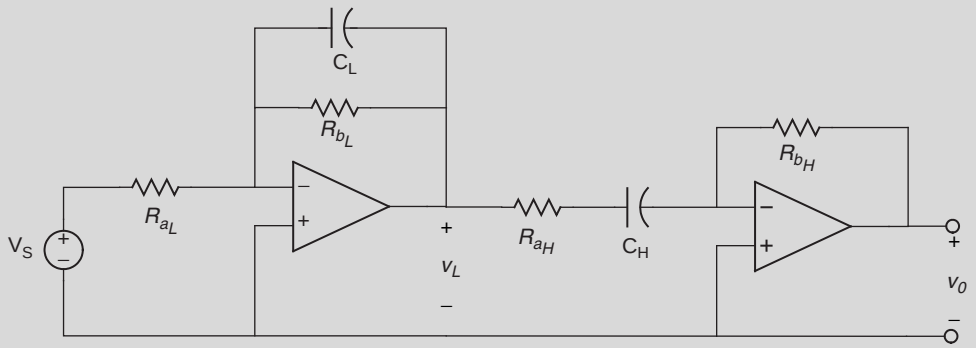
A plot of the magnitude versus frequency is shown in the following figure. As can be seen, the cutoff frequency gives a value of magnitude equal to 3.53 at 100 Hz, which is the design goal.



Example Problem 9.30 demonstrates the technique to create band-pass filters (which require two cutoff frequencies).

EXAMPLE PROBLEM 9.30

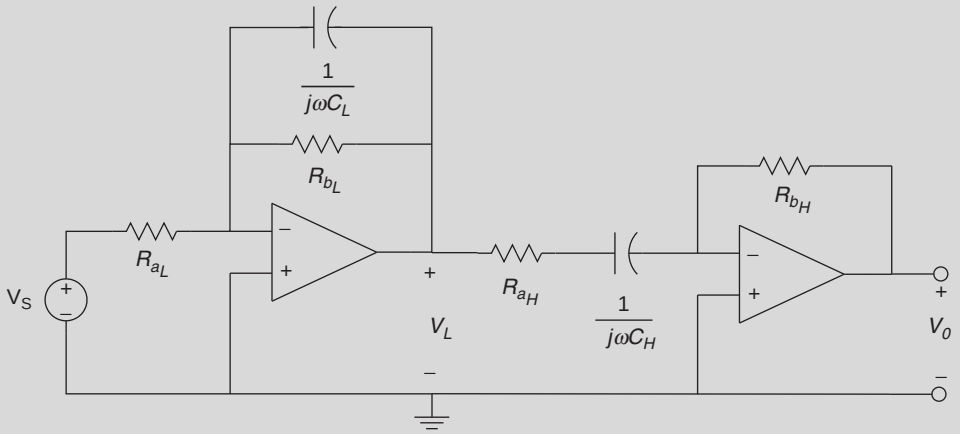
Using the band-pass filter in the following circuit, design the filter to have a gain of 5 and pass through frequencies from 100 to 500 $\frac{\text{rad}}{\text{s}}$.



Continued

Solution

As usual, the design of the filter is done in the phasor domain and uses the work done in the previous two examples. Note that the elements around the op amp on the left are the low-pass filter circuit elements, and those on the right are the high-pass filter circuit elements. In fact, when working with op amps, filters can be cascaded together to form other filters, so a low-pass and high-pass filter cascaded together will form a band-pass. The phasor domain circuit is given in the next figure.



As before the noninverting input to the op amps is connected to ground, which means that the inverting input is also connected to ground. Summing the currents leaving the inverting input for each op amp gives

$$-\frac{\mathbf{V}_s}{R_{a_L}}-\frac{\mathbf{V}_L}{\dfrac{1}{j\omega C_L}}-\frac{\mathbf{V}_L}{R_{b_L}}=0$$

$$-\frac{\mathbf{V}_L}{R_{a_H}+\dfrac{1}{j\omega C_H}}-\frac{\mathbf{V}_0}{R_{b_H}}=0$$

Solving the first equation for $\mathbf{V}_L$ gives

$$\mathbf{V}_L=-\frac{1}{R_{a_L}C_L}\left(\frac{1}{j\omega+\dfrac{1}{R_{b_L}C_L}}\right)\mathbf{V}_s$$

Solving the second equation for $\mathbf{V}_0$ gives

$$\mathbf{V}_0=-\frac{R_{b_H}}{R_{a_H}}\frac{j\omega}{j\omega+\dfrac{1}{R_{a_H}C_H}}\mathbf{V}_L$$

Substituting $\mathbf{V}_\mathrm{L}$ into the previous equation yields

$$\mathbf{V}_0 = \frac{R_{b_H}}{R_{a_H}} \frac{j\omega}{j\omega + \dfrac{1}{R_{a_H}C_H}} \times \frac{1}{R_{a_L}C_L} \left(\frac{1}{j\omega + \dfrac{1}{R_{b_L}C_L}} \right) \mathbf{V}_\mathrm{s}$$

The form of the solution is simply the product of each filter. The magnitude of the filter is

$$\left|\frac{\mathbf{V}_0}{\mathbf{V}_\mathrm{s}}\right| = \frac{R_{b_H}}{R_{a_H}} \frac{\omega}{\sqrt{\omega^2 + \left(\dfrac{1}{R_{a_H}C_H}\right)^2}} \frac{\dfrac{1}{R_{a_L}C_L}}{\sqrt{\omega^2 + \left(\dfrac{1}{R_{b_L}C_L}\right)^2}}$$

Since there are two cutoff frequencies, two equations evolve:

$$\omega_{c_H} = \frac{1}{R_{a_H}C_H} = 100 \frac{\text{rad}}{\text{s}}$$

and

$$\omega_{c_L} = \frac{1}{R_{b_L}C_L} = 500 \frac{\text{rad}}{\text{s}}$$

At either cutoff frequency, the magnitude is $\dfrac{5}{\sqrt{2}}$, such that at $\omega_{c_H} = 100 \dfrac{\text{rad}}{\text{s}}$

$$\frac{5}{\sqrt{2}} = \frac{R_{b_H}}{R_{a_H}} \frac{\omega_{c_H}}{\sqrt{\omega_{c_H}^2 + \left(\dfrac{1}{R_{a_H}C_H}\right)^2}} \frac{\dfrac{1}{R_{a_L}C_L}}{\sqrt{\omega_{C_H}^2 + \left(\dfrac{1}{R_{b_L}C_L}\right)^2}}$$

$$= \frac{R_{b_H}}{R_{a_H}} \frac{100}{\sqrt{100^2 + 100^2}} \frac{\dfrac{1}{R_{a_L}C_L}}{\sqrt{100^2 + 500^2}}$$

Therefore,

$$500\sqrt{26} = \frac{R_{b_H}}{R_{a_H}R_{a_L}C_L}$$

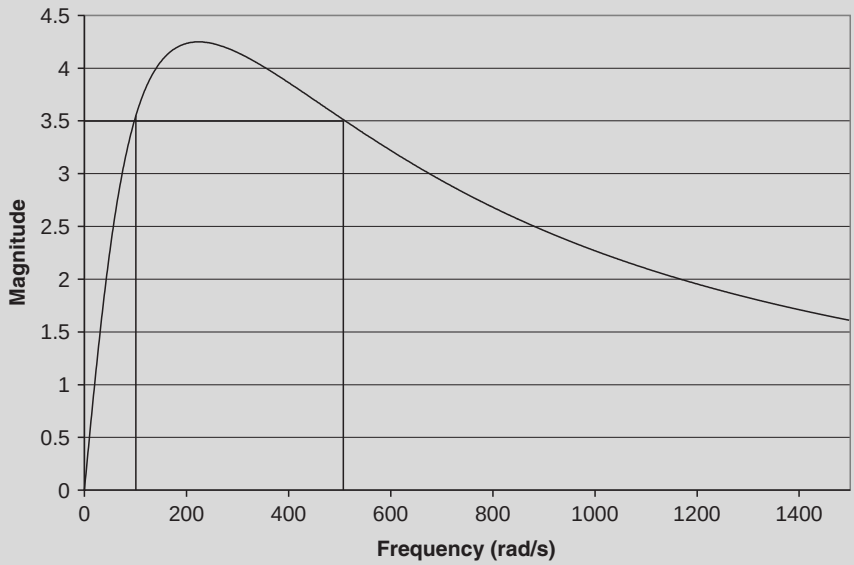
The other cutoff frequency gives the same result as the previous equation. There are now three equations $\left(\dfrac{1}{R_{a_H}C_H} = 100,\ \dfrac{1}{R_{b_L}C_L} = 500, \text{ and } 500\sqrt{26} = \dfrac{R_{b_H}}{R_{a_H}R_{a_L}C_L} \right)$, and six unknowns. For convenience, set $R_{b_L} = 100\ \text{k}\Omega$ and $R_{a_H} = 100\ \text{k}\Omega$, which gives $C_L = \dfrac{1}{500R_{b_L}} = 20\ \text{nF}$ and $C_H = \dfrac{1}{100R_{a_H}} = 0.1\ \mu\text{F}$. Now from $500\sqrt{26} = \dfrac{R_{b_H}}{R_{a_H}} \dfrac{1}{R_{a_L}C_L}$,

$$\frac{R_{b_H}}{R_{a_L}} = 500\sqrt{26}C_L R_{a_H} = 5.099$$

Once again, one can specify one of the resistors—say, $R_{a_L} = 10\ \text{k}\Omega$—giving $R_{b_H} = 50.099\ \text{k}\Omega$.

Continued

A plot of the magnitude versus frequency is shown in the following figure. As can be seen, the cutoff frequency gives a value of magnitude equal to 3.53 at 500 Hz, which is the design goal.



None of the filters in Example Problems 9.28–9.30 have the ideal characteristics of Figure 9.36. To improve the performance from the pass-band to stopband in a low-pass filter with a sharper transition, one can cascade identical filters together—that is, connect the output of the first filter to the input of the next filter and so on. The more cascaded filters, the better the performance. The magnitude of the overall filter is the product of the individual filter magnitudes.

While this approach is appealing for improving the performance of the filter, the overall magnitude of the filter does not remain a constant in the pass-band. Better filters are available with superior performance, such as a Butterworth filter. Two Butterworth filters are shown in Figure 9.38. Analysis of these filters is carried out in Exercises 53 and 55.

9.14 BIOINSTRUMENTATION DESIGN

Figure 9.2 described the various elements needed in a biomedical instrumentation system. The purpose of this type of instrument is to monitor the output of a sensor or sensors and to extract information from the signals that are produced by the sensors.

Acquiring a discrete-time signal and storing this signal in computer memory from a continuous-time signal is accomplished with an analog-to-digital (A/D) converter. The A/D converter uniformly samples the continuous-time waveform and transforms it into a sequence of numbers, one every t_k *seconds.* The A/D converter also transforms the

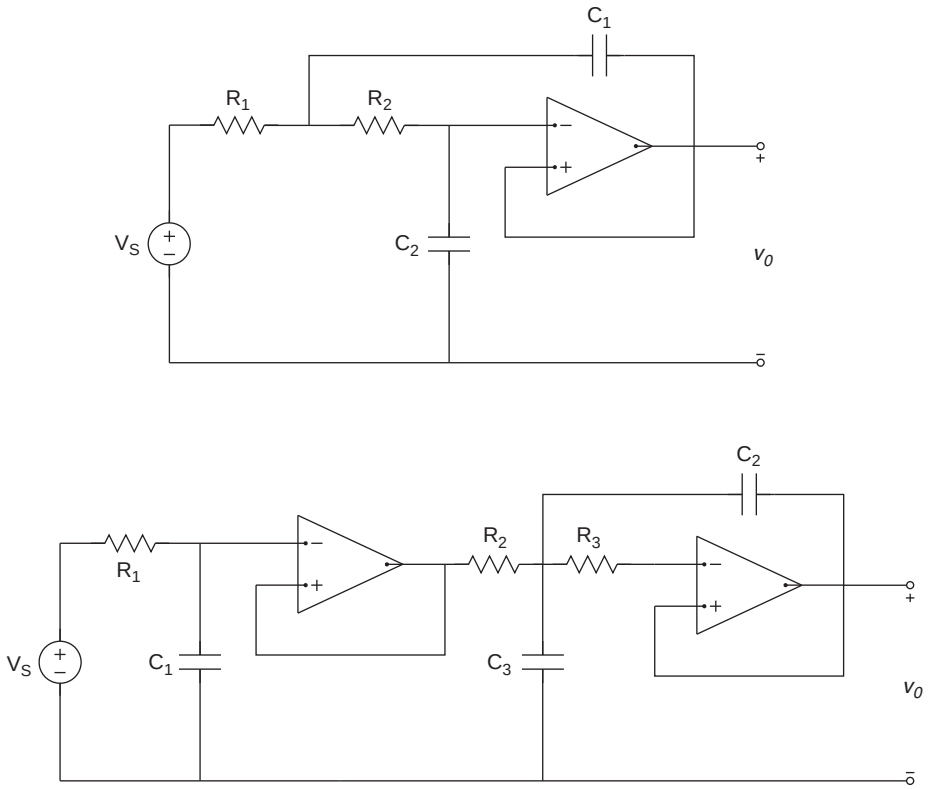


FIGURE 9.38 (Top) Second-order Butterworth low-pass filter. (Bottom) Third-order Butterworth low-pass filter.

continuous-time waveform into a digital signal (i.e., the amplitude takes one of 2^n discrete values), which is converted into computer words and stored in computer memory. To adequately capture the continuous-time signal, the sampling instants t_k must be selected carefully so information is not lost. The minimum sampling rate is twice the highest frequency content of the signal (based on the sampling theorem from communication theory). Realistically, we often sample at five to ten times the highest frequency content of the signal so as to achieve better accuracy by reducing aliasing error.

9.14.1 Noise

Measurement signals are always corrupted by noise in a biomedical instrumentation system. Interference noise occurs when unwanted signals are introduced into the system by outside sources, such as power lines and transmitted radio and television electromagnetic waves. This kind of noise is effectively reduced by careful attention to the circuit's wiring configuration to minimize coupling effects.

Interference noise is introduced by power lines (50 or 60 Hz), fluorescent lights, AM/FM radio broadcasts, computer clock oscillators, laboratory equipment, and cellular phones.

Electromagnetic energy radiating from noise sources is injected into the amplifier circuit or into the patient by capacitive and/or inductive coupling. Even the action potentials from nerve conduction in the patient generate noise at the sensor/amplifier interface. Filters are used to reduce the noise and to maximize the signal-to-noise (S/N) ratio at the input of the A/D converter.

Low-frequency noise (amplifier d.c. offsets, sensor drift, temperature fluctuations, etc.) is eliminated by a high-pass filter with the cutoff frequency set above the noise frequencies and below the biological signal frequencies. High-frequency noise (nerve conduction, radio broadcasts, computers, cellular phones, etc.) is reduced by a low-pass filter with the cutoff set below the noise frequencies and above the frequencies of the biological signal that is being monitored. Power line noise is a very difficult problem in biological monitoring, since the 50- or 60-Hz frequency is usually within the frequency range of the biological signal that is being measured. Band-stop filters are commonly used to reduce power line noise. The notch frequency in these band-stop filters is set to the power line frequency of 50 or 60 Hz with the cutoff frequencies located a few Hertz to either side.

The second type of corrupting signal is called inherent noise. Inherent noise arises from random processes that are fundamental to the operation of the circuit's elements and thus is reduced by good circuit design practice. While inherent noise can be reduced, it can never be eliminated. Low-pass filters can be used to reduce high-frequency components. However, noise signals within the frequency range of the biosignal being amplified cannot be eliminated by this filtering approach.

9.14.2 Computers

Computers consist of three basic units: the central processing unit (CPU), the arithmetic and logic unit (ALU), and memory. The CPU directs the functioning of all other units and controls the flow of information among the units during processing procedures. It is controlled by program instructions. The ALU performs all arithmetic calculations (add, subtract, multiply, and divide) as well as logical operations (AND, OR, NOT) that compare one set of information to another.

Computer memory consists of read only memory (ROM) and random access memory (RAM). ROM is permanently programmed into the integrated circuit that forms the basis of the CPU and cannot be changed by the user. RAM stores information temporarily and can be changed by the user. RAM is where user-generated programs, input data, and processed data are stored.

Computers are binary devices that use the presence of an electrical signal to represent 1 and the absence of an electrical pulse to represent 0. The signals are combined in groups of 8 bits, a byte, to code information. A word is made up of 2 bytes. Most desktop computers that are used today are 32-bit systems, which means they can address 4.295×10^9 locations in memory. Most new computers today are 64-bit systems that can address 1.8447×10^{19} locations in memory. The first microcomputers were 8-bit devices that could interact with only 256 memory locations.

Programming languages relate instructions and data to a fixed array of binary bits so the specific arrangement has only one meaning. Letters of the alphabet and other symbols such as punctuation marks are represented by special codes. ASCII stands for American Standard Code for Information Exchange. ASCII provides a common standard that allows

different types of computers to exchange information. When word processing files are saved as text files, they are saved in ASCII format. Ordinarily, word processing files are saved in special program-specific binary formats, but almost all data analysis programs can import and export data in ASCII files.

The lowest level of computer languages is machine language and consists of the 0s and 1s that the computer interprets. Machine language represents the natural language of a particular computer. At the next level, assembly languages use English-like abbreviations for binary equivalents. Programs written in assembly language can manipulate memory locations directly. These programs run very quickly and are often used in data acquisition systems that must rapidly acquire a large number of samples, perhaps from an array of sensors, at a very high sampling rate.

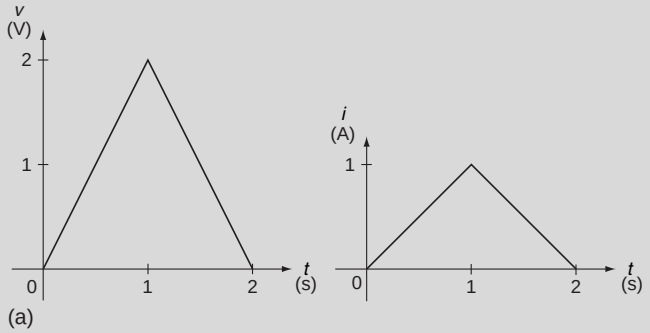
Higher-level languages such as FORTRAN, PERL, and C++ contain statements that accomplish tasks that require many machine or assembly language statements. Instructions in these languages often resemble English and contain commonly used mathematical notations. Higher-level languages are easier to learn than machine and assembly languages. Program instructions are designed to tell computers when and how to use various hardware components to solve specific problems. These instructions must be delivered to the CPU of a computer in the correct sequence in order to give the desired result. Newer programing languages such as MATLAB and LabView are easier to use and more user friendly.

When computers are used to acquire physiological data, programming instructions tell the computer when data acquisition should begin, how often samples should be taken from how many sensors, how long data acquisition should continue, and where the digitized data should be stored. The rate at which a system can acquire samples depends on the speed of the computer's clock—233 MHz—and the number of computer instructions that must be completed in order to take a sample. Some computers can also control the gain on the input amplifiers so signals can be adjusted during data acquisition. In other systems, the gain of the input amplifiers must be manually adjusted.

9.15 EXERCISES

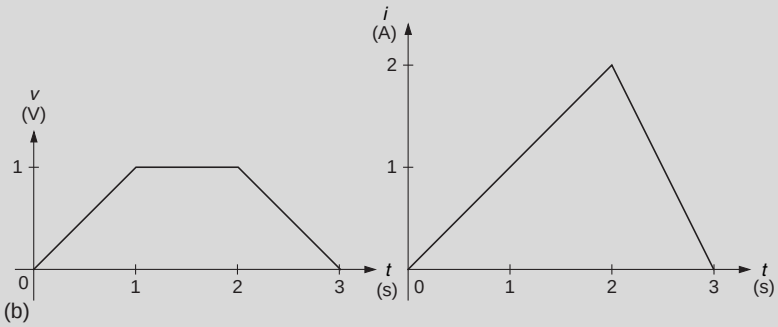
1. Find the power absorbed for the circuit element in Figure 9.7 if

(a)



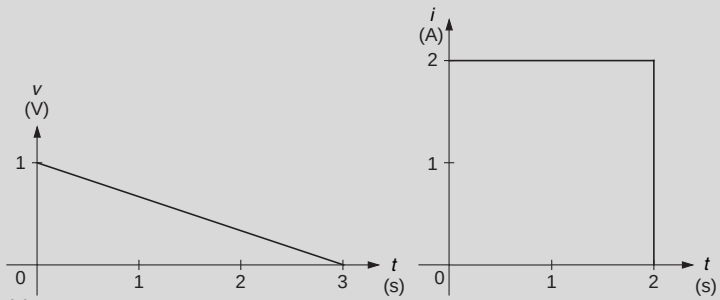
Continued

(b)


v
(V)
1
0
1
2
3
t
(s)
i
(A)
2
1
0
1
2
3
t
(s)

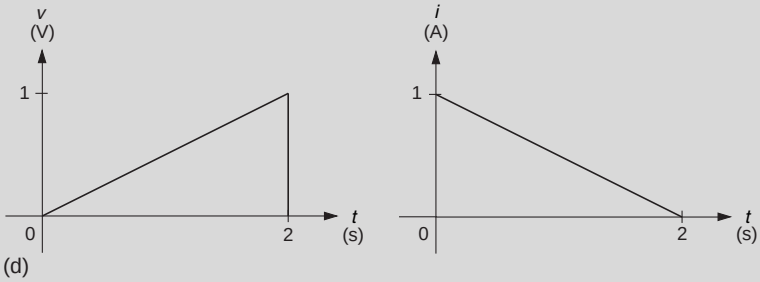
(b)

(c)


v
(V)
1
0
1
2
3
t
(s)
i
(A)
2
1
0
1
2
t
(s)

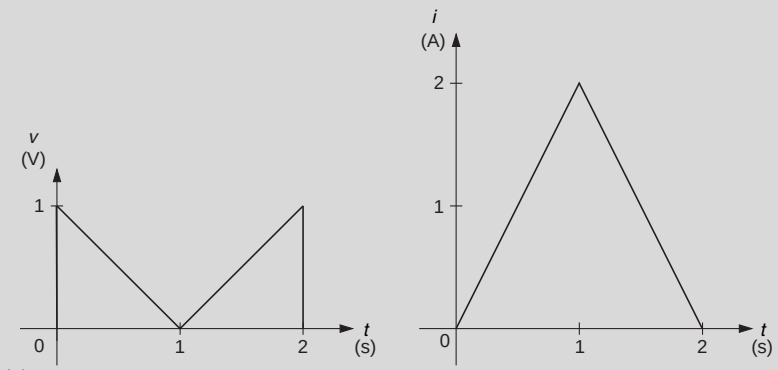
(c)

(d)


v
(V)
1
0
2
t
(s)
i
(A)
1
0
2
t
(s)

(d)

(e)


v
(V)
1
0
1
2
t
(s)
i
(A)
2
1
0
1
2
t
(s)

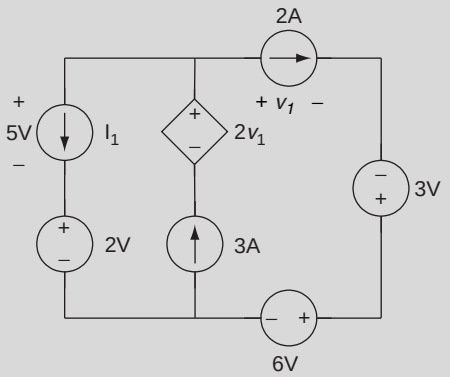
(e)

2. The voltage and current at the terminals in Figure 9.7 are

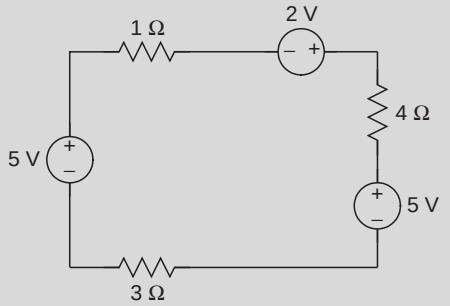
$$v = te^{-10,000t}u(t) \text{ V}$$
$$i = (t + 10)e^{-10,000t}u(t) \text{ A}$$

(a) Find the time when the power is at its maximum.
(b) Find the maximum power.
(c) Find the energy delivered to the circuit at $t = 1 \times 10^{-4}$ s.
(d) Find the total energy delivered to the circuit element.

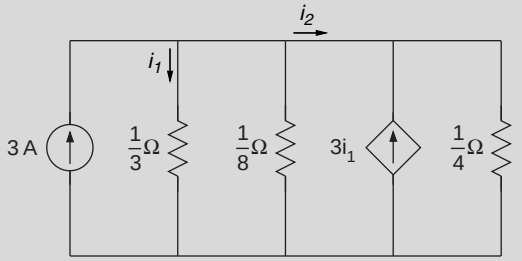
3. For the following circuit find (a) v_1, (b) the power absorbed and delivered.



4. For the following circuit, find the power in each circuit element.

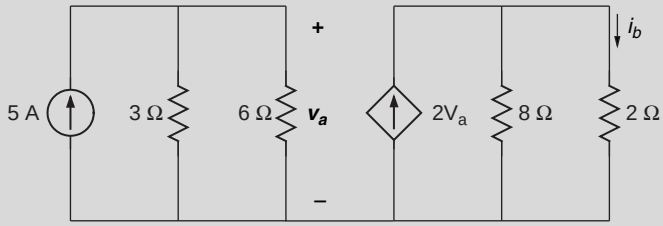


5. Find i_2 in the following circuit.

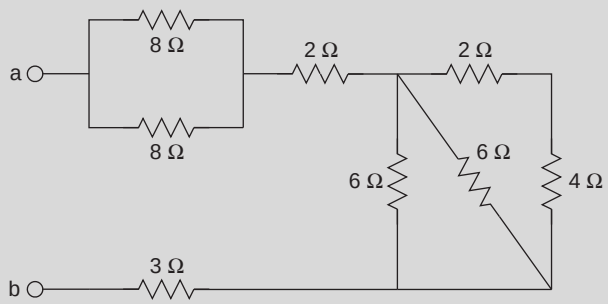


Continued

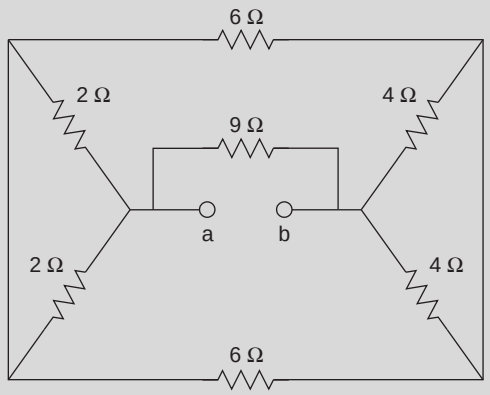
6. Find i_b for the following circuit.



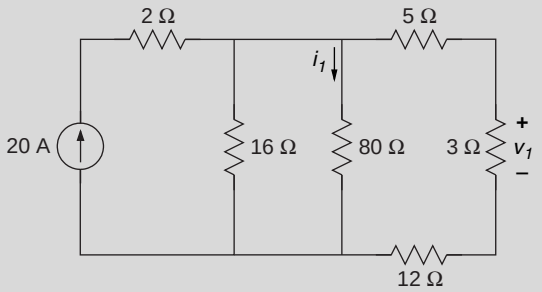
7. Find the equivalent resistance R_{ab} for the following circuit.



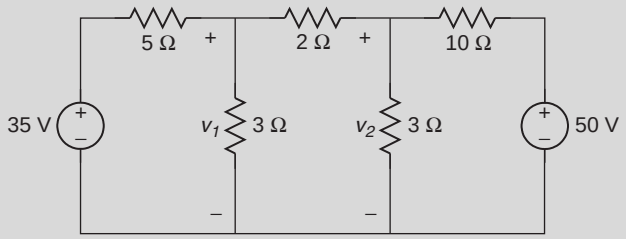
8. Find the equivalent resistance R_{ab} for the following circuit.



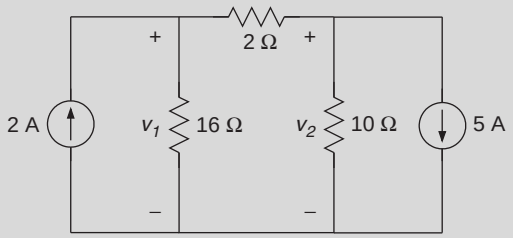
9. Find i_1 and v_1 for the following circuit.



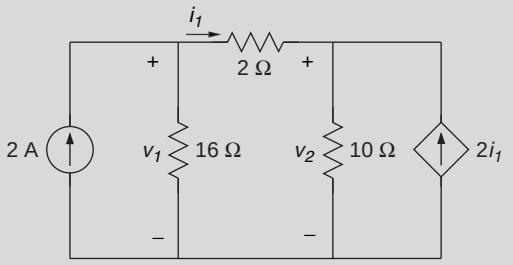
10. Use the node-voltage method to determine v_1 and v_2.



11. Use the node-voltage method to determine v_1 and v_2.

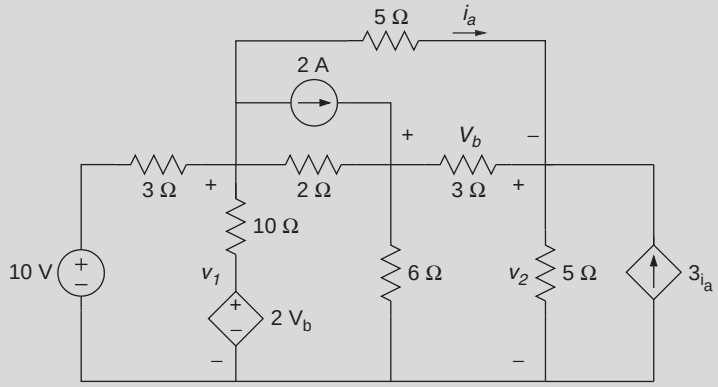


12. Use the node-voltage method to determine v_1 and v_2.

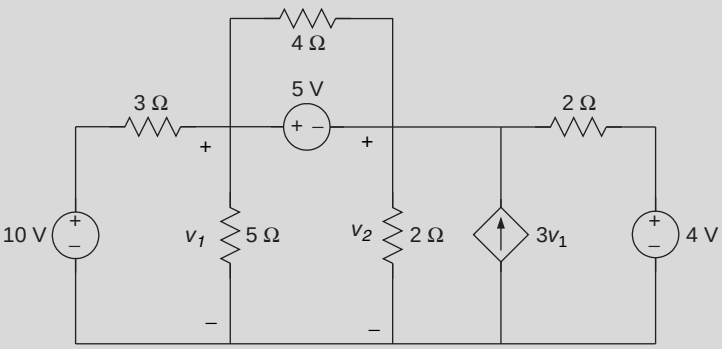


Continued

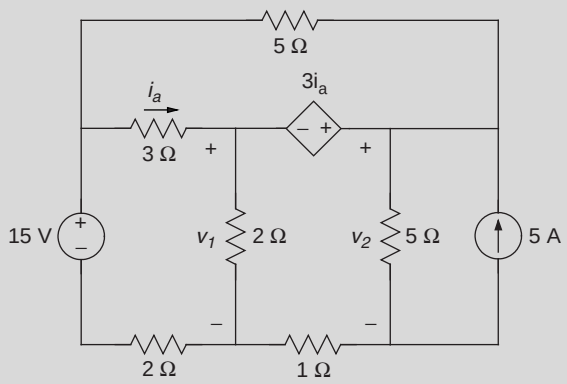
13. Use the node-voltage method to determine v_1 and v_2.



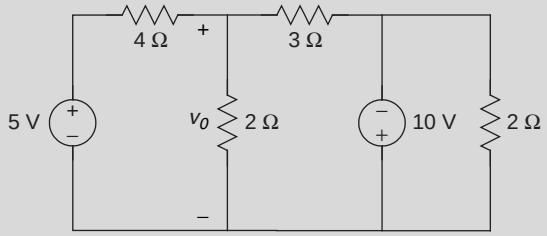
14. Use the node-voltage method to determine v_1 and v_2.



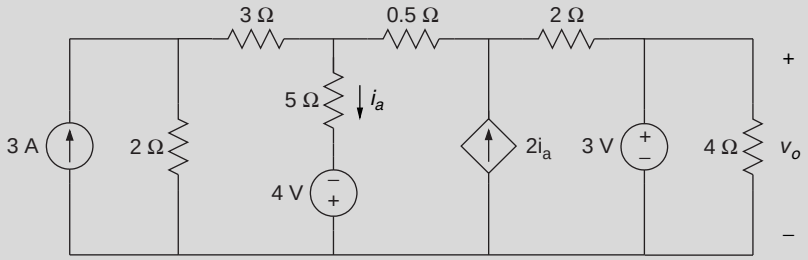
15. Use the node-voltage method to determine v_1 and v_2.



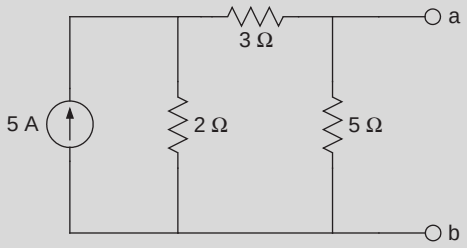
16. Use the superposition method to find v_o.



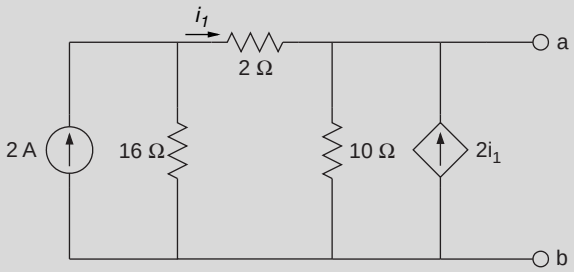
17. Use the superposition method to find v_o.



18. Find the Thévenin equivalent with respect to terminals a and b.



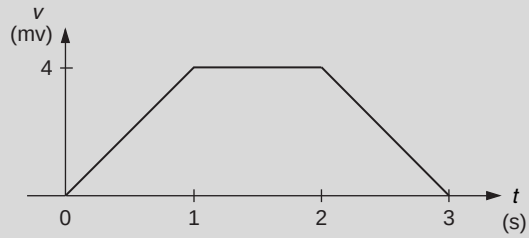
19. Find the Thévenin equivalent with respect to terminals a and b.



Continued

20. A current pulse given by $i(t) = \left(2 + 10e^{-2t}\right)u(t)$ is applied through a 10-mH inductor. (a) Find the voltage across the inductor. (b) Sketch the current and voltage. (c) Find the power as a function of time.

21. The voltage across an inductor is given by the following figure. If $L = 30$ mH and $i(0) = 0$ A, find $i(t)$ for $t \geq 0$.



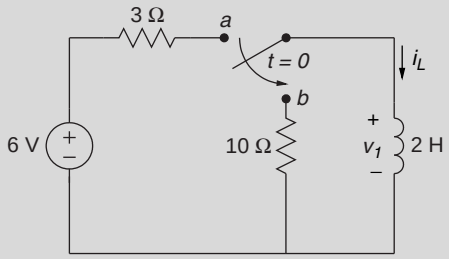
22. The voltage across a 4 μF capacitor is $v(t) = (200{,}000t - 50{,}000)e^{-2000t}u(t)$ V. Find (a) the current through the capacitor, (b) power as a function of time, and (c) energy.

23. The current through a 5 μF capacitor is

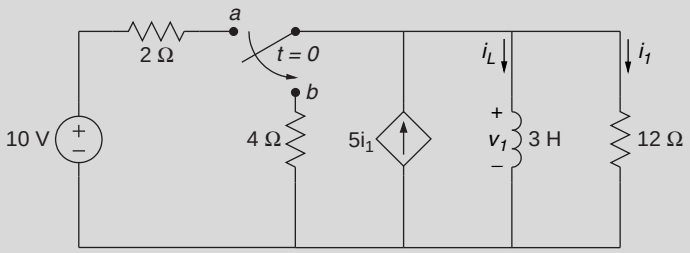
$$i(t) = \begin{cases} 0 \text{ mA} & t < 0 \text{ ms} \\ 5t^2 \text{ mA} & 0 \leq t < 1 \text{ ms} \\ 5(2\text{-}t^2) \text{ mA} & t \geq 1 \text{ ms} \end{cases}$$

Find the voltage across the capacitor.

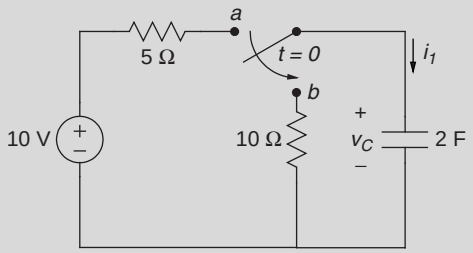
24. The switch has been in position a for a long time. At $t = 0$, the switch instantaneously moves to position b. Find i_L and v_1 for $t > 0$.



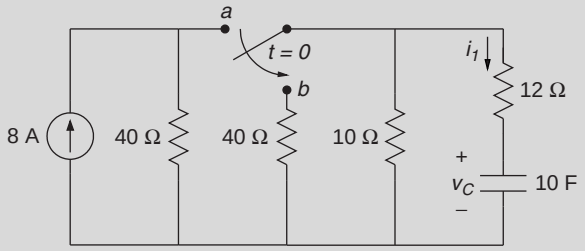
25. The switch has been in position a for a long time. At $t = 0$, the switch instantaneously moves to position b. Find i_L, i_1, and v_1 for $t > 0$.



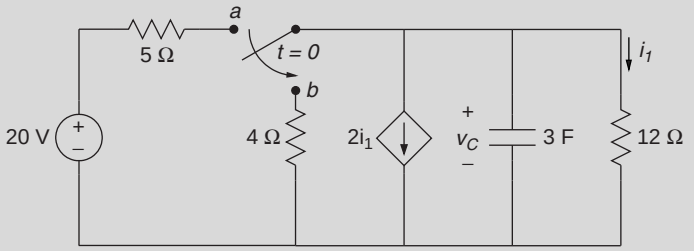
26. The switch has been in position a for a long time. At $t = 0$, the switch instantaneously moves to position b. Find v_c and i_1 for $t > 0$.



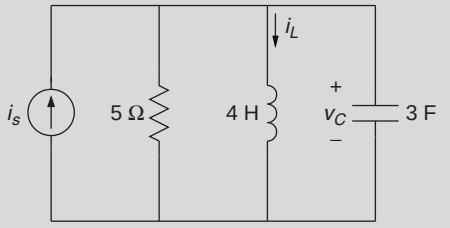
27. The switch has been in position a for a long time. At $t = 0$, the switch instantaneously moves to position b. Find v_c and i_1 for $t > 0$.



28. The switch has been in position a for a long time. At $t = 0$, the switch instantaneously moves to position b. Find v_c and i_1 for $t > 0$.

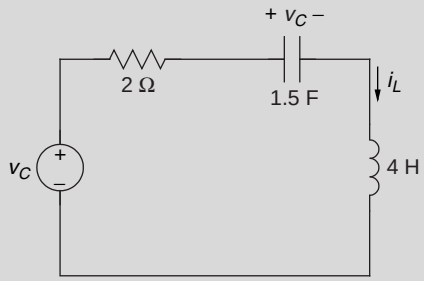


29. Find i_L and v_c for $t > 0$ for the following circuit if (a) $i_s = 3u(t)$ A; (b) $i_s = 1 + 3u(t)$ A.

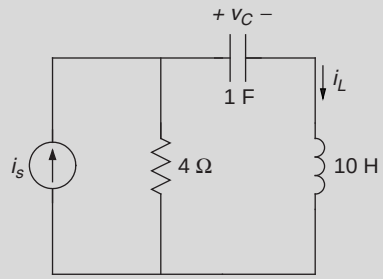


Continued

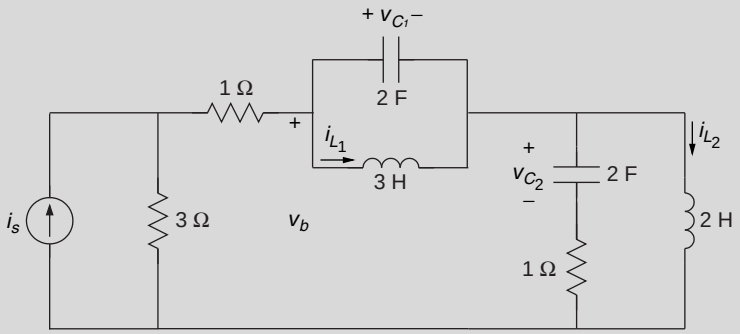
30. Find i_L and v_c for $t > 0$ for the following circuit if (a) $v_s = 5u(t)$ V; (b) $v_s = 5u(t) + 3$ V.



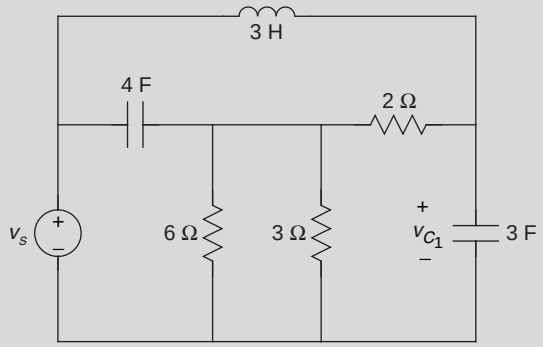
31. Find i_L and v_c for $t > 0$ for the following circuit if (a) $i_s = 3u(t)$ A; (b) $i_s = 3u(t) - 1$ A.



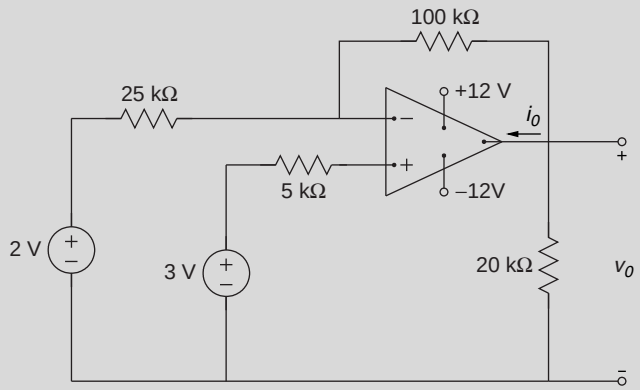
32. For the following circuit we are given that $i_{L_1}(0) = 2$ A, $i_{L_2}(0) = 5$ A, $v_{c_1}(0) = 2$ V, $v_{c_2}(0) = -3$ V, and $i_s = 2e^{-2t}u(t)$ A. Use the node-voltage method to find V_b for $t > 0$.



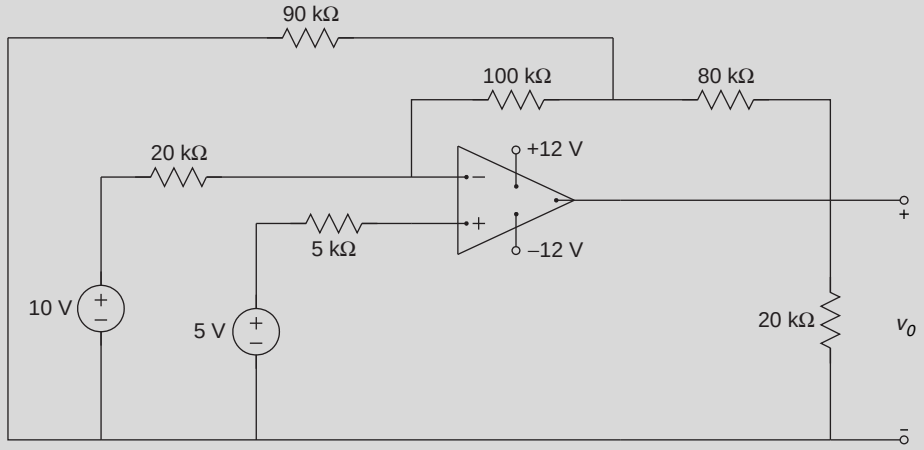
33. Use the node-voltage method to find v_{c_1} for $t > 0$ for the following circuit if (a) $v_s = 2e^{-3t}u(t)$ V; (b) $v_s = 3\cos(2t)u(t)$ V; (c) $v_s = 3u(t) - 1$ V.



34. The operational amplifier shown in the following figure is ideal. Find v_o and i_o.

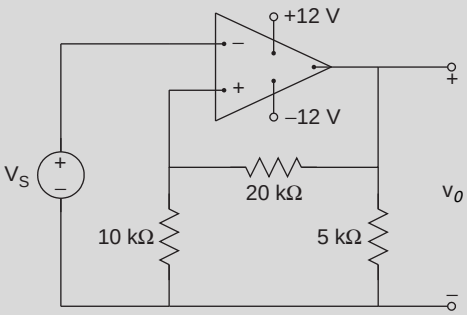


35. The operational amplifier shown in the following figure is ideal. Find v_o.

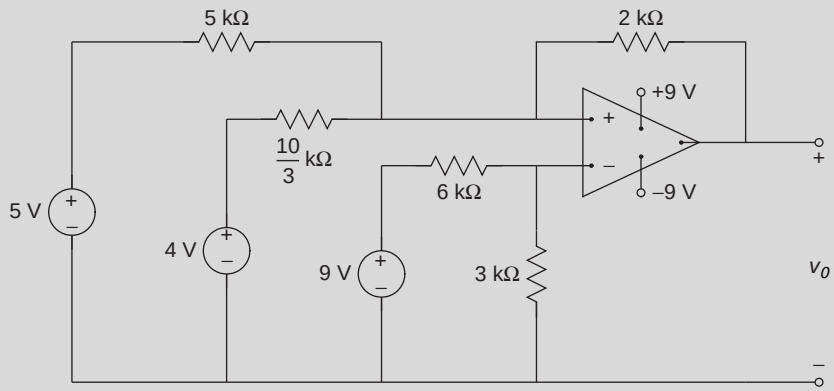


Continued

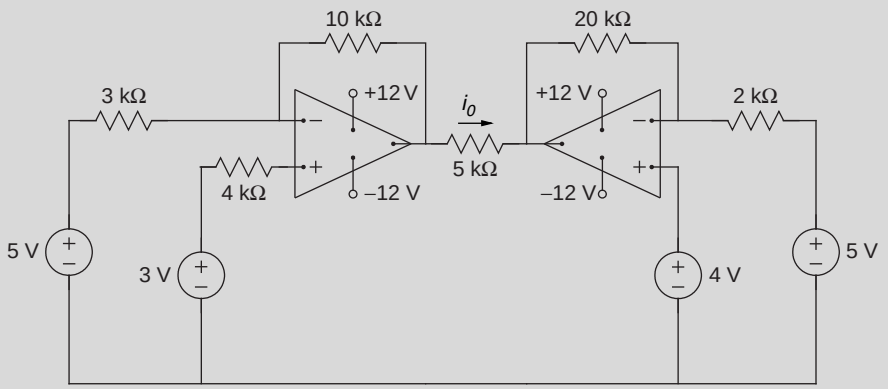
36. Find the overall gain for the following circuit if the operational amplifier is ideal. Draw a graph of v_o versus V_s if V_s varies from 0 to 10 V.



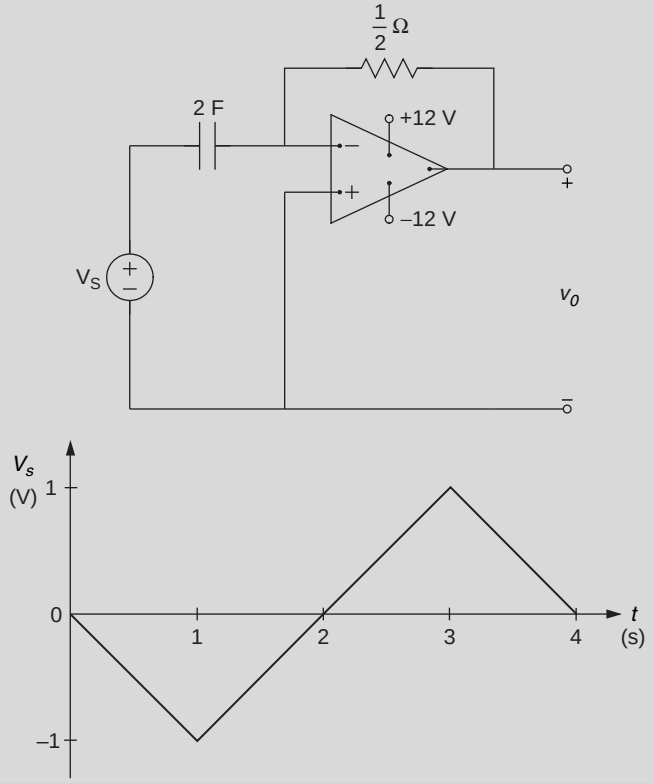
37. Find v_o in the following circuit if the operational amplifier is ideal.



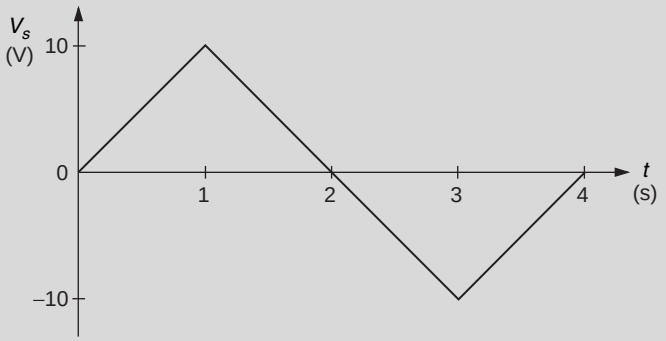
38. Find i_o in the following circuit if the operational amplifiers are ideal.



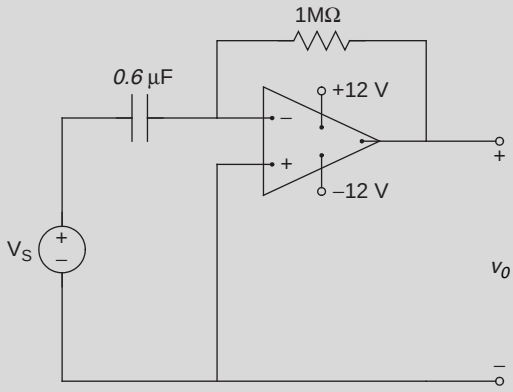
39. Suppose the input V_s is given as a triangular waveform as shown in the following figure. If there is no stored energy in the following circuit with an ideal operational amplifier, find v_o.



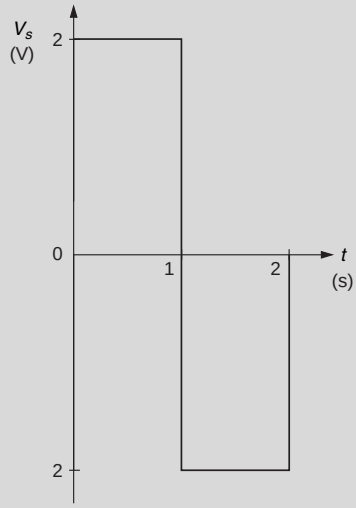
40. Suppose the input V_s is given in the following figure. If there is no stored energy in the following circuit with an ideal operational amplifier, find v_o.

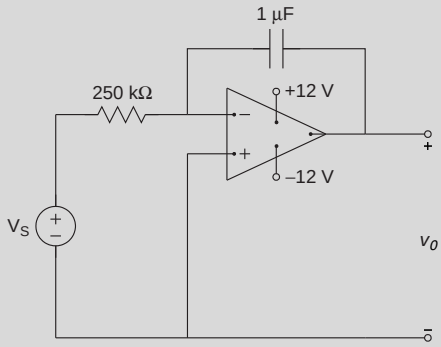


Continued

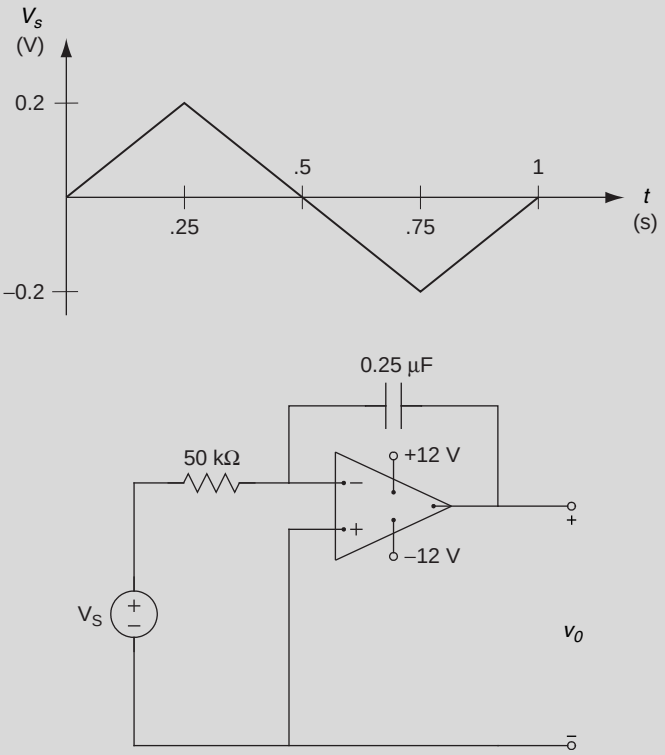


41. Suppose the input V_s is given in the following figure. If there is no stored energy in the following circuit with an ideal operational amplifier, find v_o.

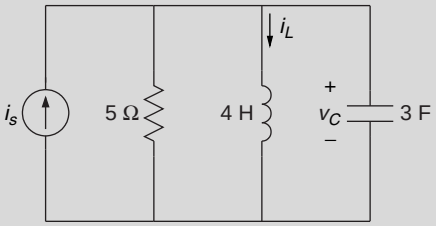




42. Suppose the input V_s is given in the following figure. If there is no stored energy in the following circuit with an ideal operational amplifier, find v_o.

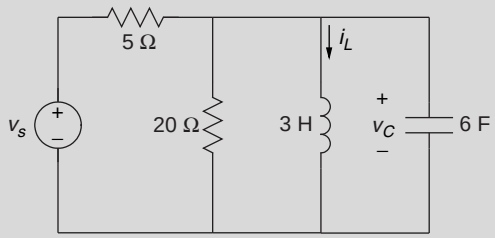


43. The following circuit is operating in the sinusoidal steady state. Find the steady-state expression for i_L if $i_s = 30\cos 20t$ A.

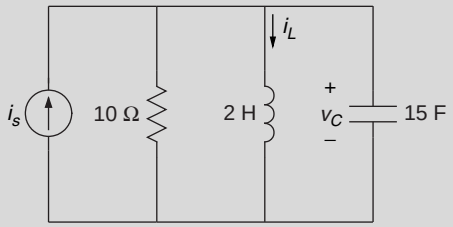


Continued

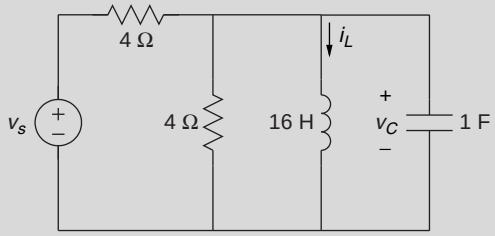
44. The following circuit is operating in the sinusoidal steady state. Find the steady-state expression for v_c if $v_s = 10 \sin 1000t$ V.



45. The following circuit is operating in the sinusoidal steady state. Find the steady-state expression for i_L if $i_s = 5 \cos 500t$ A.



46. The following circuit is operating in the sinusoidal steady state. Find the steady-state expression for v_c if $i_s = 25 \cos 4000t$ V.



47. Design a low-pass filter with a magnitude of 10 and a cutoff frequency of 250 $\frac{\text{rad}}{\text{s}}$.

48. Design a high-pass filter with a magnitude of 20 and a cutoff frequency of 300 $\frac{\text{rad}}{\text{s}}$.

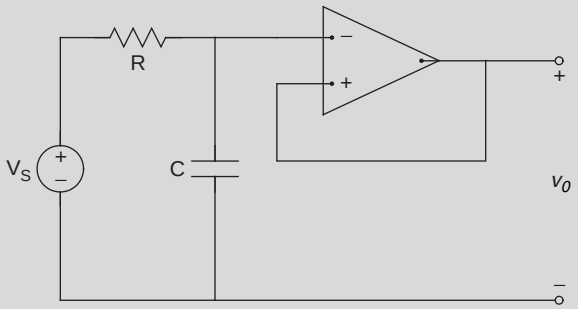
49. Design a band-pass filter with a gain of 15 and passthrough frequencies from 50 to 200 $\frac{\text{rad}}{\text{s}}$.

50. Design a low-pass filter with a magnitude of 5 and a cutoff frequency of 200 $\frac{\text{rad}}{\text{s}}$.

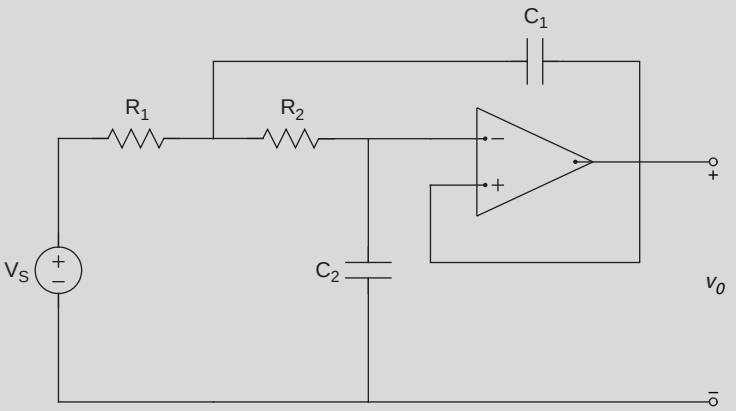
51. Design a high-pass filter with a magnitude of 10 and a cutoff frequency of 500 $\frac{\text{rad}}{\text{s}}$.

52. Design a band-pass filter with a gain of 10 and passthrough frequencies from 20 to 100 $\frac{\text{rad}}{\text{s}}$.

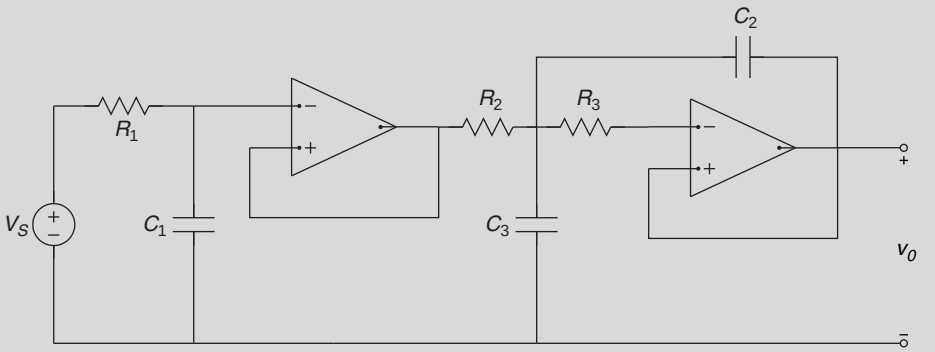
53. Suppose the operational amplifier in the following circuit is ideal. (The circuit is a low-pass first-order Butterworth filter.) Find the magnitude of the output v_o as a function of frequency.



54. With an ideal operational amplifier, the following circuit is a second-order Butterworth low-pass filter. Find the magnitude of the output v_o as a function of frequency.



55. A third-order Butterworth low-pass filter is shown in the following circuit with an ideal operational amplifier. Find the magnitude of the output v_o as a function of frequency.



Suggested Readings

R. Aston, Principles of Biomedical Instrumentation and Measurement, Macmillan Publishing Co., New York, 1990.
J.D. Bronzino, Biomedical Engineering and Instrumentation: Basic Concepts and Applications, PWS Engineering, Boston, 1986.
J.D. Bronzino, V.H. Smith, M.L. Wade, Medical Technology and Society: An Interdisciplinary Perspective, The MIT Press, Cambridge, MA, 1990.
J.J. Carr, J.M. Brown, Introduction to Biomedical Equipment Technology, fourth ed., Prentice Hall, Upper Saddle River, NJ, 2000.
J. Dempster, Computer Analysis of Electrophysiological Signals, Academic Press, London, 1993.
D.A. Johns, K. Martin, Analog Integrated Circuit Design, John Wiley & Sons, New York, 1997.
J.W. Nilsson, S. Riedel, Electric Circuits, seventh ed., Prentice Hall, Upper Saddle River, NJ, 2004.
R.B. Northrop, Noninvasive Instrumentation and Measurement in Medical Diagnosis, CRC Press, Boca Raton, 2001.
R.B. Northrop, Introduction to Instrumentation and Measurements, CRC Press, Boca Raton, 1997.
R. Perez, Design of Medical Electronic Devices, Academic Press, San Diego, 2002.
A. Rosen, H.D. Rosen (Eds.), New Frontiers in Medical Device Technology, John Wiley & Sons, Inc., New York, 1995.
J.G. Webster (Ed.), Bioinstrumentation, John Wiley & Sons, New York, 2003.
J.G. Webster (Ed.), Medical Instrumentation: Application and Design, third ed., John Wiley & Sons, New York, 1997.
W. Welkowitz, S. Deutsch, M. Akay, Biomedical Instruments—Theory and Design, second ed., Academic Press, Inc., San Diego, 1992.
D.E. Wise (Ed.), Bioinstrumentation and Biosensors, Marcel Dekker Inc., New York, 1991.
D.E. Wise (Ed.), Bioinstrumentation: Research, Developments, and Applications, Butterworth Publishers, Stoneham, MA, 1990.

CHAPTER

10

Biomedical Sensors

Yitzhak Mendelson, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Describe the different classifications of biomedical sensors.
- Describe the characteristics that are important for packaging materials associated with biomedical sensors.
- Calculate the half-cell potentials generated by different electrodes immersed in an electrolyte solution.
- Describe the electrodes that are used to record the ECG, EEG, and EMG and those that are used for intracellular recordings.
- Describe how displacement transducers, airflow transducers, and thermistors are used to make physical measurements.
- Describe how blood gases are measured.
- Describe how enzyme-based and microbial biosensors work and some of their uses.
- Explain how optical biosensors work and describe some of their uses.

10.1 INTRODUCTION

Diagnostic bioinstrumentation is used routinely in clinical medicine and biological research for measuring a wide range of physiological variables. Generally, the measurement is derived from sensors or transducers and further processed by the instrument to provide valuable diagnostic information. Biomedical sensors or transducers are the main building blocks of diagnostic medical instrumentation found in many physician offices, clinical laboratories, and hospitals. They are routinely used in vivo to perform continuous invasive and noninvasive monitoring of critical physiological variables, as well as in vitro to help clinicians in various diagnostic procedures. Similar devices are also used in nonmedical applications such as in environmental monitoring, agriculture, bioprocessing, food processing, and the petrochemical and pharmacological industries.

Increasing pressures to lower health care costs, optimize efficiency, and provide better care in less expensive settings without compromising patient care are shaping the future of clinical medicine. As part of this ongoing trend, clinical testing is rapidly being transformed by the introduction of new tests that will revolutionize the way physicians will diagnose and treat diseases in the future. Among these changes, patient self-testing and physician office screening are the two most rapidly expanding areas. This trend is driven by the desire of patients and physicians alike to have the ability to perform some types of instantaneous diagnosis right next to the patient and to move the testing apparatus from an outside central clinical laboratory closer to the point of care.

Generally, medical diagnostic instruments derive their information from sensors, electrodes, or transducers. Medical instrumentation relies on analog electrical signals for an input. These signals can be acquired directly by biopotential electrodes—for example, in monitoring the electrical signals generated by the heart, muscles or brain, or indirectly by transducers that convert a nonelectrical physical variable such as pressure, flow, or temperature, or biochemical variables, such as partial pressures of gases or ionic concentrations, to an electrical signal. Since the process of measuring a biological variable is commonly referred to as sensing, electrodes and transducers are often grouped together and are termed *sensors*.

Biomedical sensors play an important role in a wide range of diagnostic medical applications. Depending on the specific needs, some sensors are used primarily in clinical laboratories to measure in vitro physiological quantities such as electrolytes, enzymes, and other biochemical metabolites in blood. Other biomedical sensors for measuring pressure, flow, and the concentrations of gases, such as oxygen and carbon dioxide, are used in vivo to follow continuously (monitor) the condition of a patient. For real-time continuous in vivo sensing to be worthwhile, the target analytes must vary rapidly and, most often, unpredictably.

10.1.1 Sensor Classifications

Biomedical sensors are usually classified according to the quantity to be measured and are typically categorized as physical, electrical, or chemical, depending on their specific applications. Biosensors, which can be considered a special subclassification of biomedical sensors, are a group of sensors that have two distinct components: a biological recognition element, such as a purified enzyme, antibody, or receptor, that functions as a mediator and provides the selectivity that is needed to sense the chemical component (usually referred to as the

analyte) of interest, and a supporting structure that also acts as a transducer and is in intimate contact with the biological sensing sensed by the biological recognition element into a quantifiable measurement, typically in the element. The purpose of the transducer is to convert the biochemical reaction into the form of an optical, electrical, or physical signal that is proportional to the concentration of a specific chemical. Thus, a blood pH sensor is not considered a biosensor according to this classification, although it measures a biologically important variable. It is simply a chemical sensor that can be used to measure a biological quantity.

10.1.2 Sensor Packaging

Packaging of certain biomedical sensors, primarily sensors for in vivo applications, is an important consideration during the design, fabrication, and use of the device. Obviously, the sensor must be safe and remain functionally reliable. In the development of implantable biosensors, an additional key issue is the long operational lifetime and biocompatibility of the sensor. Whenever a sensor comes into contact with body fluids, the host itself may affect the function of the sensor, or the sensor may affect the site in which it is implanted. For example, protein absorption and cellular deposits can alter the permeability of the sensor packaging that is designed to both protect the sensor and allow free chemical diffusion of certain analytes between the body fluids and the biosensor. Improper packaging of implantable biomedical sensors could lead to drift and a gradual loss of sensor sensitivity and stability over time. Furthermore, inflammation of tissue, infection, or clotting in a vascular site may produce harmful adverse effects. Hence, the materials used in the construction of the sensor's outer body must be biocompatible, since they play a critical role in determining the overall performance and longevity of an implantable sensor. One convenient strategy is to utilize various polymeric covering materials and barrier layers to minimize leaching of potentially toxic sensor components into the body. It is also important to keep in mind that once the sensor is manufactured, common sterilization practices by steam, ethylene oxide, or gamma radiation must not alter the chemical diffusion properties of the sensor packaging material.

10.1.3 Sensor Specifications

The need for accurate medical diagnostic procedures places stringent requirements on the design and use of biomedical sensors. Depending on the intended application, the performance specifications of a biomedical sensor may be evaluated in vitro and in vivo to ensure that the measurement meets the design specifications.

To understand sensor performance characteristics, it is important first to understand some of the common terminology associated with sensor specifications. The following definitions are commonly used to describe sensor characteristics and selecting sensors for particular applications.

Sensitivity

Sensitivity is typically defined as the ratio of output change for a given change in input. Another way to define sensitivity is by finding the slope of the calibration line relating the input to the output (i.e., $\Delta\text{Output}/\Delta\text{Input}$), as illustrated in Figure 10.1. A high sensitivity implies that a small change in input quantity causes a large change in its output.

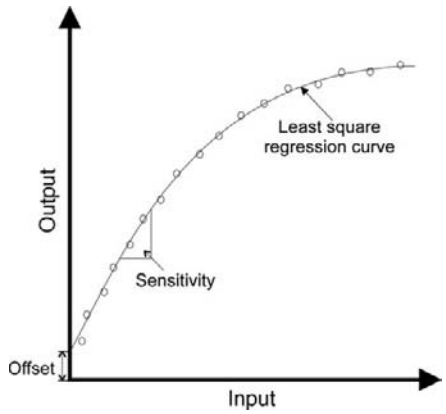


FIGURE 10.1 Input versus output calibration curve of a typical sensor.

For example, a temperature sensor may have a sensitivity of 20 μV/°C; that is, the output of this sensor will change by 20 μV for 1°C change in input temperature. Note that if the calibration line is linear, the sensitivity is constant, whereas the sensitivity will vary with the input when the calibration is nonlinear, as illustrated in Figure 10.1. Alternatively, sensitivity can also be defined as the smallest change in the input quantity that will result in a detectable change in sensor output.

Range

The *range* of a sensor corresponds to the minimum and maximum operating limits that the sensor is expected to measure accurately. For example, a temperature sensor may have a nominal performance over an operating range of −200 to +500°C.

Accuracy

Accuracy refers to the difference between the true value and the actual value measured by the sensor. Typically, accuracy is expressed as a ratio between the preceding difference and the true value and is specified as a percent of full-scale reading. Note that the true value should be traceable to a primary reference standard.[1]

Precision

Precision refers to the degree of measurement reproducibility. Very reproducible readings indicate a high precision. Precision should not be confused with accuracy. For example, measurements may be highly precise but not necessary accurate.

Resolution

When the input quantity is increased from some arbitrary nonzero value, the output of a sensor may not change until a certain input increment is exceeded. Accordingly, *resolution* is defined as the smallest distinguishable input change that can be detected with certainty.

[1]An independently calibrated reference obtained by an absolute measurement of the highest quality that is subsequently used in the calibration of similar measured quantities.

Reproducibility

Reproducibility describes how close the measurements are when the same input is measured repeatedly over time. When the range of measurements is small, the reproducibility is high. For example, a temperature sensor may have a reproducibility of $\pm 0.1^{\circ}$C for a measurement range of 20°C to 80°C. Note that reproducibility can vary depending on the measurement range. In other words, readings may be highly reproducible over one range and less reproducible over a different operating range.

Offset

Offset refers to the output value when the input is zero, as illustrated in Figure 10.1.

Linearity

Linearity is a measure of the maximum deviation of any reading from a straight calibration line. The calibration line is typically defined by the least-square regression fit of the input versus output relationship. Typically, sensor linearity is expressed as either a percent of the actual reading or a percent of the full-scale reading.

The conversion of an unknown quantity to a scaled output reading by a sensor is most convenient if the input-output calibration equation follows a linear relationship. This simplifies the measurement, since we can multiply the measurement of any input value by a constant factor rather than using a "lookup table" to find a different multiplication factor that depends on the input quantity when the calibration equation follows a nonlinear relation. Note that although a linear response is sometimes desired, accurate measurements are possible even if the response is nonlinear as long as the input-output relation is fully characterized.

Response Time

The response time indicates the time it takes a sensor to reach a certain percent (e.g., 95 percent) of its final steady-state value when the input is changed. For example, it may take 20 seconds for a temperature sensor to reach 95 percent of its maximum value when a change in temperature of 1°C is measured. Ideally, a short response time indicates the ability of a sensor to respond quickly to changes in input quantities.

Drift

Drift refers to the change in sensor reading when the input remains constant. Drift can be quantified by running multiple calibration tests over time and determining the corresponding changes in the intercept and slope of the calibration line. Sometimes, the input-output relation may vary over time or may depend on another independent variable that can also change the output reading. This can lead to a *zero* (or *offset*) *drift* or a *sensitivity drift*, as illustrated in Figure 10.2. To determine zero drift, the input is held at zero while the output reading is recorded. For example, the output of a pressure transducer may depend not only on pressure but also on temperature. Therefore, variations in temperature can produce changes in output readings even if the input pressure remains zero. Sensitivity drift may be found by measuring changes in output readings for different nonzero constant inputs. For example, for a pressure transducer, repeating the measurements over a range of temperatures will reveal how much the slope of the input-output calibration line varies with temperature. In practice, both zero

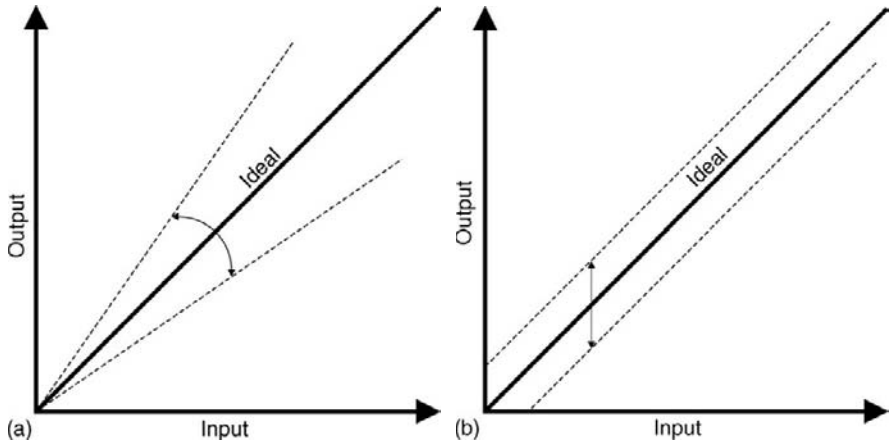


FIGURE 10.2 Changes in input versus output response caused by (a) sensitivity errors and (b) offset errors.

and sensitivity drifts specify the total error due to drift. Knowing the values of these drifts can help to compensate and correct sensor readings.

Hysteresis

In some sensors, the input-output characteristic follows a different nonlinear trend, depending on whether the input quantity increases or decreases, as illustrated in Figure 10.3. For example, a certain pressure gauge may produce a different output voltage when the input pressure varies from zero to full scale and then back to zero. When the measurement is not perfectly reversible, the sensor is said to exhibit *hysteresis*. If a sensor exhibits hysteresis, the input-output relation is not unique but depends on the direction change in the input quantity.

The following sections will examine the operation principles of different types of biomedical sensors, including examples of invasive and noninvasive sensors for measuring biopotentials and other physical and biochemical variables encountered in different clinical and research applications.

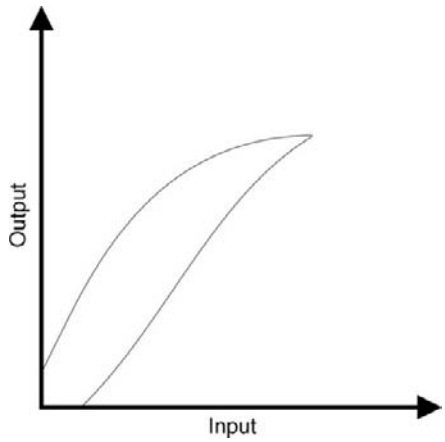


FIGURE 10.3 Input versus output response of a sensor with hysteresis.

EXAMPLE PROBLEM 10.1

A new temperature sensor produced the readings in Table 10.1.

TABLE 10.1 Sample Calibration Data for a Temperature Sensor

Temperature (°C)	Reading (mV)
0	12.3
10	18.2
20	25.4
30	37.0
40	43.6
50	55.8
60	62.0
70	67.8
80	70.4
90	72.1
100	73.0

1. Plot the input-output calibration for this sensor.
2. Find the offset and sensitivity for readings between 0 to 70°C.
3. Estimate the average sensitivity for readings ranging between 70°C to 100°C.

Solution

1.

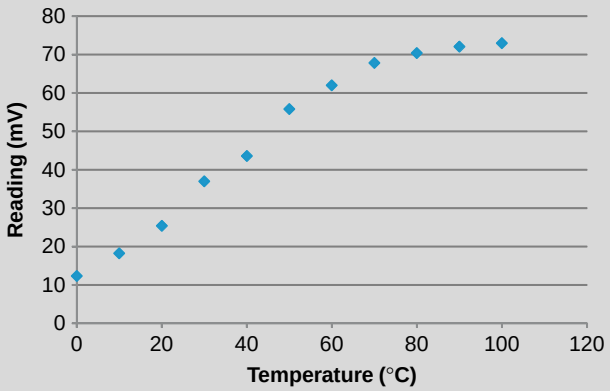


FIGURE 10.4 Input-Output Calibration for a Temperature Sensor.

Continued

2. The equation of the linear regression line describing the input-output calibration data can be written as

$$\text{Reading} = a \cdot (\text{Reference temperature}) + b$$

where, $a =$ is the slope and b is the y-intercept of the regression line. Accordingly, the offset (b) and sensitivity (a) are equal to 10.87 mV and 0.84 mV/°C, respectively.
3. From the equation of the linear regression line, the average sensitivity for readings between 70°C to 100°C is 0.17 mV/°C.

10.2 BIOPOTENTIAL MEASUREMENTS

Biopotential measurements are made using different kinds of specialized electrodes. The function of these recording electrodes is to couple the ionic potentials generated inside the body to an electronic instrument. Biopotential electrodes are classified either as noninvasive (skin surface) or invasive (e.g., microelectrodes or wire electrodes).

Biopotential measurements must be carried out using high-quality electrodes to minimize motion artifacts and ensure that the measured signal is accurate, stable, and undistorted. Body fluids are very corrosive to metals, so not all metals are acceptable for biopotential sensing. Furthermore, some materials are toxic to living tissues. For implantable applications, we typically use relatively strong metal electrodes made, for example, from stainless steel or noble materials such as gold, or from various alloys such as platinum-tungsten, platinum-iridium, titanium-nitride, or iridium-oxide. These electrodes do not react chemically with tissue electrolytes and therefore minimize tissue toxicity. Unfortunately, they give rise to large interface impedances and unstable potentials. External monitoring electrodes can use nonnoble materials such as silver with lesser concerns of biocompatibility, but they must address the large skin interface impedance and the unstable biopotential. Other considerations in the design and selection of biopotential electrodes are cost, shelf life, and mechanical characteristics.

10.2.1 The Electrolyte/Metal Electrode Interface

When a metal is placed in an electrolyte (i.e., an ionizable) solution, a charge distribution is created next to the metal/electrolyte interface, as illustrated in Figure 10.5. This localized charge distribution causes an electric potential, called a half-cell potential, to be developed across the interface between the metal and the electrolyte solution.

The half-cell potentials of several important metals are listed in Table 10.2. Note that the hydrogen electrode is considered to be the standard electrode against which the half-cell potentials of other metal electrodes are measured.

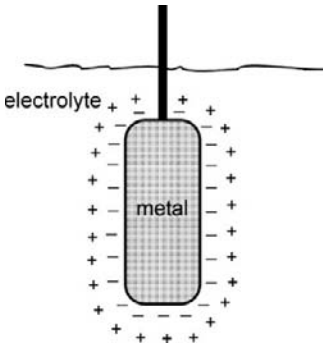


FIGURE 10.5 Distribution of charges at a metal/electrolyte interface.

TABLE 10.2 Half-Cell Potentials of Important Metals

Primary metal and chemical reaction	Half-cell potential (V)
$Al \rightarrow Al^{3+} + 3e^{-}$	−1.706
$Cr \rightarrow Cr^{3+} + 3e^{-}$	−0.744
$Cd \rightarrow Cd^{2+} + 2e^{-}$	−0.401
$Zn \rightarrow Zn^{2+} + 2e^{-}$	−0.763
$Fe \rightarrow Fe^{2+} + 2e^{-}$	−0.409
$Ni \rightarrow Ni^{2+} + 2e^{-}$	−0.230
$Pb \rightarrow Pb^{2+} + 2e^{-}$	−0.126
$H_2 \rightarrow 2H^{+} + 2e^{-}$	0.000 (standard by definition)
$Ag \rightarrow Ag^{+} + e^{-}$	+0.799
$Au \rightarrow Au^{3+} + 3e^{-}$	+1.420
$Cu \rightarrow Cu^{2+} + 2e^{-}$	+0.340
$Ag + Cl^{-} \rightarrow AgCl + 2e^{-}$	+0.223

EXAMPLE PROBLEM 10.2

Silver and zinc electrodes are immersed in an electrolyte solution. Calculate the potential drop between these two electrodes.

Solution

From Table 10.2, the half-cell potentials for the silver and zinc electrodes are 0.799 V and −0.763 V, respectively. Therefore, the potential drop between these two metal electrodes is equal to

$$0.799\text{V} - (-0.763\text{V}) = 1.562\text{V}$$

Typically, biopotential measurements are made by utilizing two similar electrodes composed of the same metal. Therefore, the two half-cell potentials for these electrodes would be equal in magnitude. For example, two similar biopotential electrodes can be taped to the chest near the heart to measure the electrical potentials generated by the heart (electrocardiogram, or ECG). Ideally, assuming that the skin-to-electrode interfaces are electrically identical, the differential amplifier attached to these two electrodes would amplify the biopotential (ECG) signal, but the half-cell potentials would be cancelled out. In practice, however, disparity in electrode material or skin contact resistance could cause a significant DC offset voltage that would cause a current to flow through the two electrodes. This current will produce a voltage drop across the body. The offset voltage will appear superimposed at the output of the amplifier and may cause instability or base line drift in the recorded biopotential.

EXAMPLE PROBLEM 10.3

Silver and aluminum electrodes are placed in an electrolyte solution. Calculate the current that will flow through the electrodes if the equivalent resistance of the solution is equal to 2kΩ.

Solution

$$0.799\ \text{V} - (-1.706\ \text{V}) = 2.505\ \text{V}$$
$$2.505\ \text{V}/2\ \text{k}\Omega = 1.252\ \text{mA}$$

10.2.2 ECG Electrodes

Examples of different types of noninvasive biopotential electrodes used primarily for ECG recording are shown in Figure 10.6. A typical flexible biopotential electrode for ECG recording is composed of certain types of polymers or elastomers that are made electrically conductive by the addition of a fine carbon or metal powder. These electrodes (Figure 10.6a) are available with prepasted AgCl gel for quick and easy application to the skin using a double-sided peel-off adhesive tape.

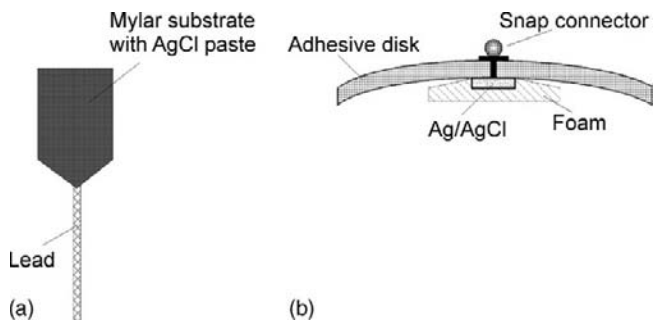


FIGURE 10.6 Biopotential skin surface ECG electrodes: (a) flexible Mylar electrode and (b) disposable snap-type Ag/AgCl electrode.

The most common type of biopotential electrode is the silver/silver chloride electrode (Ag/AgCl), which is formed by electrochemically depositing a very thin layer of silver chloride onto a silver electrode (Figure 10.6b). These electrodes are recessed from the surface of the skin and imbedded in foam that has been soaked with an electrolyte paste to provide good electrical contact with the skin. The electrolyte-saturated foam is also known to reduce motion artifacts that could be produced, for example, during stress testing when the layer of the skin moves relative to the surface of the Ag/AgCl electrode. This motion artifact could cause large interference in the recorded biopotential and, in extreme cases, could severely degrade the measurement.

10.2.3 EMG Electrodes

A number of different types of biopotential electrodes are used in recording electromyographic (EMG) signals from different muscles in the body. The shape and size of the recorded EMG signals depend on the electrical property of these electrodes and the recording location. For noninvasive recordings, proper skin preparation, which normally involves cleansing the skin with alcohol or the application of a small amount of an electrolyte paste, helps to minimize the impedance of the skin-electrode interface and improve the quality of the recorded signal considerably. The most common electrodes used for surface EMG recording and nerve conduction studies are circular discs, about 1 cm in diameter, that are made of silver or platinum.

For direct recording of electrical signals from nerves and muscle fibers, a variety of percutaneous needle electrodes are available, as illustrated in Figure 10.7. The most common type of needle electrode is the concentric bipolar electrode shown in Figure 10.7a. This electrode is made from thin metallic wires encased inside a larger canula or hypodermic needle. The two wires serve as the recording and reference electrodes. Another type of percutaneous EMG electrode is the unipolar needle electrode (Figure 10.7b). This electrode is made of a thin wire that is mostly insulated by a thin layer of Teflon, except about 300 μm near the distal tip. Unlike a bipolar electrode, this electrode requires a second unipolar reference electrode to form a closed electrical circuit. The second recording electrode is normally placed either adjacent to the recording electrode or attached to the surface of the skin.

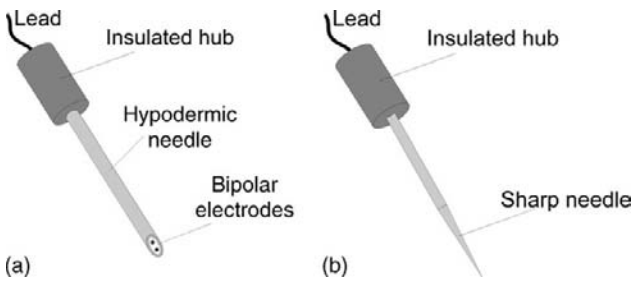


FIGURE 10.7 Intramascular biopotential electrodes: (a) bipolar and (b) unipolar configuration.

10.2.4 EEG Electrodes

The most commonly used electrodes for recording electroencephalographic signals from the brain (EEG) are cup electrodes and subdermal needle electrodes. Cup electrodes are made of platinum or tin approximately 5–10 mm in diameter. These cup electrodes are filled with a conducting electrolyte gel and can be attached to the scalp with an adhesive tape.

Recording of electrical potentials from the scalp is difficult because hair and oily skin impede good electrical contact. Therefore, clinicians sometimes prefer to use subdermal EEG electrodes instead of metal surface electrodes for EEG recording. These are basically fine platinum or stainless-steel needle electrodes about 10 mm long by 0.5 mm wide, which are inserted under the skin to provide a better electrical contact.

10.2.5 Microelectrodes

Microelectrodes are biopotential electrodes with an ultrafine tapered tip that can be inserted into individual biological cells. These electrodes serve an important role in recording action potentials from single cells and are commonly used in neurophysiological studies. The tip of these electrodes must be small with respect to the dimensions of the biological cell to avoid cell damage and at the same time sufficiently strong to penetrate the cell wall. Figure 10.8 illustrates the construction of three typical types of microelectrodes: glass micropipettes, metal microelectrodes, and solid-state microprobes.

In Figure 10.8a, a hollow glass capillary tube, typically 1 mm in diameter, is heated and softened in the middle inside a small furnace and then quickly pulled apart from both ends. This process creates two similar microelectrodes with an open tip that has a diameter on the

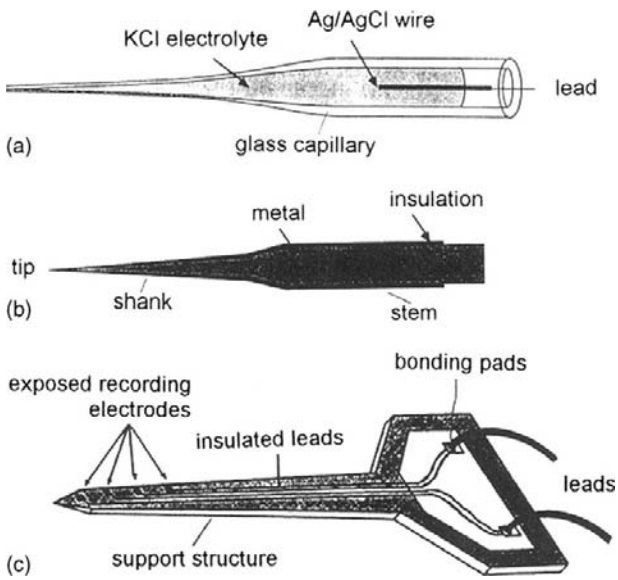


FIGURE 10.8 Biopotential microelectrodes: (a) a capillary glass microelectrode, (b) an insulated metal microelectrode, and (c) a solid-state multisite recording microelectrode.

order of 0.1 to 10 μm. The larger end of the glass tube (the stem) is then filled typically with a 3M KCl electrolyte solution. A short piece of Ag/AgCl wire is inserted through the stem to provide an electrical contact with the electrolyte solution. When the tip of the microelectrode is inserted into an electrolyte solution, such as the intracellular cytoplasm of a biological cell, ionic current can flow through the fluid junction at the tip of the microelectrode. This establishes a closed electrical circuit between the Ag/AgCl wire inside the microelectrode and the biological cell.

A different kind of microelectrode made from a small-diameter strong metal wire (e.g., tungsten or stainless steel) is illustrated in Figure 10.8b. The tip of this microelectrode is usually sharpened down to a diameter of a few micrometers by an electrochemical etching process. The wire is then insulated up to its tip.

Solid-state microfabrication techniques commonly used in the production of integrated circuits can be used to produce microprobes for multichannel recordings of biopotentials or for electrical stimulation of neurons in the brain or spinal cord. An example of such a microsensor is shown in Figure 10.8c. The probe consists of a precisely micromachined silicon substrate with four exposed recording sites. One of the major advantages of this fabrication technique is the ability to mass-produce very small and highly sophisticated microsensors with highly reproducible electrical and physical properties.

10.3 PHYSICAL MEASUREMENTS

10.3.1 Displacement Transducers

Displacement transducers are typically used to measure physical changes in the position of an object or medium. They are commonly employed in detecting changes in length, pressure, or force. Variations in these parameters can be used to quantify and diagnose abnormal physiological functions. In this section, we will describe inductive types of displacement transducers that can be used to measure blood pressure, electromagnetic transducers to measure blood flow, potentiometer transducers to measure linear or angular changes in position, and other types of elastic, strain gauge, capacitive, and piezoelectric type transducers.

Inductive Displacement Transducers

Inductive displacement transducers are based on the inductance L of a coil given by

$$L = \mu \times n^2 \times l \times A \tag{10.1}$$

where μ is the permeability of the magnetically susceptible medium inside the coil (in henry per meter), n is the number of coil turns (in turns per meter), l is the coil length (in meters), and A is the cross-sectional area of the coil (in square meters). These types of transducers measure displacement by changing either the self-inductance of a single coil or the mutual inductance coupling between two or more stationary coils, typically by the displacement of a ferrite or iron core in the bore of the coil assembly. A widely used inductive displacement transducer is the linear variable differential transformer (LVDT) shown in Figure 10.9.

This device is essentially a three-coil mutual inductance transducer that is composed of a primary coil (P) and two secondary coils (S_1 and S_2) connected in series but opposite in

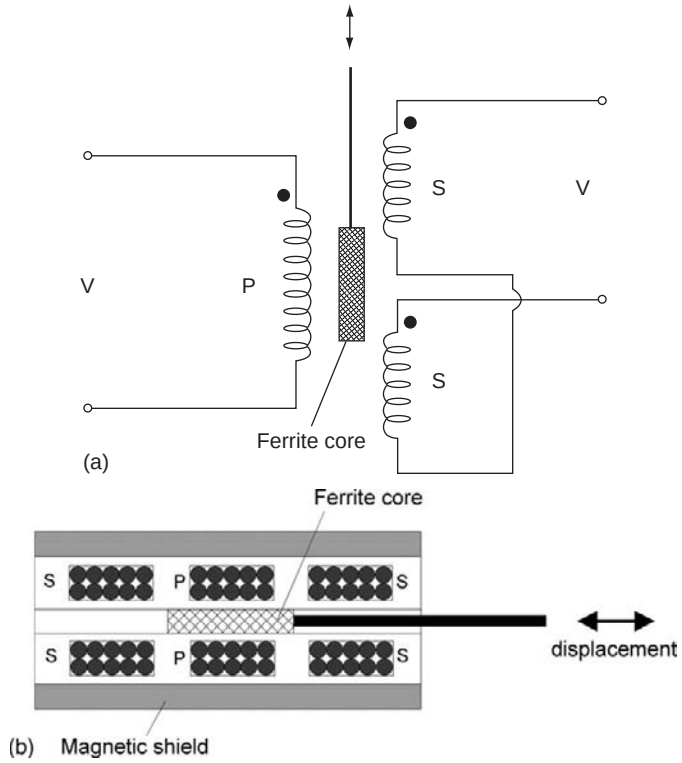


FIGURE 10.9 LVDT transducer: (a) an electric diagram and (b) a cross-section view.

polarity in order to achieve a wider linear output range. The mutual inductance coupled between the coils is changed by the motion of a high-permeability slug. The primary coil is usually excited by passing an AC current. When the slug is centered symmetrically with respect to the two secondary coils, the primary coil induces an alternating magnetic field in the secondary coils. This produces equal voltages (but of opposite polarities) across the two secondary coils. Therefore, the positive voltage excursions from one secondary coil will cancel out the negative voltage excursions from the other secondary coil, resulting in a zero net output voltage. When the core moves toward one coil, the voltage induced in that coil is increased proportionally to the displacement of the core, while the voltage induced in the other coil is decreased proportionally, leading to a typical voltage-displacement diagram, as illustrated in Figure 10.10. Since the voltages induced in the two secondary coils are out of phase, special phase-sensitive electronic circuits must be used to detect both the position and the direction of the core's displacement.

Electromagnetic Flow Transducer

Blood flow through an exposed vessel can be measured by means of an electromagnetic flow transducer. It can be used in research studies to measure blood flow in major blood vessels near the heart, including the aorta at the point where it exits from the heart.

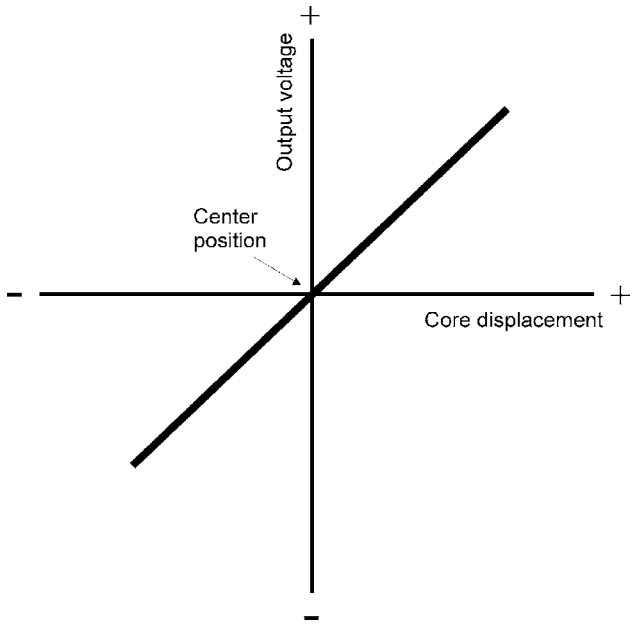


FIGURE 10.10 Output voltage versus core displacement of a typical LVDT transducer.

Consider a blood vessel of diameter, l, filled with blood flowing with a uniform velocity, $\vec{u}$. If the blood vessel is placed in a uniform magnetic flux, $\vec{B}$ (in weber), that is perpendicular to the direction of blood flow, the negatively charged anion and positively charged cation particles in the blood will experience a force, $\vec{F}$ (in newton), which is normal to both the magnetic field and blood flow directions and is given by

$$\vec{F} = q(\vec{u} \times \vec{B}) \tag{10.2}$$

where q is the elementary charge (1.6×10^{-19} C). As a result, these charged particles will be deflected in opposite directions and will move along the diameter of the blood vessels according to the direction of the force vector, $\vec{F}$. This movement will produce an opposing force, $\vec{F}o$, which is equal to

$$\vec{F}o = q \times \vec{E} = q \times \frac{V}{l} \tag{10.3}$$

where $\vec{E}$ is the net electrical field produced by the displacement of the charged particles and V is the potential produced across the blood vessel. At equilibrium, these two forces will be equal. Therefore, the potential difference, V, is given by

$$V = B \times l \times u \tag{10.4}$$

and is proportional to the velocity of blood through the vessel.

EXAMPLE PROBLEM 10.4

Calculate the voltage induced in a magnetic flow probe if the probe is applied across a blood vessel with a diameter of 5×10^{-3} m and the velocity of blood is 5×10^{-2} m/s. Assume that the magnitude of the magnetic field, B, is equal to 1.5×10^{-5} Wb/m^2.

Solution

From Eq. (10.4),

$$V = B \times l \times u = (1.5 \times 10^{-5}\ \text{Wb/m}^2) \times (5 \times 10^{-3}\ \text{m}) \times (5 \times 10^{-2}\ \text{m/s}) = 37.5 \times 10^{-10}\ \text{V}$$

(*Note*: [Wb] = [V × S])

Practically, this device consists of a clip-on probe that fits snugly around the blood vessel, as illustrated in Figure 10.11. The probe contains electrical coils to produce an electromagnetic field that is transverse to the direction of blood flow. The coil is usually excited by an AC current. A pair of very small biopotential electrodes are attached to the housing and rest against the wall of the blood vessel to pick up the induced potential. The flow-induced voltage is an AC voltage at the same frequency as the excitation voltage. Using an AC method instead of DC excitation helps to remove any offset potential error due to the contact between the vessel wall and the biopotential electrodes.

Potentiometer Transducers

A potentiometer is a resistive-type transducer that converts either linear or angular displacement into an output voltage by moving a sliding contact along the surface of a resistive element. Figure 10.12 illustrates linear and angular-type potentiometric transducers. A voltage, V_i, is applied across the resistor, R. The output voltage, Vo, between the sliding contact and one terminal of the resistor is linearly proportional to the displacement. Typically, a constant current source is passed through the variable resistor, and the small change in output voltage is measured by a sensitive voltmeter using Ohm's law (i.e., $I = V/R$).

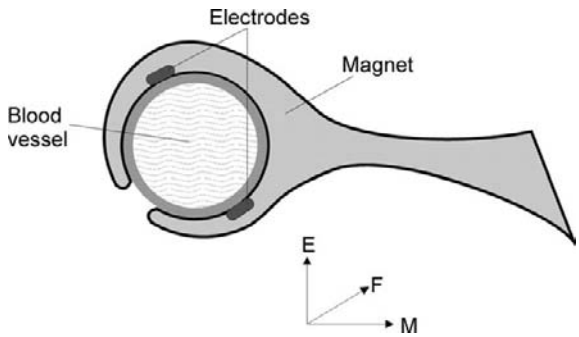


FIGURE 10.11 Electromagnetic blood flow transducer.

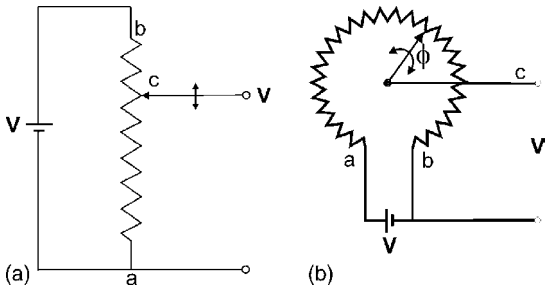


FIGURE 10.12 Linear translational (a) and angular (b) displacement transducers.

EXAMPLE PROBLEM 10.5

Calculate the change in output voltage of a linear potentiometer transducer that undergoes a 20 percent change in displacement.

Solution

Assuming that the current flowing through the transducer is constant, from Ohm's law,

$$\Delta V = I \times \Delta R$$

Hence, since the resistance between the sliding contact and one terminal of the resistor is linearly proportional to the displacement, a 20 percent change in displacement will produce a 20 percent change in the output voltage of the transducer.

Elastic Resistive Transducers

In certain clinical situations, it is desirable to measure changes in the peripheral volume of a leg when the venous outflow of blood from the leg is temporarily occluded by a blood pressure cuff. This volume-measuring method is called plethysmography and can indicate the presence of large venous clots in the legs. The measurement can be performed by wrapping an elastic resistive transducer around the leg and measuring the rate of change in resistance of the transducer as a function of time. This change corresponds to relative changes in the blood volume of the leg. If a clot is present, it will take more time for the blood stored in the leg to flow out through the veins after the temporary occlusion is removed. A similar transducer can be used to follow a patient's breathing pattern by wrapping the elastic band around the chest.

An elastic resistive transducer consists of a thin elastic tube filled with an electrically conductive material, as illustrated in Figure 10.13. The resistance of the conductor inside the flexible tubing is given by

$$R = \rho \frac{I}{A} \tag{10.5}$$

where ρ is the resistivity of the electrically conductive material (in $\Omega \times$ m), l is the length, and A is the cross-sectional area of the conductor.

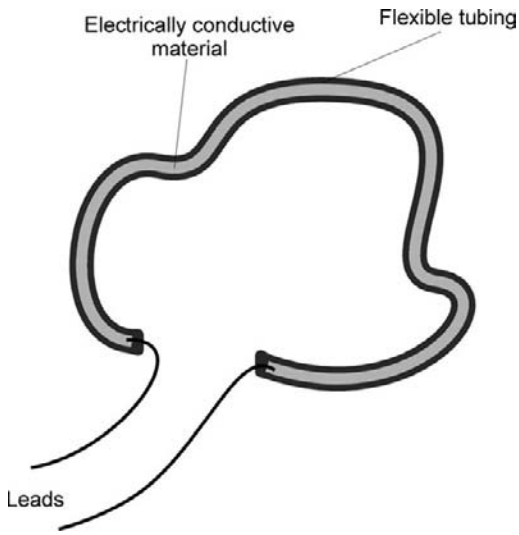


FIGURE 10.13 Elastic resistive transducer.

EXAMPLE PROBLEM 10.6

A 0.1 m long by 0.005 m diameter elastic resistive transducer has a resistance of 1 kΩ. (1) Calculate the resistivity of the electrically conductive material inside the transducer, and (2) calculate the resistance of the transducer after it has been wrapped around a patient's chest having a circumference of 1.2 m. Assume that the cross-sectional area of the transducer remains unchanged.

Solution

1. The cross-sectional area of the transducer (A) is equal to $\pi\ (0.0025)^2\ \text{m}^2 = 1.96 \cdot 10^{-5}\ \text{m}^2$. From Eq. (10.5), $\rho = \dfrac{RA}{l} = \dfrac{1 \cdot 10^3\ \Omega \cdot 1.96 \cdot 10^{-5}\ \text{m}^2}{0.1\ \text{m}} = 0.196\ \Omega \cdot \text{m}$

2.

$$R_{\text{stretched}} = 0.196\ \Omega \cdot \text{m} \cdot \left(\frac{1.2\ \text{m}}{1.96 \cdot 10^{-5}\ \text{m}^2}\right) = 12\ \text{k}\Omega$$

EXAMPLE PROBLEM 10.7

Calculate the change in voltage that is induced across the elastic transducer in Example Problem 10.6 assuming that normal breathing produces a 10 percent change in chest circumference and a constant current of 0.5 mA is passed through the transducer.

Solution

From Ohm's law ($V = I \times R$), $V = 0.5\ \text{mA} \times 12\ \text{k}\Omega = 6\ \text{V}$. If R changes by 10 percent, $\Delta V = 0.6\ \text{V}$.

Strain Gauge Transducers

Strain gauges are displacement-type transducers that measure changes in the length of an object as a result of an applied force. These transducers produce a resistance change that is proportional to the fractional change in the length of the object, also called strain, S, which is defined as

$$\mathrm{S} = \frac{\Delta l}{l} \tag{10.6}$$

where Δl is the fractional change in length, and l is the initial length of the object. Examples include resistive wire elements and certain semiconductor materials.

To understand how a strain gauge works, consider a fine wire conductor of length, l, cross-sectional area, A, and resistivity, ρ. The resistance of the unstretched wire is given by Eq. (10.5). Now suppose that the wire is stretched within its elastic limit by a small amount, Δl, such that its new length becomes $(l + \Delta l)$. Because the volume of the stretched wire must remain constant, the increase in the wire length results in a smaller cross-sectional area, $A_{\text{stretched}}$. Thus,

$$lA = (l + \Delta l) \times A_{\text{stretched}} \tag{10.7}$$

The resistance of the stretched wire is given by

$$R_{\text{stretched}} = \rho \times \frac{l + \Delta l}{A_{\text{stretched}}} \tag{10.8}$$

The increase in the resistance of the stretched wire ΔR is

$$\Delta R = R_{\text{stretched}} - \rho \times \frac{l}{A} \tag{10.9}$$

Substituting Eq. (10.8) and the value for $A_{\text{stretched}}$ from Eq. (10.7) into Eq. (10.9) gives

$$\Delta R = \rho \times \frac{(l + \Delta l)^2}{l \times A} - \rho \times \frac{l}{A} = \frac{\rho \times (l^2 + 2l\Delta l + \Delta l^2 - l^2)}{l \times A} \tag{10.10}$$

Assume that for small changes in length, $\Delta l << l$, this relationship simplifies to

$$\Delta \mathrm{R} = \rho \times \frac{2 \times \Delta l}{A} = \frac{2 \times \Delta l}{l} \times R \tag{10.11}$$

The fractional change in resistance, (ΔR/R), divided by the fractional change in length, ($\Delta l/l$), is called the gauge factor, G. Note that G is a unitless number. Accordingly, the gauge factor provides sensitivity information on the expected change in resistance for a given change in the length of a strain gauge. The gauge factor varies with temperature and the type of material. Therefore, it is important to select a material with a high gauge factor and small temperature coefficient. For a common metal wire strain gauge made of constantan, G is approximately equal to 2. Semiconductor strain gauges made of silicon have a gauge factor about 70 to 100 times higher and are therefore much more sensitive than metallic wire strain gauges.

EXAMPLE PROBLEM 10.8

Calculate the strain in a metal wire gauge for a fractional change in resistance of 10 percent.

Solution

Combine Eqs. (10.6) and (10.11) to obtain

$$\frac{\Delta R}{R} = \frac{2 \times \Delta l}{l} = 2 \times S$$

$$\frac{0.1}{R} = 2 \times S$$

$$S = \frac{0.05}{R}$$

Strain gauges typically fall into two categories: bonded or unbonded. A bonded strain gauge has a folded thin wire cemented to a semiflexible backing material, as illustrated in Figure 10.14.

An unbonded strain gauge consists of multiple resistive wires (typically four) stretched between a fixed and a movable rigid frame. In this configuration, when a deforming force is applied to the structure, two of the wires are stretched, and the other two are shortened proportionally. This configuration is used in blood pressure transducers, as illustrated in Figure 10.15. In this arrangement, a diaphragm is coupled directly by an armature to a movable frame that is inside the transducer. Blood in a peripheral vessel is coupled through a thin fluid-filled (saline) catheter to a disposable dome that is sealed by the flexible diaphragm. Changes in blood pressure during the pumping action of the heart apply a force on the diaphragm that causes the movable frame to move from its resting position. This movement causes the strain gauge wires to stretch or compress and results in a cyclical change in resistance that is proportional to the pulsatile blood pressure measured by the transducer.

In general, the change in resistance of a strain gauge is typically quite small. In addition, changes in temperature can also cause thermal expansion of the wire and thus lead to large changes in the resistance of a strain gauge. Therefore, very sensitive electronic amplifiers

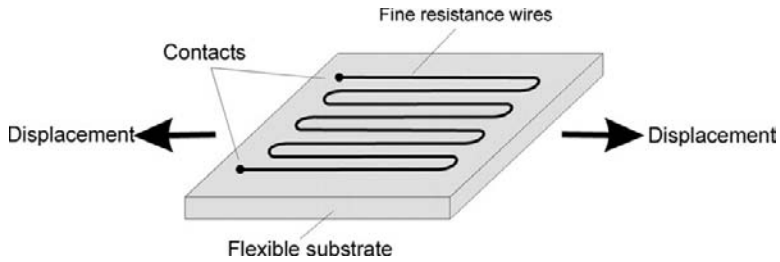


FIGURE 10.14 A bonded-type strain gauge transducer.

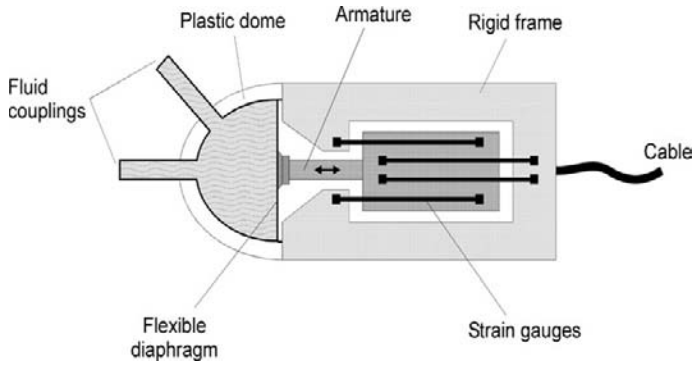


FIGURE 10.15 A resistive strain gauge (unbounded-type) blood pressure transducer.

with special temperature compensation circuits are typically used in applications involving strain gauge transducers.

Capacitive Transducers

The capacitance, C (in farad), between two equal-size parallel plates of cross-sectional area, A, separated by a distance, d, is given by

$$C = \varepsilon_o \times \varepsilon_r \times \frac{A}{d} \tag{10.12}$$

where ε_o is the dielectric constant of free space (8.85×10^{-12} F/m), and ε_r is the relative dielectric constant of the insulating material placed between the two plates. The method that is most commonly employed to measure displacement is to change the separation distance, d, between a fixed and a movable plate, as illustrated in Figure 10.16a. This arrangement can be used to measure force, pressure, or acceleration. Alternatively, it is possible to add a third plate and form a differential-type capacitance transducer (Figure 10.16b). In this configuration, two of the plates are stationary, whereas the middle plate can be moved freely relative to the position of the other plates, thus creating two variable-size capacitors.

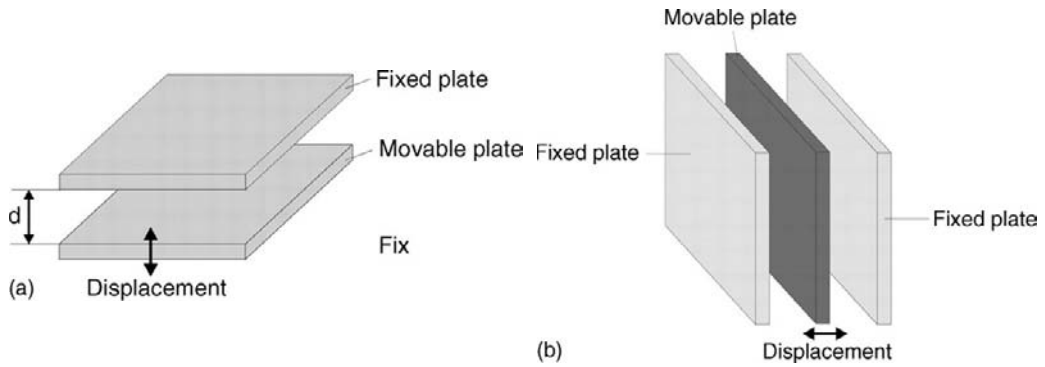


FIGURE 10.16 Capacitive displacement transducer: (a) single capacitance and (b) differential capacitance.

Accordingly, movement of the middle plate, which will change the initial distance, d, by $\pm\Delta d$, will change the distance between two adjacent plates such that one capacitor will increase while the other will decrease in value. This double-capacitor arrangement provides improved sensitivity and can be incorporated into a Wheatstone bridge configuration. Capacitance sensors can be mass-produced using solid-state microfabrication techniques that are commonly employed in making integrated circuits.

EXAMPLE PROBLEM 10.9

Two metal plates with an area of 0.4×10^{-3} m^2 and separation distance of 1×10^{-4} m are used to form a capacitance transducer. If the material between the two plates has a dielectric constant $\varepsilon_r = 2.5$, calculate the capacitance of the transducer.

Solution

$$C = \varepsilon_o \times \varepsilon_r \frac{A}{d} = 8.85 \times 10^{-12}\ \text{F/m} \times 2.5 \times 4 \times 10^{-4}\ \text{m}^2/(1 \times 10^{-4}\ \text{m}) = 0.885\ \text{F}$$

Capacitive displacement transducers can be used to measure respiration or movement by attaching multiple transducers to a mat that is placed on a bed. A capacitive displacement transducer can also be used as a pressure transducer by attaching the movable plate to a thin diaphragm that is in contact with a fluid or air. By applying a voltage across the capacitor and amplifying the small AC signal generated by the movement of the diaphragm, it is possible to obtain a signal that is proportional to the applied external pressure source.

Piezoelectric Transducers

Piezoelectric transducers are used in cardiology to listen to heart sounds (phonocardiography), in automated blood pressure measurements, and for measurement of physiological forces and accelerations. They are also commonly employed in generating ultrasonic waves (high-frequency sound waves typically above 20 kHz) that are used for measuring blood flow or imaging internal soft structures in the body.

A piezoelectric transducer consists of a small crystal (e.g., quartz) that contracts if an electric field (usually in the form of a short voltage impulse) is applied across its plates, as illustrated in Figure 10.17. Conversely, if the crystal is mechanically strained, it will generate a small electric potential. Besides quartz, several other ceramic materials, such as barium titanate and lead zirconate titanate, are also known to produce a piezoelectric effect.

The piezoelectric principle is based on the phenomenon that when an asymmetrical crystal lattice is distorted by an applied force, F, the internal negative and positive charges are reoriented. This causes an induced surface charge, Q, on the opposite sides of the crystal. The induced charge is directly proportional to the applied force and is given by

$$Q = k \times F \tag{10.13}$$

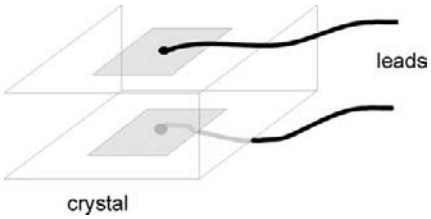


FIGURE 10.17 Ultrasonic transducer.

where k is a proportionality constant for the specific piezoelectric material. By assuming that the piezoelectric crystal acts like a parallel plate capacitor, the voltage across the crystal, V, is given by

$$\Delta V = \frac{\Delta Q}{C} \tag{10.14}$$

where C is the equivalent capacitance of the crystal.

EXAMPLE PROBLEM 10.10

Derive a relationship for calculating the output voltage across a piezoelectric transducer that has a thickness, d, and area, A, in terms of an applied force, F.

Solution

The capacitance of a piezoelectric transducer can be approximated by Eq. (10.12). Equation (10.14) is combined with the relationship given by Eq. (10.13) to give

$$\Delta V = \frac{\Delta Q}{C} = \frac{k \times F}{C} = \frac{k \times F \times d}{\varepsilon_0 \times \varepsilon_r \times A}$$

Since the crystal has an internal leakage resistance, any steady charge produced across its surfaces will eventually be dissipated. Consequently, these piezoelectric transducers are not suitable for measuring a steady or low-frequency DC force. Instead, they are used either as variable force transducers or as mechanically resonating devices to generate high frequencies (typically from 1 to 10 MHz) either in crystal-controlled oscillators or as ultrasonic pulse transducers.

Piezoelectric transducers are commonly used in biomedical applications to measure the thickness of an object or in noninvasive blood pressure monitors. For instance, if two similar crystals are placed across an object (e.g., a blood vessel), one crystal can be excited to produce a short burst of ultrasound. The time it takes for this sound to reach the other transducer can be measured. Assuming that the velocity of sound propagation in soft tissue, c_t, is known (typically 1500 m/s), the time, t, it takes the ultrasonic pulse to propagate across the object can be measured and used to calculate the separation distance, d, of the two transducers from the following relationship:

$$d = c_t \times t \tag{10.15}$$

Microelectromechanical System Transducers

Microelectromechanical system (MEMS) transducers are fabricated using solid-state micromachining techniques commonly used by the semiconductor industry in the production of integrated circuits. A pressure sensor based on MEMS technology is based on the deflection of a micromachined silicon diaphragm mounted on a piezoresistive transducer that changes its output voltage with corresponding variations in the applied pressure. Common commercial applications of MEMS sensors include automobile airbag restraints and fuel injection systems.

To date, the biggest success in medical MEMS technology is the development as a disposable transducer for use in invasive blood pressure monitoring (Figure 10.18). The MEMS transducer measures blood pressure through a silicon-based dielectric gel applied between the sensor element and the saline solution filling the intravascular catheter. The function of the gel is to protect the sensitive electrical circuitry by providing electrical isolation between the MEMS sensor and the saline solution.

MEMS-based medical sensors are used, for example, in intrauterine pressure measurement to monitor contractions during delivery, automatic noninvasive blood pressure cuffs, respiratory monitors, infusion pumps, and kidney dialysis machines. MEMS transducers are also used as accelerometers in implantable pacemakers to monitor body motion to determine patient exertion level and suitably adjust the pacing rate to match changes in metabolic demand.

10.3.2 Airflow Transducers

One of the most common airflow transducers is the Fleish pneumotachometer, shown in Figure 10.19. The device consists of a straight short-tube section with a fixed screen

FIGURE 10.18 An MEMS pressure transducer. *Courtesy of Measurement Specialties, Inc.*

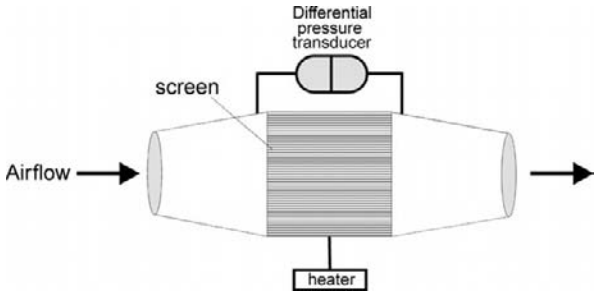


FIGURE 10.19 Fleish airflow transducer.

obstruction in the middle that produces a slight pressure drop as the air is passed through the tube. The pressure drop created across the screen is measured by a differential pressure transducer. The signal produced by the pressure transducer is proportional to the air velocity. The tube is normally shaped in a cone to generate a laminar airflow pattern. A small heater heats the screen so water vapor does not condense on it over time and produce an artificially high pressure drop. Fleish-type pneumotachometers are used to monitor volume, flow, and breathing rates of patients on mechanical ventilators.

10.3.3 Temperature Measurement

Body temperature is one of the most tightly controlled physiological variables and one of the four basic vital signs used in the daily assessment of a patient's health. The interior (core) temperature in the body is remarkably constant—about 37°C for a healthy person—and is normally maintained within ±0.5°C. Therefore, elevated body temperature is a sign of disease or infection, whereas a significant drop in skin temperature may be a clinical indication of shock.

There are two distinct areas in the body where temperature is measured routinely: the surface of the skin under the armpit or inside a body cavity such as the mouth or the rectum. Several sensors exist to measure body temperature.

Thermistors

Thermistors are temperature-sensitive transducers made of compressed sintered metal oxides (such as nickel, manganese, or cobalt) that change their resistance with temperature. Commercially available thermistors range in shape from small beads to large disks, as shown in Figure 10.20.

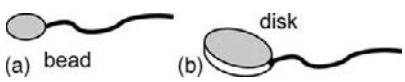


FIGURE 10.20 Common forms of thermistors.

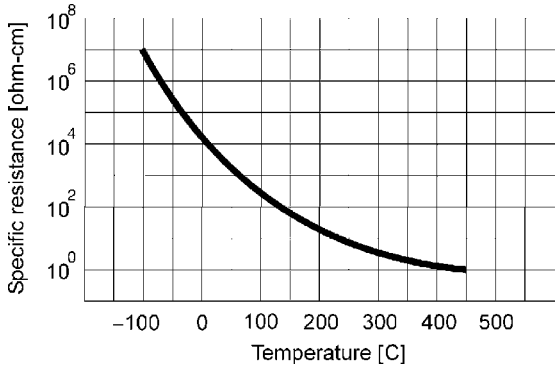


FIGURE 10.21 Resistivity versus temperature characteristics of a typical thermistor.

Mathematically, the resistance-temperature characteristic of a thermistor can be approximated by

$$R_T = R_0 \times \exp\left[\beta \times \left(\frac{1}{T} - \frac{1}{T_0}\right)\right] \tag{10.16}$$

where R_0 is the resistance at a reference temperature, T_0 (in degrees K), R_T is the resistance at temperature, T (in degrees K), and β is a material constant, typically between 2,500 and 5,500 K. A typical resistance-temperature characteristic of a thermistor is shown in Figure 10.21. Note that unlike metals and conventional resistors that have a positive temperature coefficient (as the temperature increases, the resistance increases), thermistors have a nonlinear relationship between temperature and resistance and a negative temperature coefficient. Increasing the temperature decreases the resistance of the thermistor.

EXAMPLE PROBLEM 10.11

A thermistor with a material constant β of 4,500 K is used as a thermometer. Calculate the resistance of this thermistor at 25°C. Assume that the resistance of this thermistor at body temperature (37°C) is equal to 85 Ω.

Solution

Using the resistance-temperature characteristic of a thermistor (Eq. (10.16)) gives

$$R_T = 85 \times \exp\left[4500 \times \left(\frac{1}{298} - \frac{1}{310}\right)\right] = 152.5\ \Omega$$

The size and mass of a thermistor probe in a medical thermometer must be small in order to produce a rapid response time to temperature variations. The probe is normally covered with a very thin sterile plastic sheet that is also disposable to prevent cross-contamination between patients.

A thermistor sensor can be employed in a Swan-Ganz thermodilution technique for measuring cardiac output (the volume of blood ejected by the heart each minute) and assessing ventricular function. The procedure is normally performed in the operating room or the intensive care unit. It involves a rapid bolus injection of a cold indicator solution, usually 3–5 ml of a sterile saline or dextrose solution kept at 0°C, into the right atrium via a flexible pulmonary artery catheter (Figure 10.22).

The 5 or 7 French-size thermodilution catheter contains a small balloon and is normally inserted into either the femoral or internal jugular veins. The catheter is constructed of a radiopaque material to enable easy visualization by an x-ray machine. It contains three ports: a balloon inflation port to guide the flexible tip to the right location, a proximal central venous port, and a distal pulmonary artery port. After the balloon is inflated, the tip of the flexible catheter is passed across the tricuspid valve through the right ventricle, across the pulmonary valve, and into the pulmonary artery. The proximal and distal ports can

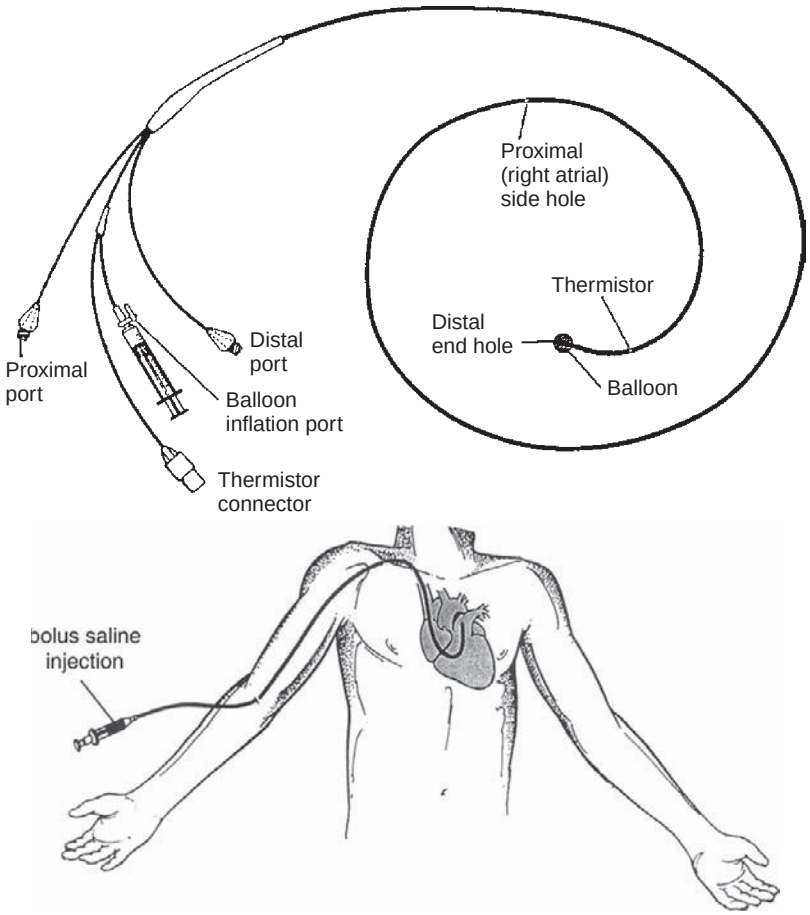


FIGURE 10.22 A Swan-Ganz thermodilution catheter.

be connected to pressure transducers that measure blood pressure inside the right side of the heart while the catheter is advanced into the right atrium.

After the catheter is inserted, a cold bolus is injected into the right atrium through the proximal lumen of the catheter. The bolus solution mixes with the venous blood in the right atrium and causes the blood to cool slightly. The cooled blood is ejected by the right ventricle into the pulmonary artery, where it contacts a thermistor that is located in the wall of the catheter near its distal tip. The thermistor measures the change in blood temperature as the blood passes on to the lungs. An instrument computes the cardiac output by integrating the change in blood temperature immediately following the bolus injection, which is inversely proportional to cardiac output.

Thermocouples

Thermocouples are temperature transducers formed by joining together two dissimilar metals, based on the discovery by Seebeck in 1821. When the two junctions of these dissimilar materials are maintained at different temperatures, an electromotive force (EMF) is generated. The magnitude of the EMF is dependent on the temperature at the junctions and the properties of the materials. This means that the thermocouple is only capable of recognizing a temperature difference between two points and it cannot measure absolute temperature directly. To determine the absolute temperature of the measured environment, the temperature of the reference junction must be determined independently and factored into the absolute temperature calculation. In principle, one junction can be maintained at a constant reference temperature (for example, 0°C if the reference junction is placed in an insulated ice water bath) while the other junction is used to measure the unknown temperature of the environment. However, the use of an ice water bath is not very practical and complicates the measurement considerably. To simplify this process, special integrated circuits have been developed and are available commercially to perform "cold junction compensation." If a closed circuit is formed, as shown in Figure 10.23, the current flowing in the

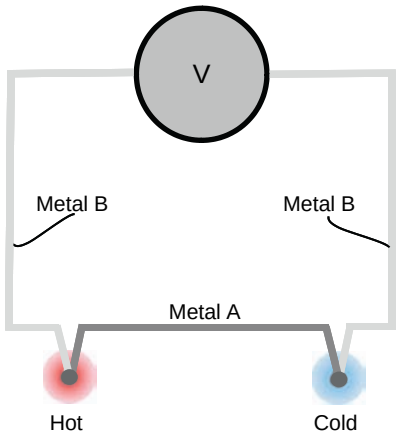


FIGURE 10.23 Principle of a thermocouple-type temperature transducer.

circuit is proportional to the temperature difference between the two junctions over a reasonable wide range of temperatures.

The relationship between the EMF across a junction of two dissimilar metals, E, and the temperature of the measurement junction, T, assuming the cold reference junction is maintained at 0°C, can be approximated using the following truncated power series expansion:

$$E = c_o + c_1 T + c_2 T^2 + \cdots \tag{10.17}$$

where c_i are empirically derived calibration coefficients, T is given in degrees Centigrade, and E is in mV. The Seebeck coefficient α, which describes the temperature sensitivity of the thermocouple, can be derived by differentiating Eq. (10.17) with respect to T:

$$\alpha = \frac{dE}{dT} = c_1 + 2c_2 T + \cdots \tag{10.18}$$

Note that α is a function of temperature. The properties of commonly used thermocouple materials are given in Table 10.3.

The small size, fast response, and rugged design of thermocouple probes make them very attractive for in vivo applications. They can be inserted into the body through a hypodermic needle or a catheter. Examples of medical applications of thermocouples include medical equipment, deep-tissue hyperthermia, and cryogenic therapy.

EXAMPLE PROBLEM 10.12

A Chromel/Alumel thermocouple has the following empirical coefficients:

$$C_0 = -1.76004 \times 10^{-2}$$
$$C_1 = 3.89212 \times 10^{-2}$$
$$C_2 = 1.85587 \times 10^{-5}$$

Find the EMF generated by this thermocouple at a temperature of 500°C.

Solution

Substituting the calibration coefficients just given into Eq. (10.17) yields E ≅ 24.1 mV.

TABLE 10.3 Properties of Selected Thermocouple Materials

Thermocouple	Sensitivity μV/°C (@25°C)	Operating Range (°C)
Chromel/Alumel	40.6	−270 to 1,300
Copper/Constantan	40.9	−270 to 600
Iron/Constantan	51.7	−270 to 1,000
Chromel/Constantan	60.9	−200 to 1,000

EXAMPLE PROBLEM 10.13

Find the Seebeck coefficient for the Chromel/Alumel thermocouple at a temperature of 500°C.

Solution

Using Eq. (10.18) with the coefficients given in Example Problem 10.12 yields $\alpha \cong 57\ \mu V/°C$.

Tympanic Thermometer

Noncontact thermometers measure the temperature of the ear canal near the tympanic membrane, which is known to track the core temperature by about 0.5–1.0°C. Basically, as shown in Figure 10.24, infrared radiation from the tympanic membrane is channeled to a heat-sensitive detector through a metal waveguide that has a gold-plated inner surface for better reflectivity. The detector, which is either a thermopile or a pyroelectric sensor that converts heat flow into an electric current, is normally maintained at a constant temperature environment to minimize inaccuracies due to fluctuation in ambient temperature. A disposable speculum is used on the probe to protect patients from cross-contamination.

Temporal Artery Thermometer

A noninvasive scanning thermometer was developed by Exergen Corporation as an alternative to the tympanic thermometer for measuring core body temperature, essentially eliminating the discomfort caused by a mouth, ear, or rectal thermometer (Figure 10.25). The measurement is based on scanning the area above the temporal artery using an IR detector similar to the sensor used in the tympanic thermometer. The superficial temporal artery extends directly from the external carotid artery and travels in front of the ear. Anatomically, it is lying approximately 1 mm below the skin, readily accessible, and maintains good blood perfusion. Assuming zero thermal loss to the environment, the skin surface over the temporal artery area would be at the same temperature as the arterial blood in the aorta, which is essentially equal to core body temperature. Since ambient temperature

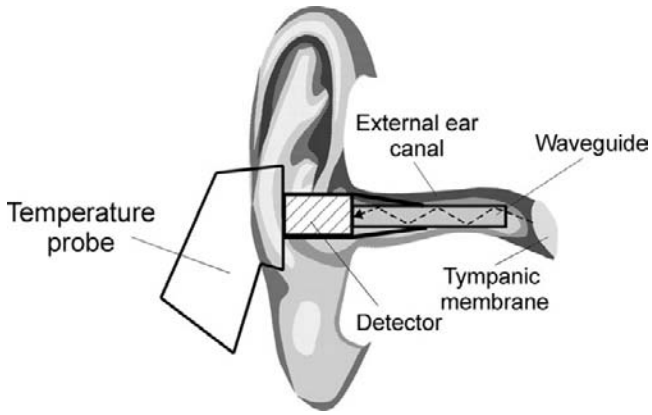


FIGURE 10.24 Noncontact-type infrared ear thermometer.

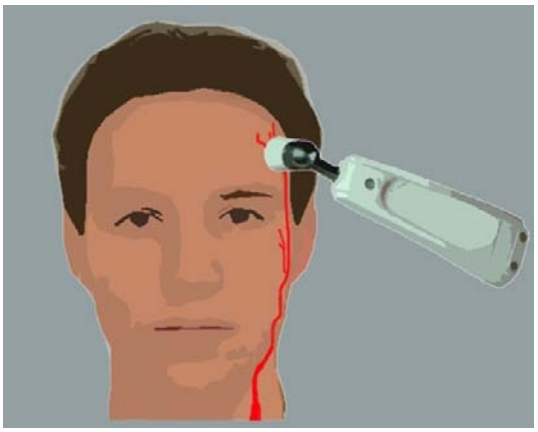


FIGURE 10.25 Temporal artery thermometry.

is normally lower than core body temperature, there is a cooling effect at the skin surface due to the radiative heat loss to the surrounding air. To account for errors due to the natural heat loss, the hand-held scanning thermometer measures ambient temperature at the same time it measures the absolute temperature of the skin surface over the temporal artery and computes arterial temperature using a heat balance equation.

Ingestible Temperature Pill

Heat exhaustion (also known as hyperthermia or heatstroke) occurs when the body cannot adequately dissipate an internal rise in core body temperature. It can be caused by an excessive exposure to heat or dehydration—particularly in football players during sporting activities, astronauts during space flights, or soldiers, and can ultimately lead to life-threatening brain damage or even death. To minimize the potential of heatstroke, an ingestible thermometer pill was developed in the mid-1980s in collaboration with NASA (Figure 10.26) to monitor core body temperature in real-time with an accuracy of 0.1°C. Once ingested into the body, the $^3/_4$-inch-long battery-operated pill transmits wirelessly core body temperature as it travels harmlessly through the gastrointestinal tract. A small quartz crystal oscillator inside the pill vibrates at a frequency that is proportional to core body temperature. The pill converts this change in crystal frequency to a magnetic field with a radius of about 1 meter that can be picked up wirelessly by an external data recorder. The silicone-coated pill remains in the body for about 24–36 hours before it is excreted.

10.4 BLOOD GAS SENSORS

Measurements of arterial blood gases (pO_2, pCO_2 and pH) are frequently performed on critically ill patients in both the operating room and the intensive care unit. They are used by the physician to adjust mechanical ventilation or to administer pharmacological agents. These measurements provide information about the respiratory and metabolic imbalances in the body and reflect the adequacy of blood oxygenation and CO_2 elimination.

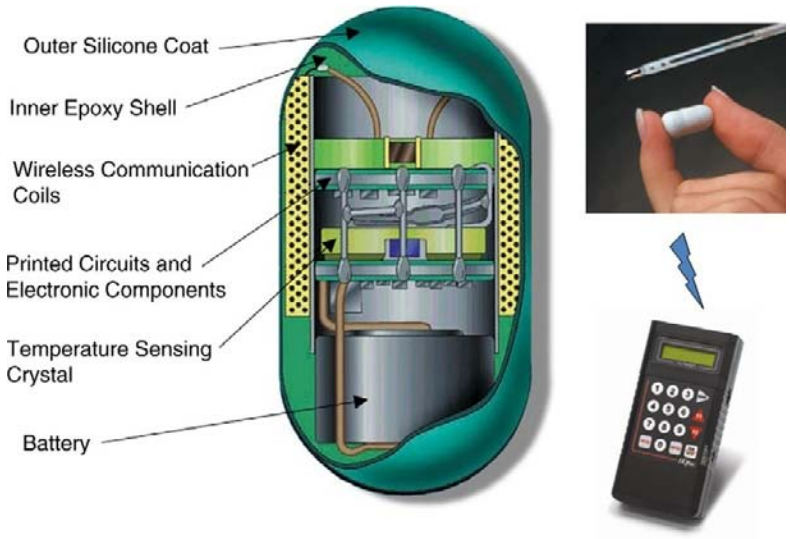


FIGURE 10.26 Ingestible temperature pill. *Courtesy of HQ, Inc., Palmetto, FL.*

Traditionally, arterial blood gas analysis has been performed by withdrawing blood from a peripheral artery. The blood sample is then transported to a clinical laboratory for analysis. The need for rapid test results in the management of unstable, critically ill patients has led to the development of newer methods for continuous noninvasive blood gas monitoring. This allows the physician to follow trends in the patient's condition as well as receive immediate feedback on the adequacy of certain therapeutic interventions.

Noninvasive sensors for measuring O_2 and CO_2 in arterial blood are based on the discovery that gases, such as O_2 and CO_2, can easily diffuse through the skin. Diffusion occurs due to a partial pressure difference between the blood in the superficial layers of the skin and the outermost surface of the skin. This concept has been used to develop two types of noninvasive electrochemical sensors for transcutaneous monitoring of pO_2 and pCO_2. Furthermore, the discovery that blood changes its color depending on the amount of oxygen chemically bound to the hemoglobin in the erythrocytes has led to the development of several optical methods to measure the oxygen saturation in blood.

10.4.1 Oxygen Measurement

A quantitative method for measuring blood oxygenation is of great importance in assessing the circulatory and respiratory condition of a patient. Oxygen is transported by the blood from the lungs to the tissues in two distinct states. Under normal physiological conditions, approximately 2 percent of the total amount of oxygen carried by the blood is dissolved in the plasma. This amount is linearly proportional to the blood pO_2. The remaining 98 percent is carried inside the erythrocytes in a loose, reversible chemical combination with hemoglobin (Hb) as oxyhemoglobin (HbO_2). Thus, there are two options for measuring blood oxygenation: either using a polarographic pO_2 sensor or measuring oxygen saturation (the relative amount of HbO_2 in the blood) by means of an optical oximeter.

A pO_2 sensor, also widely known as a Clark electrode, is used to measure the partial pressure of O_2 gas in a sample of air or blood. This sensor is categorized as an amperometric (i.e., the measurement is based on the production of a current when a voltage is applied between two electrodes) sensor and requires an external polarizing bias voltage source. The measurement is based on the principle of polarography, as shown in Figure 10.27. The electrode utilizes the ability of O_2 molecules to react chemically with H_2O in the presence of electrons to produce hydroxyl (OH^-) ions. This electrochemical reaction, called an oxidation/reduction or redox reaction, generates a small current and requires an externally applied constant polarizing voltage source of about 0.6V.

Oxygen is reduced (consumed) at the surface of a noble metal (e.g., platinum or gold) cathode (the electrode connected to the negative side of the voltage source) according to the following chemical reaction:

$$O_2 + 2H_2O + 4e^- \leftrightarrow 4OH^-$$

In this reduction reaction, an O_2 molecule takes four electrons and reacts with two water molecules, generating four hydroxyl ions. The resulting OH^- ions migrate and react with a reference Ag/AgCl anode (the electrode connected to the positive side of the voltage source), causing a two-step oxidation reaction to occur as follows:

$$Ag \leftrightarrow Ag^+ + e^-$$
$$Ag^+ + Cl^- \leftrightarrow AgCl\downarrow$$

In this oxidation reaction, silver from the electrode is first oxidized to silver ions, and electrons are liberated to the anode. These silver ions are immediately combined with

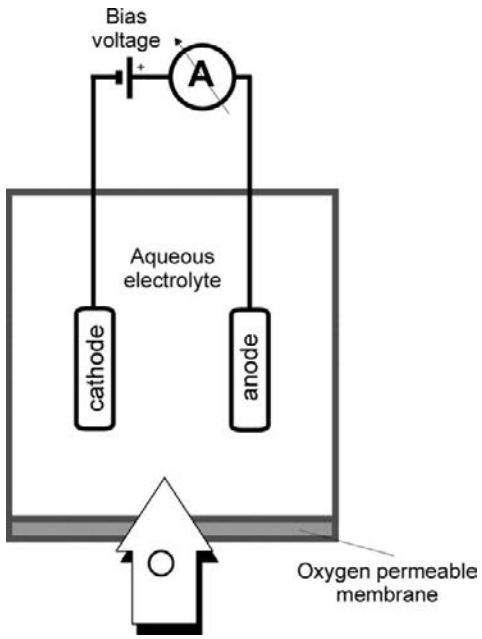


FIGURE 10.27 Principle of a polarographic Clark-type pO_2 sensor.

chloride ions to form silver chloride that precipitates on the surface of the anode. The current flowing between the anode and the cathode in the external circuit produced by this reaction is directly (i.e., linearly) proportional to the number of O_2 molecules constantly reduced at the surface of the cathode. The electrodes in the polarographic cell are immersed in an electrolyte solution of potassium chloride and surrounded by an O_2-permeable Teflon or polypropylene membrane that permits gases to diffuse slowly into the electrode. Thus, by measuring the change in current between the cathode and the anode, the amount of oxygen that is dissolved in the solution can be determined.

With a rather minor change in the configuration of a polarographic pO_2 sensor, it is also possible to measure the pO_2 transcutaneously. Figure 10.28 illustrates a cross section of a Clark-type transcutaneous pO_2 sensor. This sensor is essentially a standard polarographic pO_2 electrode that is attached to the surface of the skin by double-sided adhesive tape. It measures the partial pressure of oxygen that diffuses from the blood through the skin into the Clark electrode, similar to the way it measures the pO_2 in a sample of blood. However, since the diffusion of O_2 through the skin is normally very low, a miniature heating coil is incorporated into the housing of this electrode to cause gentle vasodilatation (increased local blood flow) of the capillaries in the skin. By raising the local skin temperature to about 43°C, the pO_2 measured by the transcutaneous sensor approximates that of the underlying arterial blood. This electrode has been used extensively in monitoring newborn babies in the intensive care unit. However, as the skin becomes thicker and matures in adult patients, the gas diffusion properties of the skin change significantly and cause large errors that result in inconsistent readings.

Various methods for measuring the oxygen saturation, SO_2 (the relative amount of oxygen carried by the hemoglobin in the erythrocytes), of blood in vitro or in vivo in arterial blood (S_aO_2) or mixed venous blood (S_vO_2), have been developed. This method, referred to as oximetry, is based on the light absorption properties of blood and, in particular, the

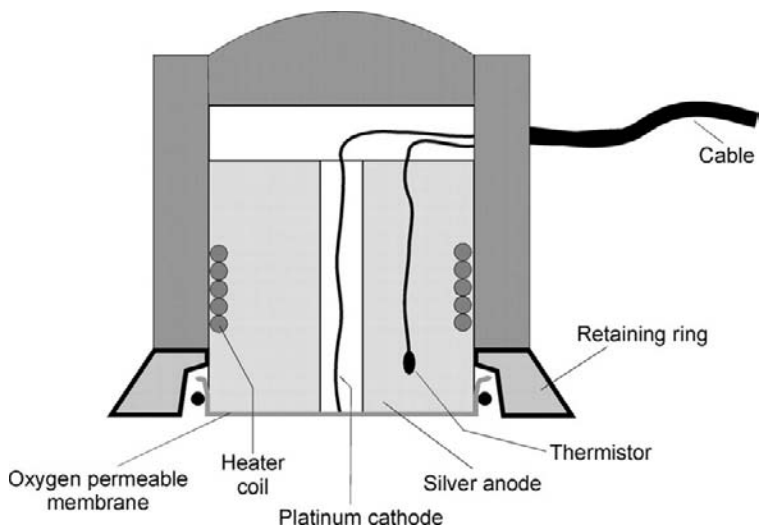


FIGURE 10.28 Transcutaneous pO_2 sensor.

relative concentration of Hb and HbO_2, since the characteristic color of deoxygenated blood is blue, whereas fully oxygenated blood has a distinct bright red color.

The measurement is performed at two specific wavelengths: a red wavelength, λ_1, where there is a large difference in light absorbance between Hb and HbO_2 (e.g., 660 nm), and a second wavelength, λ_2, in the near-infrared region of the spectrum. The second wavelength can be either isobestic (a region of the spectrum around 805 nm, where the absorbencies of Hb and HbO_2 are equal) or around 940–960 nm, where the absorbance of Hb is slightly smaller than that of HbO_2. Figure 10.29 shows the optical absorption spectra of blood in the visible and near-infrared region.

The measurement is based on Beer-Lambert's law that relates the transmitted light power, P_t, to the incident light power, P_0, according to the following relationship:

$$P_t = P_0 \times 10^{-abc} \tag{10.19}$$

where a is a wavelength-dependent constant called the extinction coefficient (or molar absorptivity) of the sample, b is the light path length through the sample, and c is the concentration of the sample.

Assuming for simplicity that (i) $\lambda_1 = 660$ nm and $\lambda_2 = 805$ nm (i.e., isobestic), (1) the hemolyzed blood sample (blood in which the erythrocytes have been ruptured—i.e., the hemoglobin has been released and uniformly mixed with the plasma) consists of a two-component mixture of Hb and HbO_2, and (2) the total light absorbance by the mixture of these two components is additive, a simple mathematical relationship can be derived for computing the oxygen saturation of blood:

$$SO_2 = A - B \times \left[\frac{OD(\lambda_1)}{OD(\lambda_2)}\right] \tag{10.20}$$

where A and B are two coefficients that are functions of the specific absorptivity of Hb and HbO_2, OD (or absorbance) is defined as the optical density—that is, $\log_{10}(1/T)$—where

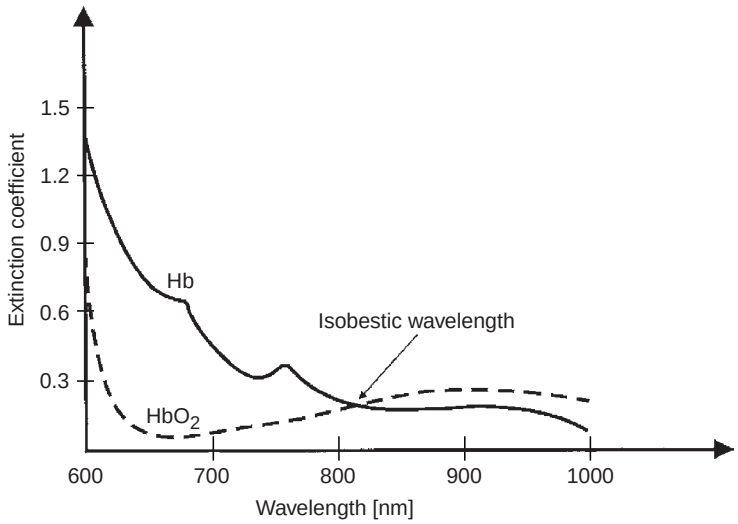


FIGURE 10.29 Optical extinction coefficients of Hb and HbO_2.

T represents the light transmission through the sample and is given by P_t/P_0, and SO_2 is defined as $c_{HB}/(c_{HB} + c_{HBO_2})$.

The measurement of SO_2 in blood can be performed either in vitro or in vivo. In vitro measurement using a bench top oximeter requires a sample of blood, usually withdrawn from a peripheral artery. The sample is transferred into an optical cuvette (a parallel-wall glass container that holds the sample), where it is first hemolyzed and then illuminated sequentially by light from an intense white source after proper wavelength selection using narrow-band optical filters.

SO_2 can also be measured in vivo using a noninvasive pulse oximeter. Noninvasive optical sensors for measuring S_aO_2 by a pulse oximeter consist of a pair of small and inexpensive light-emitting diodes (LEDs)—typically a red (R) LED around 660 nm and an infrared (IR) LED around 940–960 nm—and a single, highly sensitive silicon photodetector. These components are typically mounted inside a reusable spring-loaded clip or a disposable adhesive wrap (Figure 10.30). Electronic circuits inside the pulse oximeter generate digital switching signals to turn on and off the two LEDs in a sequential manner and synchronously

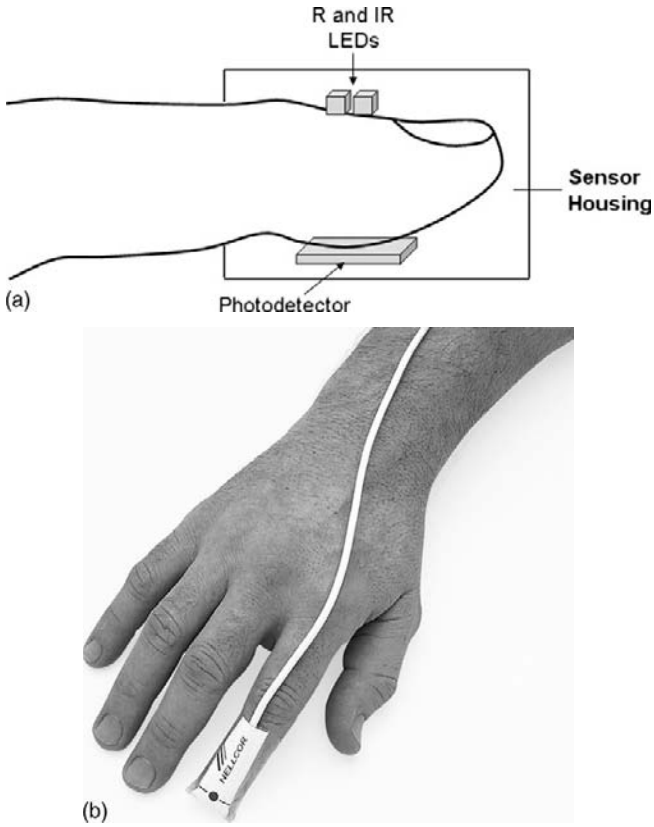


FIGURE 10.30 (a) Transmission-type pulse oximeter finger probe, and (b) disposable finger sensor. *Courtesy of Nellcor Puritan Bennett, Inc., Pleasanton, CA.*

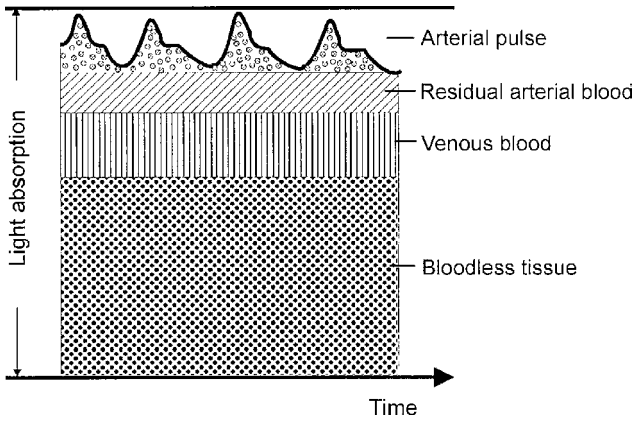


FIGURE 10.31 Time dependence of light absorption by a peripheral vascular tissue bed showing the effect of arterial pulsation.

measure the photodetector output when the corresponding LEDs are turned on. The sensor is usually attached either to the fingertip or earlobe so the tissue is sandwiched between the light source and the photodetector.

Pulse oximetry relies on the detection of a photoplethysmographic signal, as shown in Figure 10.31. This signal is caused by changes in the arterial blood volume associated with periodic contraction of the heart during systole. The magnitude of this signal depends on the amount of blood ejected from the heart into the peripheral vascular bed with each cardiac cycle, the optical absorption of the blood, the composition and color of the skin and underlying tissues, and the wavelengths used to illuminate the blood. S_aO_2 is derived by analyzing the magnitude of the red and infrared photoplethysmograms measured by the photodetector. Electronic circuits separate the red and infrared photoplethysmograms into their pulsatile (AC) and nonpulsatile (DC) signal components. An algorithm inside the pulse oximeter performs a mathematical normalization by which the AC signal at each wavelength is divided by the corresponding DC component that is mainly due to the light absorbed by the bloodless tissue, residual arterial blood when the heart is in diastole, venous blood, and skin pigmentation. Since it is assumed that the AC portion in the photoplethysmogram results only from the pulsatile arterial blood component, this scaling process provides a normalized red/infrared ratio, R, which is highly dependent on the color of the arterial blood and is therefore related to S_aO_2 but is largely independent of the volume of arterial blood entering the tissue during systole, skin pigmentation, and thickness. Hence, the instrument does not need to be recalibrated for measurements on different patients. The mathematical relationship between S_aO_2 and R is programmed by the manufacturer into the pulse oximeter.

10.4.2 pH Electrodes

pH describes the balance between acid and base in a solution. Acid solutions have an excess of hydrogen ions (H^+), whereas basic solutions have an excess of hydroxyl ions (OH^-). In a dilute solution, the product of these ion concentrations is a constant (1.0×10^{-14}).

Therefore, the concentration of either ion can be used to express the acidity or alkalinity of a solution. All neutral solutions have a pH of 7.0.

The measurement of blood pH is fundamental to many diagnostic procedures. In normal blood, pH is maintained under tight control and is typically around 7.40 (slightly basic). By measuring blood pH, it is possible to determine whether the lungs are removing sufficient CO_2 gas from the body or how well the kidneys regulate the acid-base balance.

pH electrodes belong to group of potentiometric sensors (i.e., electrochemical sensors producing a voltage). These sensors generate a small potential difference without the need to polarize the electrochemical cell. A pH electrode essentially consists of two separate electrodes: a reference electrode and an active (indicator) electrode, as shown in Figure 10.32. The two electrodes are typically made of an Ag/AgCl wire dipped in a KCl solution and encased in a glass container. A salt bridge, which is essentially a glass tube containing an electrolyte enclosed in a membrane that is permeable to all ions, maintains the potential of the reference electrode at a constant value regardless of the solution under test. Unlike the reference electrode, the active electrode is sealed with hydrogen-impermeable glass except at the tip. The reference electrode may also be combined with the indicator electrode in a single glass housing.

The boundary separating two solutions has a potential proportional to the hydrogen ion concentration of one solution and, at a constant temperature of 25°C, is given by

$$V = -59\ \text{mV} \times \log_{10}[H^+] + C \tag{10.21}$$

where C is a constant. Since pH is defined as

$$\text{pH} = -\log_{10}[H^+] \tag{10.22}$$

the potential of the active pH electrode, V, is proportional to the pH of the solution under test and is equal to

$$V = 59\ \text{mV} \times \text{pH} + C \tag{10.23}$$

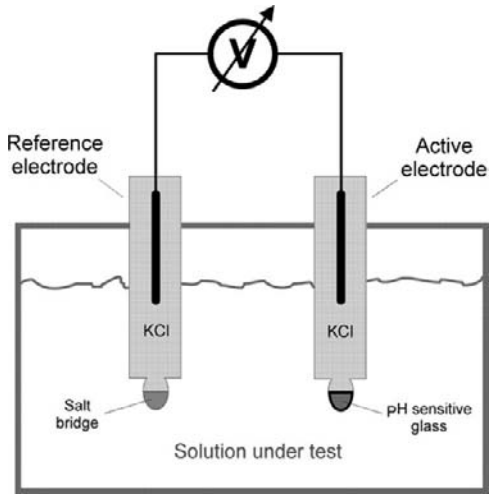


FIGURE 10.32 Principle of a pH electrode.

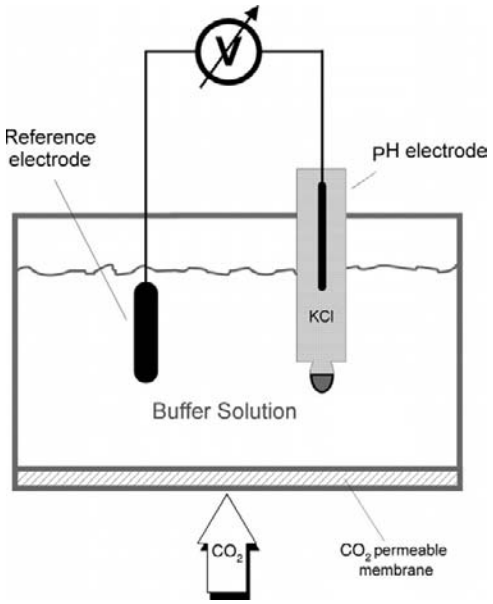


FIGURE 10.33 Principle of a pCO_2 electrode.

The value of *C* is usually compensated for electronically when the pH electrode is calibrated by placing the electrode inside different buffer solutions with known pH values.

10.4.3 Carbon Dioxide Sensors

Electrodes for measurement of partial pressure of CO_2 in blood or other liquids are based on measuring the pH, as illustrated in Figure 10.33. The measurement is based on the observation that when CO_2 is dissolved in water, it forms a weakly dissociated carbonic acid (H_2CO_3) that subsequently forms free hydrogen and bicarbonate ions according to the following chemical reaction:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

As a result of this chemical reaction, the pH of the solution is changed. This change generates a potential between the glass pH and a reference (e.g., Ag/AgCl) electrode that is proportional to the negative logarithm of the pCO_2.

10.5 BIOANALYTICAL SENSORS

Biomolecules such as enzymes, antibodies, and microorganisms, as well as animal and plant cells, have been used as biological sensing elements. Among these, microorganisms offer unique advantages by their ability to detect a wide range of chemical substances, amenability to genetic modification, and broad operating pH and temperature range, making them ideal as biological sensing materials.

The number of analytes that can be measured with biosensors can be increased significantly by adding biologically specific mediators (reagents that either undergo reactions or act as catalysts) to the semipermeable membrane structure. Several biosensors that have been constructed and used mainly for research applications have different enzymes and bacteria as the primary sensing elements. Although these biosensors have been used successfully in vitro to demonstrate unique medical and industrial applications, further technical improvements are necessary to make these sensors robust and reliable enough to fulfill the demanding requirements of routine analytical and clinical applications. Examples of some interesting sensor designs are given in the following sections.

10.5.1 Enzyme-Based Biosensors

Enzymes are the most widely used biological sensing element in the fabrication of various biosensors. Enzymes constitute a group of more than 2,000 proteins having so-called biocatalytic properties. These properties give the enzymes the unique and powerful ability to accelerate chemical reactions inside biological cells. Most enzymes react only with specific substrates even though they may be contained in a complicated mixture with other substances. It is important to keep in mind, however, that soluble enzymes are very sensitive both to temperature and pH variations, and they can be inactivated by many chemical inhibitors. For practical biosensor applications, these enzymes are normally immobilized by insolubilizing the free enzymes via entrapment into an inert and stable matrix such as starch gel, silicon rubber, or polyacrylamide. This process is important to ensure that the enzyme retains its catalytic properties and can be reusable.

The action of specific enzymes can be utilized to construct a range of different biosensors. A typical example of an enzyme-based sensor is a glucose sensor that uses the enzyme glucose oxidase. Glucose plays an important role in metabolic processes. In patients suffering from diabetes mellitus, the pancreas does not produce sufficient amounts of insulin to control adequately the level of glucose in their blood. Therefore, to manage the disease, these patients must monitor and regulate their blood glucose level on a regular basis by medication and insulin injections. Currently available glucose sensors are based on an immobilized enzyme, such as glucose oxidase, that acts as a catalyst. Glucose is detected by measuring electrochemically either the amount of gluconic acid or hydrogen peroxide (H_2O_2) produced or by measuring the amount of oxygen consumed, according to the following chemical reaction:

$$\text{Glucose} + O_2 + H_2O \xleftarrow{\text{glucose oxidase}} \text{gluconic acid} + H_2O_2$$

A glucose sensor is similar to a pO_2 sensor and is shown in Figure 10.34. Glucose and oxygen enter through the outside membrane to allow glucose to interact with the glucose oxidase enzyme. The remaining oxygen penetrates through the second oxygen-permeable membrane and is measured by the oxygen electrode.

Biocatalytic enzyme-based sensors generally consist of an electrochemical gas-sensitive transducer or an ion-selective electrode with an enzyme immobilized in or on a membrane that serves as the biological mediator. The analyte diffuses from the bulk sample solution into the biocatalytic layer, where an enzymatic reaction takes place. The electroactive

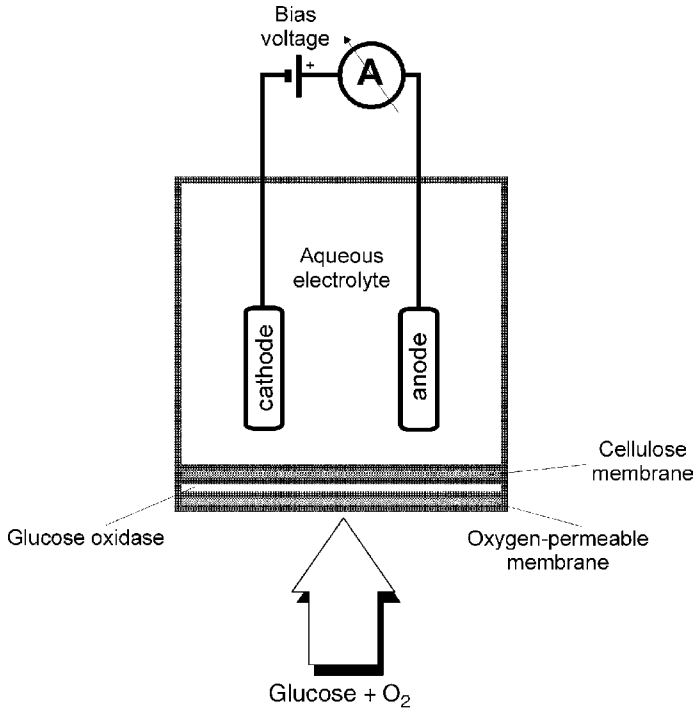


FIGURE 10.34 Principle of a glucose sensor.

product that is formed (or consumed) is usually detected by an ion-selective electrode. A membrane separates the basic sensor from the enzyme if a gas is consumed (such as O_2) or is produced (such as CO_2 or NH_3). Although the concentration of the bulk substrate drops continuously, the rate of consumption is usually negligible. The decrease is detected only when the test volume is very small or when the area of the enzyme membrane is large enough. Thus, this electrochemical analysis is nondestructive, and the sample can be reused. Measurements are usually performed at a constant pH and temperature either in a stirred medium solution or in a flow-through solution.

10.5.2 Microbial Biosensors

The fundamental basis of a microbial biosensor is the close proximity between an immobilized microorganism that serves as a specific recognition element and an electrochemical or optical sensing transducer that is used to convert the biochemical signal into an electronic signal that can be processed.

The fabrication of a microbial biosensor requires the immobilization of the microorganisms on a transducer by chemical or physical methods. Since the response, operational stability, and long-term use of a microbial biosensor is a function of the immobilization strategy used, immobilization technology plays a very important role in the successful design of microbial biosensors, and the choice of immobilization technique is critical.

Chemical methods of microbe immobilization include covalent binding and cross-linking. Covalent binding methods rely on the formation of a stable covalent bond between functional groups of the cell wall components of the microorganism and the transducer. To successfully achieve this goal, whole cells are exposed to harsh chemical reactions that can damage the microbial cell membrane and decrease the biological viability of the cells. Determining how to overcome this drawback remains a practical challenge. Cross-linking, on the other hand, involves bridging between functional groups on the outer cell membrane by multifunctional reagents (e.g., glutaraldehyde) to form a network. Because of the speed and simplicity, the method has found wide acceptance for the immobilization of microorganisms. The cells may be cross-linked directly onto the transducer surface or on a removable support membrane that can then be placed on the transducer. While cross-linking has advantages over covalent binding, the cell viability can be affected by the cross-linking agents. Therefore, cross-linking is suitable in constructing microbial biosensors where cell viability is not important and only the intracellular enzymes are involved in the detection.

Physical methods of microbe immobilization include adsorption and entrapment. Because these methods do not involve covalent bond formation with microbes and provide relatively small perturbation of microorganism native structure and function, these methods are preferred when viable cells are required. Physical adsorption is the simplest method for microbe immobilization. Typically, a microbial suspension is incubated with the electrode or an immobilization matrix, such as glass bead. The microbes are immobilized due to adsorptive interactions (i.e., ionic or polar bonding) and hydrophobic interaction. However, immobilization using adsorption alone generally leads to poor long-term stability because of desorption of microbes. The immobilization of microorganisms by entrapment can be achieved, for example, by the retention of the cells in close proximity of the transducer surface using a dialysis membrane. However, a major disadvantage of entrapment immobilization is the additional diffusion resistance offered by the entrapment material, which will result in a lower sensitivity and detection limit.

A number of microbial sensors have been developed mainly for online control of biochemical processes in various environmental, agricultural, food, and pharmaceutical applications. Microbial biosensors typically involve the assimilation of organic compounds by the microorganisms, followed by a change in respiration activity (metabolism) or the production of specific electrochemically active metabolites, such as H_2, CO_2, or NH_3, that are secreted by the microorganism.

Examples of microbial biosensors include ammonia (NH_3) and nitrogen dioxide (NO_2) sensors that utilize nitrifying bacteria as the biological sensing component. An ammonia biosensor can be constructed based on nitrifying bacteria, such as *Nitrosomonas sp.*, that use ammonia as a source of energy and oxidize ammonia as follows:

$$NH_3 + 1.5O_2 \xrightarrow{\textit{Nitrosomonas sp.}} NO_2 + H_2O + H^+$$

This oxidation process proceeds at a high rate, and the amount of oxygen consumed by the immobilized bacteria can be measured directly by a polarographic oxygen electrode placed behind the bacteria.

Nitric oxide (NO) and NO_2 are the two principal pollution gases of nitrogen in the atmosphere. The principle of a NO_2 biosensor is shown in Figure 10.35. When a sample of

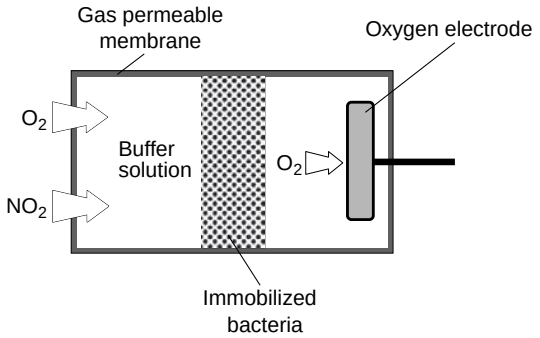


FIGURE 10.35 Principle of an NO_2 microbial-type biosensor.

NO_2 gas diffuses through the gas-permeable membrane, it is oxidized by the *Nitrobacter sp.* bacteria as follows:

$$2NO_2 + O_2 \xrightarrow{\textit{Nitrosomonas sp.}} 2NO_3$$

Similar to an ammonia biosensor, the consumption of O_2 around the membrane is determined by an electrochemical oxygen electrode.

The use of microbial cells in electrochemical sensors offers several advantages over enzyme-based electrodes, the principal one being the increased electrode lifetime to several weeks. On the other hand, microbial sensors may be less favorable compared with enzyme electrodes with respect to specificity and response time.

10.6 OPTICAL SENSORS

Optical sensors play an important role in the development of highly sensitive and selective methods for biochemical analysis. The fundamental principle that is employed is based on the change in optical properties of a biological or physical medium. The change produced can be the result of physical perturbations or intrinsic changes in absorbance, reflectance, scattering, fluorescence, polarization, or refractive index of the biological medium.

10.6.1 Optical Fibers

Optical fibers can be used to transmit light from one location to another with minimal attenuation and without any transport of heat from the light source. Therefore, they are used in a whole range of miniature sensors for biomedical applications. Optical fibers are small, flexible, and intrinsically immune to electromagnetic and radio frequency interferences. They can produce an instantaneous response to subtle changes in the microenvironments that surround their optical surface. Therefore, optical fibers permit in vivo measurements in small blood vessels or in delicate tissues such as the brain.

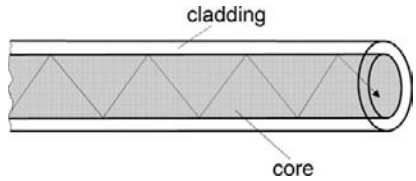


FIGURE 10.36 Principle of optical fibers.

Optical fibers are typically made from two concentric and transparent glass or plastic materials, as shown in Figure 10.36. The center piece is known as the core, and the outer layer, which serves as a coating material, is called the cladding.

The core and cladding of an optical fiber have a different index of refraction, n. The index of refraction is a number that expresses the ratio of the light velocity in free space to its velocity in a specific material. For instance, the refractive index for air is equal to 1.0, whereas the refractive index for water is equal to 1.33. Assuming that the refractive index of the core material is n_1 and the refractive index of the cladding is n_2 (where $n_1 > n_2$), according to Snell's law,

$$n_1 \times \sin\phi_1 = n_2 \times \sin\phi_2 \tag{10.24}$$

where ϕ is the angle of incidence, as shown in Figure 10.37.

Accordingly, any light passing from a lower refractive index to a higher refractive index is bent toward the line perpendicular to the interface of the two materials. For small incident angles, ϕ_1, the light ray enters the fiber core and bends inward at the first core/cladding interface. For larger incident angles, ϕ_2, the ray exceeds a minimum angle required to bend it back into the core when it reaches the core/cladding boundary. Consequently, the light escapes into the cladding. By setting $\sin\phi_2 = 1.0$, the critical angle, ϕ_{cr}, is given by

$$\sin\phi_{cr} = \frac{n_2}{n_1} \tag{10.25}$$

Any light rays entering the optical fiber with incidence angles greater than ϕ_{cr} are internally reflected inside the core of the fiber by the surrounding cladding. Conversely, any entering light rays with incidence angles smaller than ϕ_{cr} escape through the cladding and are therefore not transmitted by the core.

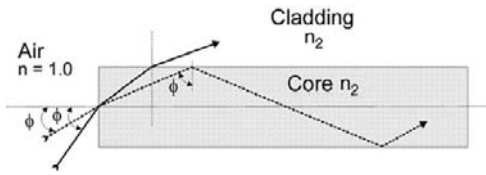


FIGURE 10.37 Optical fiber illustrating the incident and refracted light rays. The solid line shows the light ray escaping from the core into the cladding. The dashed line shows the ray undergoing total internal reflection inside the core.

EXAMPLE PROBLEM 10.14

Assume that a beam of light passes from a layer of glass with a refractive index $n_1 = 1.47$ into a second layer of glass with a refractive index of $n_2 = 1.44$. Using Snell's law, calculate the critical angle for the boundary between these two glass layers.

Solution

$$\phi_{cr} = \arcsin \times \left(\frac{n_2}{n_1}\right) = \arcsin(0.9796)$$

$$\phi_{cr} = 78.4^\circ$$

Therefore, light that strikes the boundary between these two glasses at an angle greater than 78.4° will be reflected back into the first layer.

The propagation of light along an optical fiber is not confined to the core region. Instead, the light penetrates a characteristic short distance (on the order of one wavelength) beyond the core surface into the less optically dense cladding medium. This effect causes the excitation of an electromagnetic field, called the "evanescent-wave," that depends on the angle of incidence and the incident wavelength. The intensity of the evanescent-wave decays exponentially with distance, according to Beer-Lambert's law. It starts at the interface and extends into the cladding medium.

10.6.2 Sensing Mechanisms

Optical sensors are typically interfaced with an optical module, as shown in Figure 10.38. The module supplies the excitation light, which may be from a monochromatic source such as a diode laser or from a broadband source (e.g., quartz-halogen) that is filtered to provide a narrow bandwidth of excitation. Typically, two wavelengths of light are used: one that is sensitive to changes in the species to be measured and one that is unaffected by changes in the analyte concentration. This wavelength serves as a reference and is used to compensate for fluctuations in source output and detector stability. The light output from the optic module is coupled into a fiber optic cable through appropriate lenses and an optical connector.

Several optical techniques are commonly used to sense the optical change across a biosensor interface. These are usually based on evanescent wave spectroscopy, which plays a major role in fiber optic sensors, and a surface plasmon resonance principle.

In fluorescence-based sensors, the incident light excites fluorescence emission, which changes in intensity as a function of the concentration of the analyte to be measured. The emitted light travels back down the fiber to the monitor, where the light intensity is measured by a photodetector. In other types of fiber optic sensors, the light-absorbing properties of the sensor chemistry change as a function of analyte chemistry. In the absorption-based design, a reflective surface near the tip or some scattering material within the sensing chemistry itself is usually used to return the light back through the same optical

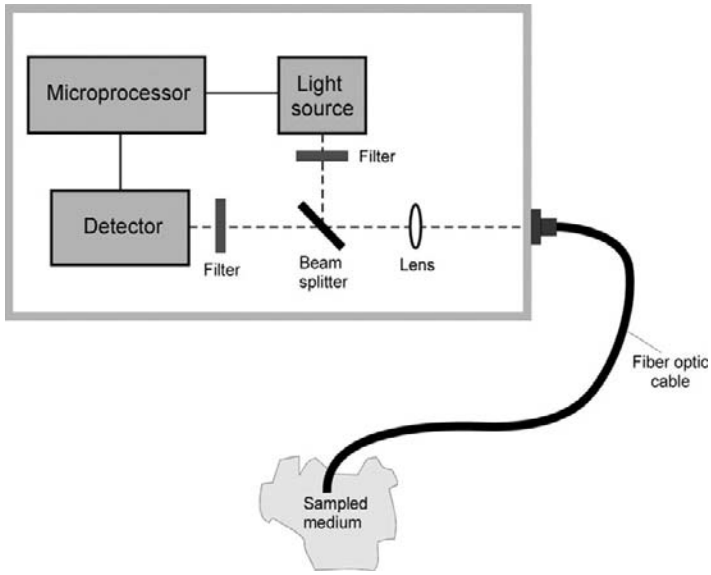


FIGURE 10.38 General principle of a fiber optic–based sensor.

fiber. Other sensing mechanisms exploit the evanescent-wave interaction with molecules that are present within the penetration depth distance and lead to attenuation in reflectance related to the concentration of the molecules. Because of the short penetration depth and the exponentially decaying intensity, the evanescent-wave is absorbed by compounds that must be present very close to the surface. The principle has been used to characterize interactions between receptors that are attached to the surface and ligands that are present in solution above the surface.

The key component in the successful implementation of evanescent-wave spectroscopy is the interface between the sensor surface and the biological medium. Receptors must retain their native conformation and binding activity, and sufficient binding sites must be present for multiple interactions with the analyte. In the case of particularly weak absorbing analytes, sensitivity can be enhanced by combining the evanescent-wave principle with multiple internal reflections along the sides of an unclad portion of a fiber optic tip. Alternatively, instead of an absorbing species, a fluorophore[2] can also be used. Light that is absorbed by the fluorophore emits detectable fluorescent light at a higher wavelength, thus providing improved sensitivity.

10.6.3 Intravascular Fiber Optic Blood Gas Sensors

Considerable effort has been devoted over the last three decades to develop disposable extracorporeal sensors (for ex vivo applications) or intra-arterial fiber optic sensors that

[2]A compound that produces a fluorescent signal in response to light.

can be placed in the arterial line (for in vivo applications) to enable continuous trending of arterial blood gases. With the advent of continuous arterial blood gas monitoring, treatment modalities can be proactive rather than reactive, which is vital for therapeutic interventions in ICU patients who may experience spontaneous and often unexpected changes in acid-base status.

Intra-arterial blood gas sensors typically employ a single- or a double-fiber configuration. Typically, the matrix containing the indicator is attached to the end of the optical fiber. Since the solubility of O_2 and CO_2 gases, as well as the optical properties of the sensing chemistry itself, is affected by temperature variations, fiber optic intravascular sensors include a thermocouple or thermistor wire running alongside the fiber optic cable to monitor and correct for temperature fluctuations near the sensor tip. A nonlinear response is characteristic of most chemical indicator sensors. Therefore, the operating range of these sensors is typically optimized to match the range of concentrations according to the intended application.

Intra-arterial fiber optic blood gas sensors are normally placed inside a standard 20-gauge arterial cannula that is sufficiently small, thus allowing adequate spacing between the sensor and the catheter wall. The resulting lumen is large enough to permit the withdrawal of blood samples, introduction of a continuous or intermittent anticoagulant (e.g., heparin) flush, and the recording of a blood pressure waveform. In addition, the optical fibers are encased in a protective tubing to contain any fiber fragments in case they break off. The material in contact with the blood is typically treated with a covalently bonded layer of heparin, resulting in low susceptibility to fibrin deposition. Despite excellent accuracy of indwelling intra-arterial catheters in vitro compared to blood gas analyzers, when these multiparameter probes were first introduced into the vascular system, it quickly became evident that the readings (primarily pO_2) vary frequently and unpredictably, mainly due to the sensor tip intermittently coming in contact with the wall of the arterial blood vessel and intermittent reductions in blood flow due to arterial vasospasm.

A more advanced multiparameter disposable probe (Figure 10.39) consisting of pO_2, pCO_2, and pH sensors was developed by Diametrics Medical, Inc. The sensor has a diameter of 0.5 mm and can be inserted intravascularly through a 20-gauge indwelling cannula. Clinical studies confirmed that the system is adequate for trend monitoring, eliminating invasive blood sampling, potential errors in analysis, and significant delays in obtaining results, which may affect treatment. The device has been evaluated in neurosurgical patients for continuous monitoring in the brain and in critically ill pediatric patients.

Fiber Optic pO_2 Sensors

Various fiber optic sensors were developed to measure pO_2 in blood based on the principle of fluorescence quenching. Quenching reduces the intensity of the emitted fluorescence light and is related to the concentration of the quenching molecules. For example, quenching can result from collisions encountered between the fluorophore (a fluorescent substance) and the quencher. For quenching to occur, the fluorophore and quencher must be in contact. When light is absorbed by a molecule, the absorbed energy is held as an excited electronic state of the molecule. It is then lost by coupling to the mechanical movement of the molecule (heat), reradiated from the molecule in a mean time of about 10 ns (fluorescence), or converted to another excited state with a much longer mean lifetime

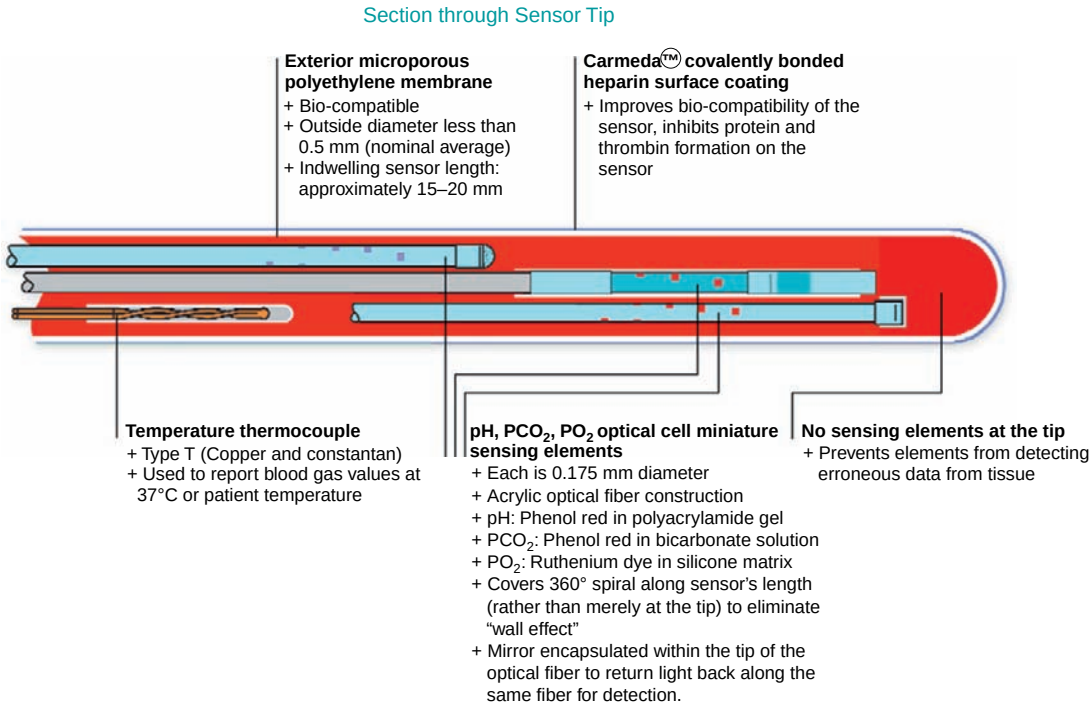


FIGURE 10.39 Principle of an indwelling arterial optical blood gas catheter. A heparin-coated porous polyethylene membrane encapsulates the optical fibers and a thermistor. *Courtesy of Diametrics, Inc., St. Paul, MN.*

(phosphorescence). A wide variety of substances act as fluorescence quenchers. One of the best-known quenchers is molecular oxygen.

A typical fiber optic sensor for measuring pO_2 using the principle of fluorescence quenching consists of a dye that is excited at 470 nm (blue) and fluoresces at 515 nm (green) with an emitted intensity that depends on the pO_2. If the excited dye encounters an oxygen molecule, the excess energy will be transferred to the oxygen molecule, decreasing the fluorescence signal. The degree of quenching depends on the concentration of oxygen. The optical information is derived from the ratio of light intensities measured from the green fluorescence and the blue excitation light, which serves as an internal reference signal. The ratio of green to blue intensity is described by the Stern-Volmer equation

$$I_o/I = 1 + KpO_2 \tag{10.26}$$

where I_o and I are the fluorescence emission intensities in the absence (i.e., $pO_2 = 0$) and presence of the oxygen quencher, respectively. K is the Stern-Volmer quenching coefficient, which is dependent on temperature. The method provides a nearly linear readout of pO_2 over the range of 0–150 mmHg (0–20 kPa), with a precision of about 1 mmHg (0.13 kPa). The slope of the plot described by Eq. (10.26) is a measure of the oxygen sensitivity of the sensor. Note that the sensor will be most sensitive to low levels of oxygen.

Fiber Optic pH Sensors

A fiber optic pH sensor can be designed by placing a reversible color-changing dye as an indicator at the end of a pair of optical fibers. For example, the popular indicator phenol red can be used because this dye changes its absorption properties from the green to the blue part of the spectrum as the acidity is increased. The dye can be covalently bound to a hydrophilic polymer in the form of water-permeable microbeads to stabilize the indicator concentration. The indicator beads are contained in a sealed hydrogen ion-permeable envelope made out of hollow cellulose tubing, forming a miniature spectrophotometric cell at the end of the optical fibers.

The phenol red dye indicator is a weak organic acid, and its unionized acid and base forms are present in a concentration ratio that is determined by the ionization constant of the acid and the pH of the medium according to the familiar Henderson-Hasselbach equation.[3] The two forms of the dye have different optical absorption spectra. Hence, the relative concentration of one form, which varies as a function of pH, can be measured optically and related to variations in pH. In the pH sensor, green and red lights that emerge from the distal end of one fiber pass through the dye, where it is backscattered into the other fiber by the light-scattering beads. The base form of the indicator absorbs the green light. The red light is not absorbed by the indicator and is used as an optical reference. The ratio of green to red light is related to the pH of the medium.

A similar principle can also be used with a reversible fluorescent indicator where the concentration of one indicator form is measured by its fluorescence rather than by the absorbance intensity. Light, typically in the blue or UV wavelength region, excites the fluorescent dye to emit longer-wavelength light. The concept is based on the fluorescence of weak acid dyes that have different excitation wavelengths for the basic and acidic forms but the same emitted fluorescent wavelength. The dye is encapsulated in a sample chamber that is permeable to hydrogen ions. When the dye is illuminated with the two different excitation wavelengths, the ratio of the emitted fluorescent intensities can be used to calculate the pH of the solution that is in contact with the encapsulated dye.

Fiber Optic pCO_2 Sensors

The pCO_2 of a sample is typically determined by measuring changes in the pH of a bicarbonate solution that is isolated from the sample by a CO_2-permeable membrane but remains in equilibrium with the CO_2 gas. The bicarbonate and CO_2, as carbonic acid, form a pH buffer system. By the Henderson-Hasselbach equation, the hydrogen ion concentration is proportional to the pCO_2 of the sample. This measurement can be done with either a pH electrode or a dye indicator.

Mixed Venous Oxygen Saturation Sensors

Fiber optic catheters can be used in vivo to measure mixed venous oxygen saturation (SvO_2) inside the pulmonary artery, which represents the blood outflow from all tissue

[3] $pH = 6.1 + \log \frac{HCO_3^-}{CO_2}$

beds. Under normal conditions, oxygen saturation in the pulmonary artery is normally around 75 percent, and oxygen consumption is less than or equal to the amount of oxygen delivered. However, in critically ill patients, oxygen delivery is often insufficient for the increased tissue demands because many such patients have compromised compensatory mechanisms. If tissue oxygen demands increase and the body's compensatory mechanisms are overwhelmed, the venous oxygen reserve will be tapped, and that change will be reflected as a decrease in SvO_2. For this reason, SvO_2 is regarded as a reliable indicator of tissue oxygenation and, therefore, can be used to indicate the effectiveness of the cardiopulmonary system during cardiac surgery and in the ICU.

The fiber optic SvO_2 catheter consists of two separate optical fibers; one fiber is used for transmitting incident light to the flowing blood, and the other directs the backscattered light to a photodetector. The catheter is introduced into the vena cava and further advanced through the heart into the pulmonary artery by inflating a small balloon located at the distal end. The flow-directed catheter also contains a small thermistor for measuring cardiac output by thermodilution.

The principle is based on the relationship between SvO_2 and the ratio of the infrared-to-red (IR/R) light backscattered from the red blood cell in blood

$$SvO_2 = A - B(\mathrm{IR/R}) \tag{10.27}$$

where, A and B are empirically derived calibration coefficients.

Several problems limit the wide clinical application of intravascular fiber optic oximeters. These include the dependence of the optical readings on motion artifacts due to catheter tip "whipping" against the blood vessel wall. Additionally, the introduction of the catheter into the heart requires an invasive procedure and can sometimes cause arrhythmias.

10.6.4 Intravascular Fiber Optic Pressure Sensors

Pressure measurements provide important diagnostic information. For example, pressure measurements inside the heart, cranium, kidneys, and bladder can be used to diagnose abnormal physiological conditions that are otherwise not feasible to ascertain from imaging or other diagnostic modalities. In addition, intracranial hypertension resulting from injury or other causes can be monitored to assess the need for therapy and its efficacy. Likewise, dynamic changes of pressure measured inside the heart, cranial cavities, uterus, and bladder can help to assess the efficiency of these organs during contractions.

Several approaches can be used to measure pressure using minimally invasive sensors. The most common technique involves the use of a fiber optic catheter. Fiber optic pressure sensors have been known and widely investigated since the early 1960s. The major challenge is to develop a small enough sensor with high sensitivity, high fidelity, and adequate dynamic response that can be inserted either through a hypodermic needle or in the form of a catheter. Additionally, for routine clinical use, the device must be cost-effective and disposable.

A variety of ideas have been exploited for varying a light signal in a fiber optic probe with pressure. Most designs utilize either an interferometer principle or measure changes in light intensity. Interferometric-based pressure sensors are known to have a high sensitivity, but they involve complex calibration and require complicated fabrication. On the other

FIGURE 10.40 Fiber optic in vivo pressure sensor. *Courtesy of Fiso Technologies, Quebec, Canada.*

hand, fiber optic pressure sensors based on light intensity modulation have a lower sensitivity but involve simpler construction.

The basic operating principle of a fiber optic pressure sensor is based on light intensity modulation. Typically, white light or light produced by a light emitting diode (LED) is carried by an optical fiber to a flexible mirrored surface located inside a pressure-sensing element. The mirror is part of a movable membrane partition that separates the fiber end from the fluid chamber. Changes in the hydrostatic fluid pressure cause a proportional displacement of the membrane relative to the distal end of the optical fiber. This in turn modulates the amount of light coupled back into the optical fiber. The reflected light is measured by a sensitive photodetector and converted to a pressure reading.

A fiber optic pressure transducer for in vivo application based on optical interferometry using white light was developed by Fiso Technologies (Figure 10.40). The sensing element is based on a Fabry-Pérot principle. A miniaturized Fabry-Pérot cavity is defined on one end by a micromachined silicon diaphragm membrane that acts as the pressure sensing element and is bonded on a cup-shaped glass base attached to the opposite side of the optical fiber. When external pressure is applied to the transducer, the deflection of the diaphragm causes variation of the cavity length that in turn is converted to a pressure reading. Due to its extremely small size (dia: 550 μm), the sensor can be inserted through a hypodermic needle.

10.6.5 Intravascular Fiber Optic Temperature Sensors

Miniature fiber optic temperature sensors, also commercialized by Fiso Technologies, are based on a similar Fabry-Pérot principle utilized in the construction of a miniaturized fiber optic pressure transducer. The Fabry-Pérot cavity in these designs is formed by two optical fibers assembled into a glass capillary tube or a transparent semiconductor material. The cavity length changes with temperature variations due to differences in the thermal expansion coefficient between the glass capillary and optical fibers. Due to their miniature construction (dia: 210–800 μm), the thermal inertia is close to zero, allowing ultrafast temperature response. The miniature size of the sensor allows the integration into minimally invasive medical devices for direct in situ measurement in space-restricted cavities.

10.6.6 Indicator-Mediated Fiber Optic Sensors

Since only a limited number of biochemical substances have an intrinsic optical absorption or fluorescence property that can be measured directly with sufficient selectivity by standard spectroscopic methods, indicator-mediated sensors have been developed to use specific reagents that are immobilized either on the surface or near the tip of an optical

fiber. In these sensors, light travels from a light source to the end of the optical fiber, where it interacts with a specific chemical or biological recognition element. These transducers may include indicators and ion-binding compounds (ionophores), as well as a wide variety of selective polymeric materials. After the light interacts with the biological sample, it returns either through the same optical fiber (in a single-fiber configuration) or a separate optical fiber (in a dual-fiber configuration) to a detector, which correlates the degree of light attenuation with the concentration of the analyte.

Typical indicator-mediated sensor configurations are shown schematically in Figure 10.41. The transducing element is a thin layer of chemical material that is placed near the sensor tip and is separated from the blood medium by a selective membrane. The chemical-sensing material transforms the incident light into a return light signal with a magnitude that is proportional to the concentration of the species to be measured. The stability of the sensor is determined by the stability of the photosensitive material that is used and also by how effective the sensing material is protected from leaching out of the probe. In Figure 10.41a, the indicator is immobilized directly on a membrane that is positioned at the end of the fiber. An indicator in the form of a powder can also be physically retained in position at the end of the fiber by a special permeable membrane, as shown in Figure 10.41b, or a hollow capillary tube, as shown in Figure 10.41c.

10.6.7 Immunoassay Sensors

The development of immunosensors is based on the observation of ligand-binding reaction products between a target analyte and a highly specific binding reagent. The key

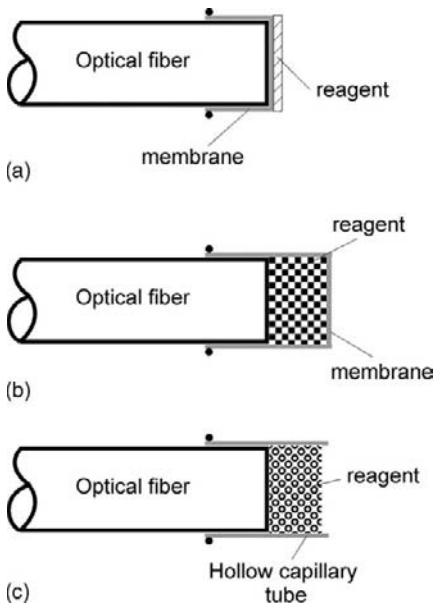


FIGURE 10.41 Different indicator-mediated fiber optic sensor configurations.

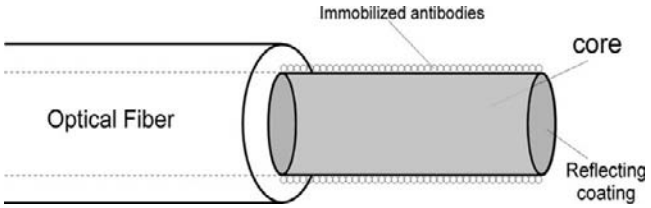


FIGURE 10.42 Principle of a fiber optic immunoassay biosensor.

component of an immunosensor is the biological recognition element, which typically consists of antibodies or antibody fragments. Immunological techniques offer outstanding selectivity[4] and sensitivity through the process of antibody-antigen interaction. This is the primary recognition mechanism by which the immune system detects and fights foreign matter and has therefore allowed the measurement of many important compounds at micromolar and even picomolar concentrations in complex biological samples.

Evanescent-type biosensors can be used in immunological diagnostics to detect antibody-antigen binding. Figure 10.42 shows a conceptual diagram of an immunoassay biosensor. The immobilized antibody on the surface of the unclad portion of the fiber captures the antigen from the sample solution, which is normally introduced into a small flow through a chamber where the fiber tip is located. The sample solution is then removed and a labeled antibody is added into the flow chamber. A fluorescent signal is excited and measured when the labeled antibody binds to the antigen that is already immobilized by the antibody.

10.6.8 Surface Plasmon Resonance Sensors

When monochromatic polarized light (e.g., from a laser source) impinges on a transparent medium having a conducting metallized surface (e.g., Ag or Au), there is a charge density oscillation at the interface. When light at an appropriate wavelength interacts with the dielectric-metal interface at a defined angle, called the resonance angle, there is a match of resonance between the energy of the photons and the electrons at the metal interface. As a result, the photon energy is transferred to the surface of the metal as packets of electrons, called plasmons, and the light reflection from the metal layer will be attenuated. This results in a phenomenon known as surface plasmon resonance (SPR) and is shown schematically in Figure 10.43. The resonance is observed as a sharp dip in the reflected light intensity when the incident angle is varied. The resonance angle depends on the incident wavelength, the type of metal, the polarization state of the incident light, and the nature of the medium in contact with the surface. Any change in the refractive index of the medium will produce a shift in the resonance angle and thus provide a highly sensitive means of monitoring surface interactions.

[4]The sensor's ability to detect a specific substance in a mixture containing other substances.

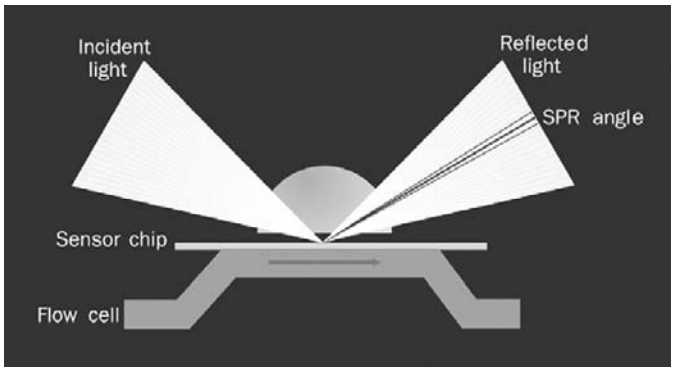


FIGURE 10.43 Principle of a surface plasmon resonance (SPR) detection system. *Courtesy of Biacore AB, Uppsala, Sweden.*

SPR is generally used for sensitive measurement of variations in the refractive index of the medium immediately surrounding the metal film. For example, if an antibody is bound to or absorbed into the metal surface, a noticeable change in the resonance angle can be readily observed because of the change of the refraction index at the surface if all other parameters are kept constant. The advantage of this concept is the improved ability to detect the direct interaction between antibody and antigen as an interfacial measurement.

10.7 EXERCISES

1. Give an example of a biomedical transducer that is used to monitor patients in the intensive care unit.
2. Discuss the important considerations in the selection of materials for packaging of an implantable biosensor.
3. Estimate the response time of an airflow transducer to monitor changes in breathing rate.
4. Explain why low drift is an important specification for implantable sensors.
5. The calibration tests of a new pressure transducer produced the readings in Table 10.4.
   (a) Plot the input-output calibration for this transducer.
   (b) Find the offset for readings between 0 to 200 mmHg.
   (c) Find the sensitivity for readings between 0 to 200 mmHg.
   (d) Estimate the average sensitivity for readings ranging between 200 to 300 mmHg.
   (e) State whether the response of this transducer over the entire measurement range is linear or nonlinear.
6. Suggest a method to measure hysteresis in a blood flow transducer.
7. Discuss the problem of using a pressure transducer with hysteresis to monitor blood pressure.
8. Explain how the accuracy of a new temperature sensor can be determined.
9. Two identical silver electrodes are placed in an electrolyte solution. Calculate the potential drop between the two electrodes.

TABLE 10.4 Sample Calibration Data for a Pressure Sensor

Pressure (mmHg)	Reading (μV)
20	0
40	20
60	40
80	60
100	80
120	100
140	120
160	135
180	150
200	165
220	180
240	190
260	200
280	210
300	220
320	225
340	230
360	235
380	237
400	239
420	240
440	240

10. Cadmium and zinc electrodes are placed in an electrolyte solution. Calculate the current that will flow through the electrodes if the equivalent resistance of the solution is equal to 14 kΩ.

11. Explain what will happen when the Ag/AgCl gel of an ECG electrode used to monitor a patient in the ICU dries out over time.

12. By how much would the inductance of an inductive displacement transducer coil change if the number of coil turns is decreased by a factor of 6?

Continued

13. Determine the ratio between the cross-sectional areas of two blood vessels assuming that the voltage ratio induced in identical magnetic flow probes is equal to 2:3 and the ratio of blood flows through these vessels is 1:5.
14. A 4.5 kΩ linear rotary transducer is used to measure the angular displacement of the knee joint. Calculate the change in output voltage for a 165° change in the angle of the knee. Assume that a constant current of 14 mA is supplied to the transducer.
15. Provide a step-by-step derivation of Eq. (10.11).
16. An elastic resistive transducer with an initial resistance, R_o, and length, l_o, is stretched to a new length. Assuming that the cross-sectional area of the transducer changes during stretching, derive a mathematical relationship for the change in resistance ΔR as a function of the initial length, l_o; the change in length, Δl; the volume of the transducer, V; and the resistivity, ρ.
17. The area of each plate in a differential capacitor sensor is equal to 5.6 cm^2. Calculate the equilibrium capacitance in air for each capacitor assuming that the equilibrium displacement for each capacitor is equal to 3 mm.
18. Plot the capacitance (y-axis) versus displacement (x-axis) characteristics of a capacitance transducer.
19. Calculate the sensitivity of a capacitive transducer (i.e., $\Delta C/\Delta d$) for small changes in displacements.
20. A capacitive transducer is used in a mattress to measure changes in breathing patterns of an infant. During inspiration and expiration, the rate of change (i.e., dV/dt) in voltage across the capacitor is equal to ±1V/s, and this change can be modeled by a triangular waveform. Plot the corresponding changes in current flow through this transducer.
21. Derive the relationship for the current through the capacitor-equivalent piezoelectric crystal as a function of V and C.
22. Two identical ultrasonic transducers are positioned across a blood vessel, as shown in Figure 10.44. Calculate the diameter of the blood vessel if it takes 380 ns for the ultrasonic sound wave to propagate from one transducer to the other.
23. Discuss the advantages of MEMS-type sensors.

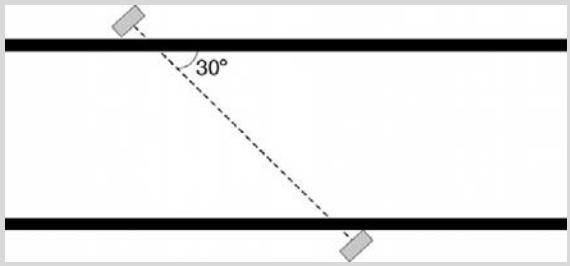


FIGURE 10.44 Two identical ultrasonic transducers positioned across a blood vessel.

24. Calculate the resistance of a thermistor at 98°F assuming that the resistance of this thermistor at 12°C is equal to 7.0 kΩ and $\beta = 4{,}600$.

25. The resistance of a thermistor with a $\beta = 5{,}500$ measured at 18°C is equal to 250 Ω. Find the temperature of the thermistor when the resistance is doubled.

26. Calculate the β of a thermistor assuming that it has a resistance of 4.4 kΩ at 21°C (room temperature) and a resistance of 2.85 kΩ when the room temperature increases by 20 percent.

27. A Chromel/Constantan thermocouple has the following empirical coefficients:

$C_0 = -2.340 \times 10^{-2}$

$C_1 = 4.221 \times 10^{-2}$

$C_2 = 3.284 \times 10^{-5}$

Find the EMF generated by this thermocouple at a temperature of 250°C.

28. Find the Seebeck coefficient for the Chromel/Alumel thermocouple at a temperature of 200°C.

29. Explain why the temporal artery thermometer is not used to measure core body temperature over the radial artery.

30. Compare and contrast the temporal artery and tympanic thermometers.

31. Compare and contrast a temperature pill with a temporal artery thermometer.

32. Sketch the current (*y*-axis) versus pO_2 (*x*-axis) characteristics of a typical polarographic Clark electrode.

33. Explain why the value of the normalized ratio (*R*) in a pulse oximeter is independent of the volume of arterial blood entering the tissue during systole.

34. Explain why the value of the normalized ratio (*R*) in a pulse oximeter is independent of skin pigmentation.

35. Explain the difference between a potentiometric and amperometric sensor.

36. Explain the difference between intravascular fiber optic pO_2 and SvO_2 sensors.

37. A pH electrode is attached to a sensitive voltmeter that reads 0.652 V when the electrode is immersed in a buffer solution with a pH of 6.7. After the pH electrode is moved to an unknown buffer solution, the reading of the voltmeter is decreased by 20 percent. Calculate the pH of the unknown buffer solution.

38. Plot the optical density, OD, of an absorbing solution (*y*-axis) as a function of the concentration of this solution (*x*-axis). What is the slope of this curve?

39. An unknown sample solution whose concentration is 1.55×10^{-3} g/L is placed in a 1 cm clear holder and found to have a transmittance of 44 percent. The concentration of this sample is changed such that its transmittance has increased to 57 percent. Calculate the new concentration of the sample.

40. Calculate the angle of the refracted light ray if an incident light ray passing from air into water has a 75-degree angle with respect to the normal.

41. Explain why fiber optic sensors typically require simultaneous measurements using two wavelengths of light.

42. Plot the fluorescence intensity of a fiber optic pO_2 sensor (*y*-axis) as a function of oxygen concentration (*x*-axis).

Continued

43. A chemical sensor is used to measure the pH of a dye with an absorbance spectrum shown in Figure 10.45. Assume that the absorbance of each form of the dye is linearly related to its pH. Devise a method to measure the pH of the dye.

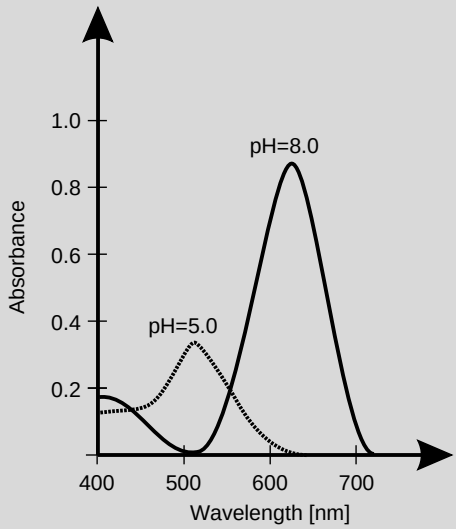


FIGURE 10.45 Optical absorbance spectra of a dye in its acid (pH = 5.0) and base (pH = 8.0) forms.

44. Explain the difference between absorption-based and fluorescence-based measurements.

Suggested Readings

J.A. Allocca, A. Stuart, Transducers: Theory and Applications, Reston Publishing, Reston, VA, 1984.
R. Aston, Principles of Biomedical Instrumentation and Measurement, Macmillan, New York, 1990.
D. Buerk, Biosensors: Theory and Applications, CRC Press, Boca Raton, FL, 1995.
R.S.C. Cobbold, Transducers for Biomedical Measurement: Principles and Applications, Wiley, New York, 1974.
L. Cromwell, F.J. Weibell, E.J. Pfeiffer, Biomedical Instrumentation and Measurements, Prentice Hall, Englewood Cliffs, NJ, 1980.
B. Eggins, Biosensors: An Introduction, Wiley, New York, 1997.
B.R. Eggins, Chemical Sensors and Biosensors for Medical and Biological Applications, John Wiley, New York, 2002.
L.A. Geddes, L.E. Baker, Principles of Applied Biomedical Instrumentation, third ed., Wiley-Interscience, New York, 1989.
E.A.H. Hall, Biosensors, Prentice Hall, Englewood Cliffs, NJ, 1991.
G. Harsanyi, Sensors in Biomedical Applications: Fundamental, Technology and Applications, CRC Press, Boca Raton, FL, 2000.
M.R. Neuman, Biomedical Sensors, in: J.D. Bronzino (Ed.), The Biomedical Engineering Handbook, second ed., CRC/IEEE Press, Boca Raton, FL, 1999.
T. Togawa, T. Tamura, P.A. Oberg, Biomedical Transducers and Instruments, CRC Press, Boca Raton, FL, 1997.
J.G. Webster, Encyclopedia of Medical Devices and Instrumentation, John Wiley, New York, 1988.
J.G. Webster, Medical Instrumentation: Application and Design, third ed., John Wiley, New York, 1998.
D.L. Wise, Bioinstrumentation and Biosensors, Marcel Dekker, New York, 1991.
J. Cooper, A.E.G. Cass, Biosensors, Oxford University Press, 2004.

CHAPTER

11

Biosignal Processing

Monty Escabí, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Describe the different origins and types of biosignals.
- Distinguish between deterministic, periodic, transient, and random signals.
- Explain the process of A/D conversion.
- Define the sampling theorem.
- Describe the main purposes and uses of the Fourier transforms.
- Define the Z-transform.
- Describe the basic properties of a linear system.
- Describe the concepts of filtering and signal averaging.
- Explain the basic concepts and advantages of fuzzy logic.
- Describe the basic concepts of artificial neural networks.

11.1 INTRODUCTION

Biological signals, or biosignals, are space, time, or space-time records of a biological event such as a beating heart or a contracting muscle. The electrical, chemical, and mechanical activity that occurs during this biological event often produces signals that can be measured and analyzed. Biosignals, therefore, contain useful information that can be used to understand the underlying physiological mechanisms of a specific biological event or system and that may be useful for medical diagnosis.

Biological signals can be acquired in a variety of ways—for example, by a physician who uses a stethoscope to listen to a patient's heart sounds or with the aid of technologically advanced biomedical instruments. Following data acquisition, biological signals are analyzed in order to retrieve useful information. Basic methods of signal analysis, such as amplification, filtering, digitization, processing, and storage, can be applied to many biological signals. These techniques are generally accomplished with simple electronic circuits or with digital computers. In addition to these common procedures, sophisticated digital processing methods are quite common and can significantly improve the quality of the retrieved data. These include signal averaging, wavelet analysis, and artificial intelligence techniques.

11.2 PHYSIOLOGICAL ORIGINS OF BIOSIGNALS

11.2.1 Bioelectric Signals

Nerve and muscle cells generate bioelectric signals that are the result of electrochemical changes within and between cells (see Chapter 5). If a nerve or muscle cell is stimulated by a stimulus that is strong enough to reach a necessary threshold, the cell will generate an action potential. The action potential, which represents a brief flow of ions across the cell membrane, can be measured with intracellular or extracellular electrodes. Action potentials generated by an excited cell can be transmitted from one cell to adjacent cells via its axon. When many cells become activated, an electric field is generated that propagates through the biological tissue. These changes in extracellular potential can be measured on the surface of the tissue or organism by using surface electrodes. The electrocardiogram (ECG), electrogastrogram (EGG), electroencephalogram (EEG), and electromyogram (EMG) are all examples of this phenomenon (Figure 11.1).

11.2.2 Biomagnetic Signals

Different organs, including the heart, brain, and lungs, also generate weak magnetic fields that can be measured with magnetic sensors. Typically, the strength of the magnetic field is much weaker than the corresponding physiological bioelectric signals. Biomagnetism is the measurement of the magnetic signals that are associated with specific physiological activity and that are typically linked to an accompanying electric field from a specific tissue or organ. With the aid of very precise magnetic sensors or SQUID (superconducting quantum interference device) magnetometers, it is possible to directly monitor

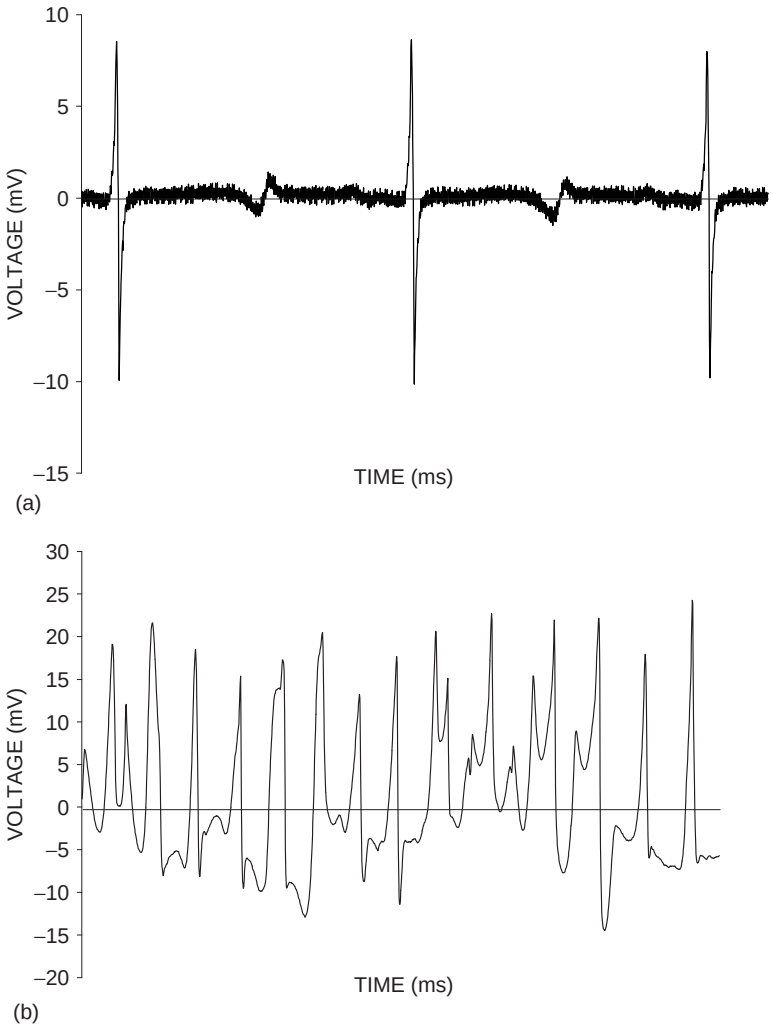


FIGURE 11.1 (a) Electrogram recorded from the surface of a pig's heart during normal sinus rhythm. (b) Electrogram recorded from the surface of the same pig's heart during ventricular fibrillation (VF). (Sampled at 1,000 samples/s.)

magnetic activity from the brain (magnetoencephalography, MEG), peripheral nerves (magnetoneurography, MNG), gastrointestinal tract (magnetogastrography, MGG), and the heart (magnetocardiography, MCG).

11.2.3 Biochemical Signals

Biochemical signals contain information about changes in concentration of various chemical agents in the body. The concentration of various ions, such as calcium and potassium, in cells can be measured and recorded. Changes in the partial pressures of oxygen (P_{O_2}) and

carbon dioxide (P_{CO_2}) in the respiratory system or blood are often measured to evaluate normal levels of blood oxygen concentration. All of these constitute biochemical signals. These biochemical signals can be used for a variety of purposes, such as determining levels of glucose, lactate, and metabolites and providing information about the function of various physiological systems.

11.2.4 Biomechanical Signals

Mechanical functions of biological systems, which include motion, displacement, tension, force, pressure, and flow, also produce measurable biological signals. Blood pressure, for example, is a measurement of the force that blood exerts against the walls of blood vessels. Changes in blood pressure can be recorded as a waveform (Figure 11.2). The upstrokes in the waveform represent the contraction of the ventricles of the heart as blood is ejected from the heart into the body and blood pressure increases to the systolic pressure, the maximum blood pressure (see Chapter 3). The downward portion of the waveform depicts ventricular relaxation as the blood pressure drops to the minimum value, better known as the diastolic pressure.

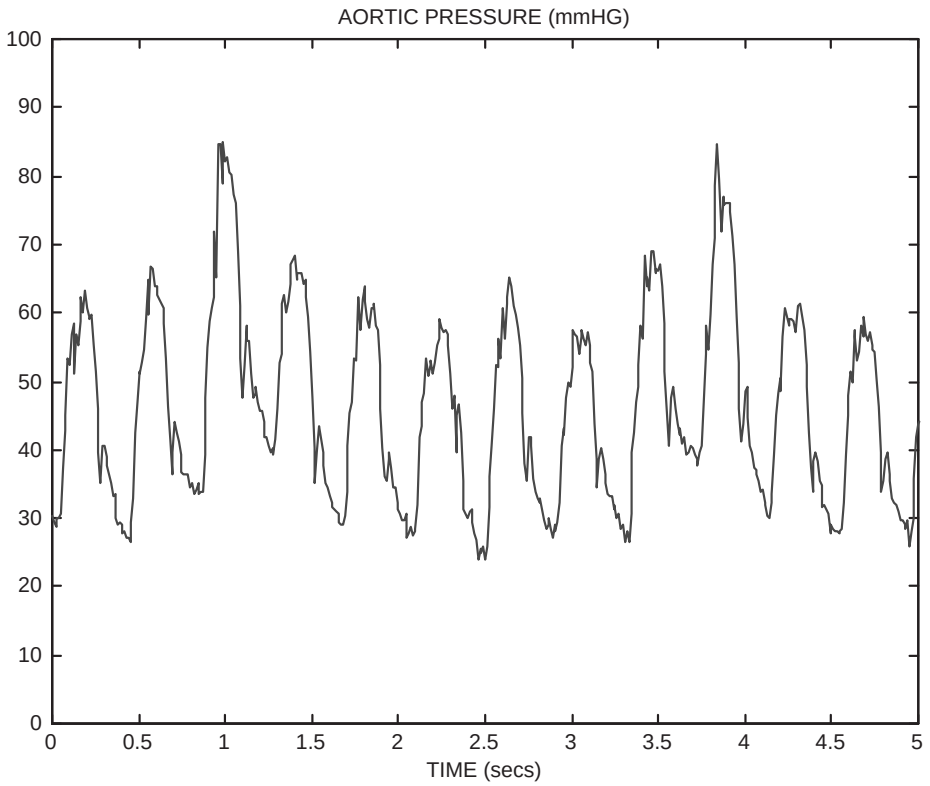


FIGURE 11.2 Blood pressure waveform recorded from the aortic arch of a 4-year-old child. (Sampled at 200 samples/s.)

11.2.5 Bioacoustic Signals

Bioacoustic signals are a special subset of biomechanical signals that involve vibrations (motion). Many biological events produce acoustic noise. For instance, the flow of blood through the valves in the heart has a distinctive sound. Measurements of the bioacoustic signal of a heart valve can be used to determine whether it is operating properly. The respiratory system, joints, and muscles also generate bioacoustic signals that propagate through the biological medium and can often be measured at the skin surface by using acoustic transducers such as microphones and accelerometers.

11.2.6 Biooptical Signals

Biooptical signals are generated by the optical or light induced attributes of biological systems. Biooptical signals can occur naturally, or in some cases, the signals may be introduced to measure a biological parameter with an external light medium. For example, information about the health of a fetus may be obtained by measuring the fluorescence characteristics of the amniotic fluid. Estimates of cardiac output can be made by using the dye dilution method that involves monitoring the concentration of a dye as it recirculates through the bloodstream. Finally, red and infrared light are used in various applications, such as to obtain precise measurements of blood oxygen levels by measuring the light absorption across the skin or a particular tissue.

EXAMPLE PROBLEM 11.1

What types of biosignals would the muscles in your lower legs produce if you were to sprint across a paved street?

Solution

Motion of the muscles and the external forces imposed as your feet hit the pavement produce biomechanical signals. Muscle stimulation by nerves and the contraction of muscle cells produce bioelectric signals. Metabolic processes in the muscle tissue could be measured as biochemical signals.

11.3 CHARACTERISTICS OF BIOSIGNALS

Biological signals can be classified according to various characteristics of the signal, including the waveform shape, statistical structure, and temporal properties. Two broad classes of signals that are commonly encountered include continuous and discrete signals. *Continuous* signals are defined over a continuum of time or space and are described by continuous variables. The notation $x(t)$ is used to represent a signal, x, that varies as a function of continuous time, t. Signals that are produced by biological phenomena are almost always continuous signals. Some examples include voltage measurements from the heart (see Figure 11.1), arterial blood pressure measurements (see Figure 11.2), and measurements of electrical activity from the brain.

Discrete signals are also commonly encountered in today's clinical setting. Unlike continuous signals, which are defined along a continuum of points in space or time, discrete signals are defined only at a subset of regularly spaced points in time and/or space. Discrete signals are therefore represented by arrays or sequences of numbers. The notation, $x(n)$, is used to represent a discrete sequence, x, that exists only at a subset of points in discrete time, n. Here, $n = 0, 1, 2, 3 \ldots$ is always an integer that represents the nth element of the discrete sequence. Although most biological signals are not discrete per se, discrete signals play an important role due to today's advancements in digital technology. Sophisticated medical instruments are commonly used to convert continuous signals from the human body to discrete digital sequences (see Chapter 7) that can be analyzed and interpreted with a computer. Computer axial tomography (CAT) scans, for instance, take digital samples from continuous x-ray images of a patient that are obtained from different perspective angles (see Chapter 15). These digitized or discrete image slices are then digitally enhanced, manipulated, and processed to generate a full three-dimensional computer model of a patient's internal organs. Such technologies are indispensable tools for clinical diagnosis.

Biological signals can also be classified as being either *deterministic* or *random*. Deterministic signals can be described by mathematical functions or rules. *Periodic* and *transient* signals make up a subset of all deterministic signals. Periodic signals are usually composed of the sum of different sine waves or sinusoid components and can be expressed as

$$x(t) = x(t + kT) \tag{11.1}$$

where $x(t)$ is the signal, k is an integer, and T is the period. The period represents the distance along the time axis between successive copies of the periodic signal. Periodic signals have a basic waveshape with a duration of T units that repeats indefinitely. Transient signals are nonzero or vary only over a finite time interval and subsequently decay to a constant value as time progresses. The sine wave, shown in Figure 11.3a, is a simple example of a periodic signal, since it repeats indefinitely with a repetition interval of 1 second. The product of a decaying exponential and a sine wave, as shown in Figure 11.3b, is a transient signal, since the signal amplitude approaches zero as time progresses.

Real biological signals almost always have some unpredictable noise or change in parameters and, therefore, are not entirely deterministic. The ECG of a normal beating heart at rest is an example of a signal that appears to be almost periodic but has a subtle unpredictable component. The basic waveshape consists of the P wave, QRS complex, and T wave and repeats (see Figure 3.22). However, the precise shapes of the P waves, QRS complexes, and T are somewhat irregular from one heartbeat to another. The length of time between QRS complexes, which is known as the R-R interval, also changes over time as a result of heart rate variability (HRV). HRV is used as a diagnostic tool to predict the health of a heart that has experienced a heart attack. The extended outlook for patients with low HRV is generally worse than it is for patients with high HRV.

Random signals, also called *stochastic* signals, contain uncertainty in the parameters that describe them. Because of this uncertainty, mathematical functions cannot be used to precisely describe random signals. Instead, random signals are most often analyzed using statistical techniques that require the treatment of the random parameters of the signal with probability distributions or simple statistical measures such as the mean and standard deviation. The electromyogram (EMG), an electrical recording of electrical activity in skeletal muscle that is

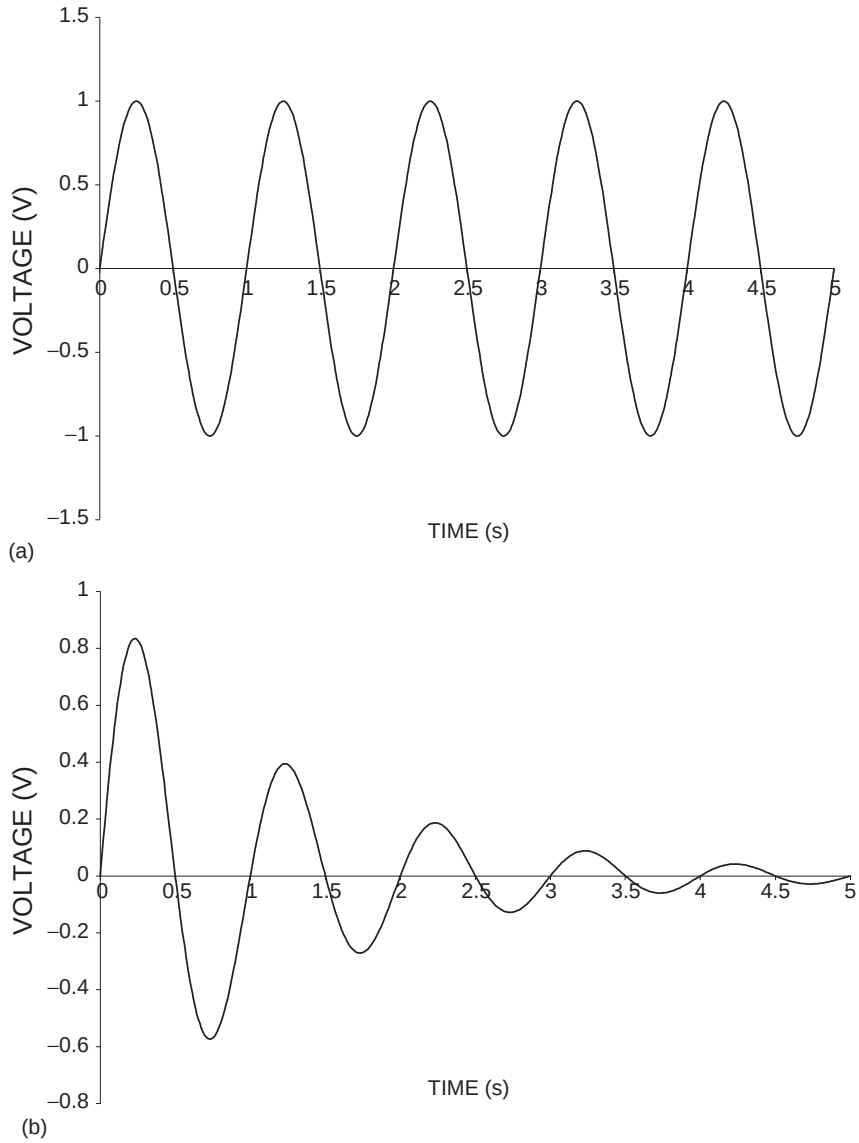


FIGURE 11.3 (a) Periodic sine wave signal $x(t) = \sin(\omega t)$ with period of 1 Hz. (b) Transient signal $y(t) = e^{-0.75t}\sin(\omega t)$ for the same 1 Hz sine wave.

used for the diagnosis of neuromuscular disorders, is a random signal. Stationary random signals have statistical properties, such as a mean and variance, that remain constant over time. Conversely, nonstationary random signals have statistical properties that vary with time. In many instances, the identification of stationary segments of random signals is important for proper signal processing, pattern analysis, and clinical diagnosis.

EXAMPLE PROBLEM 11.2

Ventricular fibrillation (VF) is a cardiac arrhythmia in which there are no regular QRS complexes, T waves, or rhythmic contractions of the heart muscle (see Figure 11.1b). VF often leads to sudden cardiac death, which is one of the leading causes of death in the United States. What type of biosignal would most probably be recorded by an ECG when a heart goes into VF?

Solution

An ECG recording of a heart in ventricular fibrillation will be a random, continuous, bioelectric signal.

11.4 SIGNAL ACQUISITION

11.4.1 Overview of Biosignal Data Acquisition

Biological signals are often very small and typically contain unwanted interference or noise. Such interference has the detrimental effect of obscuring relevant information that may be available in the measured signal. Noise can be extraneous in nature, arising from sources outside the body, such as thermal noise in sensors or 60-cycle noise in the electronic components of the acquisition system. Noise can also be intrinsic to the biological media, meaning it can arise from adjacent tissues or organs. ECG measurements from the heart, for instance, can be affected by bioelectric activity from adjacent muscles.

In order to extract meaningful information from biological signals sophisticated data acquisition techniques and equipment are commonly used. High-precision low-noise equipment is often necessary to minimize the effects of unwanted noise. Figure 11.4 shows the basic components in a bioinstrumentation system.

Throughout the data acquisition procedure, it is critical that the information and structure of the original biological signal of interest be faithfully preserved. Since these signals are often used to aid the diagnosis of pathological disorders, the procedures of amplification, analog filtering, and A/D conversion should not generate misleading or untraceable distortions. Distortions in a signal measurement could lead to an improper diagnosis.

11.4.2 Sensors, Amplifiers, and Analog Filters

Signals are first detected in the biological medium, such as a cell or on the skin's surface, by using a sensor (see Chapter 6). A sensor converts a physical measurand into an electric output and provides an interface between biological systems and electrical recording instruments. The type of biosignal determines what type of sensor will be used. ECGs, for example, are measured with electrodes that have a silver-silver chloride (Ag-AgCl) interface attached to the body that detects the movement of ions. Arterial blood pressure is measured with a sensor that detects changes in pressure. It is very important that the sensor used to detect the biological signal of interest does not adversely affect the properties and characteristics of the signal it is measuring.

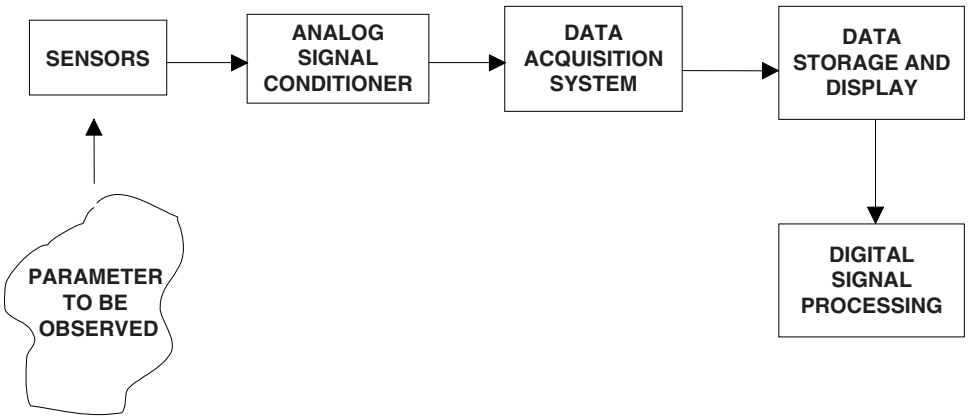


FIGURE 11.4 Sensors adapt the signal that is being observed into an electrical analog signal that can be measured with a data acquisition system. The data acquisition system converts the analog signal into a calibrated digital signal that can be stored. Digital signal processing techniques are applied to the stored signal to reduce noise and extract additional information that can improve understanding of the physiological meaning of the original parameter.

After the biosignal has been detected with an appropriate sensor, it is usually amplified and filtered. Operational amplifiers are electronic circuits that are used primarily to increase the amplitude or size of a biosignal. Bioelectric signals, for instance, are often faint and require up to a thousand-fold boosting of their amplitude with such amplifiers. An analog filter may then be used to remove noise or to compensate for distortions caused by the sensor. Amplification and filtering of the biosignal may also be necessary to meet the hardware specifications of the data acquisition system. Continuous signals may need to be limited to a certain band of frequencies before the signal can be digitized with an analog-to-digital converter, prior to storing in a digital computer.

11.4.3 A/D Conversion

Analog-to-digital (A/D) converters are used to transform biological signals from continuous analog waveforms to digital sequences. An A/D converter is a computer-controlled voltmeter, which measures an input analog signal and gives a numeric representation of the signal as its output. Figure 11.5a shows an analog signal, and Figure 11.5b shows a digital version of the same signal. The analog waveform, originally detected by the sensor and subsequently amplified and filtered, is a continuous signal. The A/D converter transforms the continuous, analog signal into a discrete, digital signal. The discrete signal consists of a sequence of numbers that can easily be stored and processed on a digital computer. A/D conversion is particularly important because storage and analysis of biosignals are becoming increasingly computer based.

The digital conversion of an analog biological signal does not produce an exact replica of the original signal. The discrete, digital signal is a digital approximation of the original, analog signal that is generated by repeatedly sampling the amplitude level of the original signal at fixed time intervals. As a result, the original, analog signal is represented as a sequence of numbers: the digital signal.

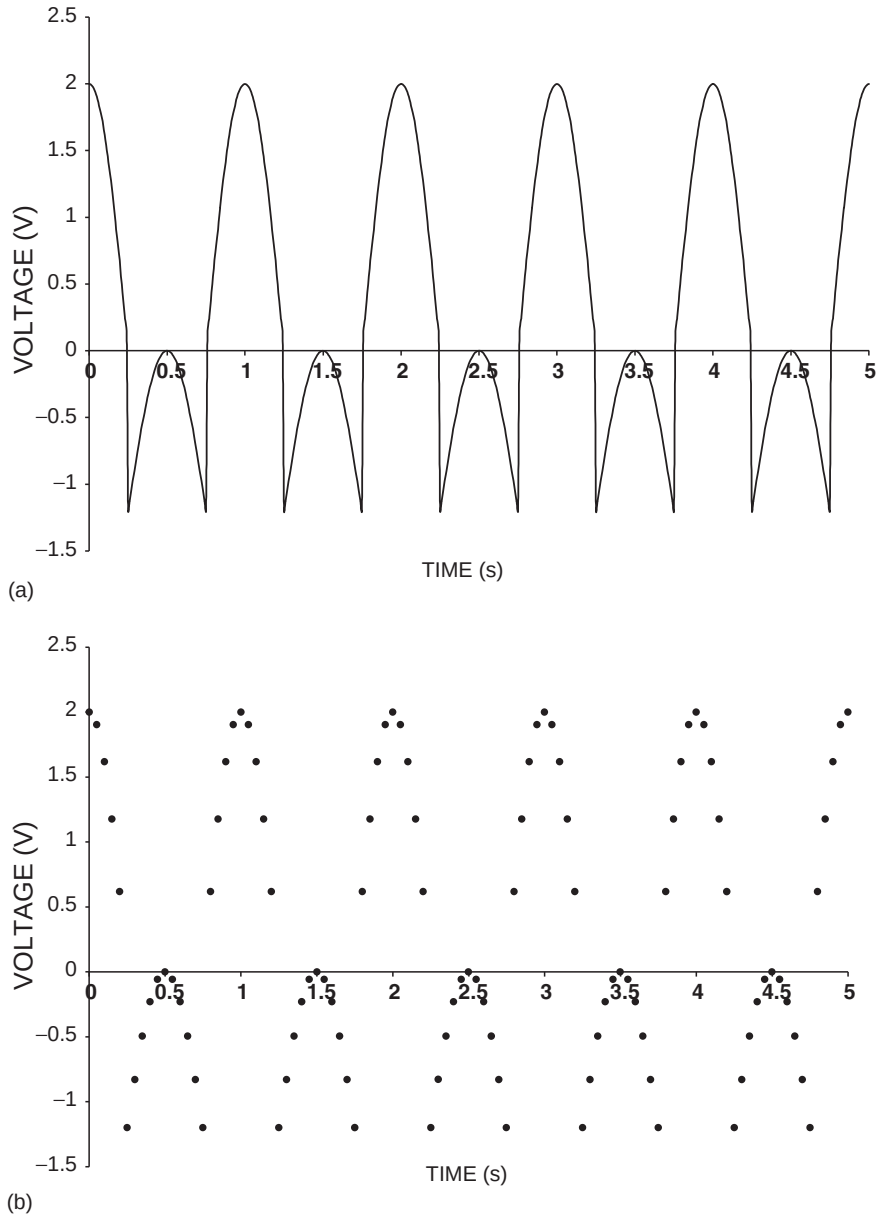


FIGURE 11.5 (a) Analog version of a periodic signal. (b) Digital version of the analog signal.

The two main processes involved in A/D conversion are *sampling* and *quantization*. Sampling is the process by which a continuous signal is first converted into a discrete sequence in time. If $x(t)$ is an analog signal, sampling involves recording the amplitude value of $x(t)$ every T seconds. The amplitude value is denoted as $x(kT)$ where $k = 0, 1, 2, 3, \ldots$ is an integer that denotes the position or the sample number from the sample set or data sequence.

T represents the sampling interval or the time between adjacent samples. In real applications, finite data sequences are generally used in digital signal processing. Therefore, the range of a data points is $k = 0, 1, \ldots N\text{-}1$, where N is the total number of discrete samples. The sampling frequency, f_s, or the sampling rate, is equal to the inverse of the sampling period, $1/T$, and is measured in units of Hertz (s^{-1}).

The following are digital sequences that are of particular importance:

The unit-sample of impulse sequence:

$$\delta(k) = \begin{array}{l} 1 \text{ if } k = 0 \\ 0 \text{ if } k \neq 0 \end{array}$$

The unit-step sequence:

$$u(k) = \begin{array}{l} 1 \text{ if } k > 0 \\ 0 \text{ if } k < 0 \end{array}$$

The exponential sequence:

$$a^k u(k) = \begin{array}{l} a^k \text{ if } k > 0 \\ 0 \text{ if } k < 0 \end{array}$$

The sampling rate used to discretize a continuous signal is critical for the generation of an accurate digital approximation. If the sampling rate is too low, distortions will occur in the digital signal. Nyquist's theorem states that the minimum sampling rate used, f_s, should be at least twice the maximum frequency of the original signal in order to preserve all of the information of the analog signal. The Nyquist rate is calculated as

$$f_{nyquist} = 2 \cdot f_{max} \tag{11.2}$$

where f_{max} is the highest frequency present in the analog signal. The Nyquist theorem therefore states that f_s must be greater than or equal to $2 \cdot f_{max}$ in order to fully represent the analog signal by a digital sequence. Practically, sampling is usually done at five to ten times the highest frequency, f_{Max}.

The second step in the A/D conversion process involves signal quantization. Quantization is the process by which the continuous amplitudes of the discrete signal are digitized by a computer. In theory, the amplitudes of a continuous signal can be any of an infinite number of possibilities. This makes it impossible to store all the values, given the limited memory in computer chips. Quantization overcomes this by reducing the number of available amplitudes to a finite number of possibilities that the computer can handle.

Since digitized samples are usually stored and analyzed as binary numbers on computers, every sample generated by the sampling process must be quantized. During quantization, the series of samples from the discretized sequence are transformed into binary numbers. The resolution of the A/D converter determines the number of bits that are available for storage. Typically, most A/D converters approximate the discrete samples with 8, 12, or 16 bits. If the number of bits is not sufficiently large, significant errors may be incurred in the digital approximation.

A/D converters are characterized by the number of bits that they use to generate the numbers of the digital approximation. A quantizer with N bits is capable of representing a total of 2^N possible amplitude values. Therefore, the resolution of an A/D converter

increases as the number of bits increases. A 16-bit A/D converter has better resolution than an 8-bit A/D converter, since it is capable of representing a total of 65,536 amplitude levels, compared to 256 for the 8-bit converter. The resolution of an A/D converter is determined by the voltage range of the input analog signal divided by the numeric range (the possible number of amplitude values) of the A/D converter.

EXAMPLE PROBLEM 11.3

Find the resolution of an 8-bit A/D converter when an input signal with a 10 V range is digitized.

Solution

$$\frac{\textit{input voltage range}}{2^N} = \frac{10\ V}{256} = 0.0391\ V/bit = 39.1\ mV/bit$$

EXAMPLE PROBLEM 11.4

The frequency content of an analog EEG signal is 0.5–100 Hz. What is the lowest rate at which the signal can be sampled to produce an accurate digital signal?

Solution

Highest frequency in analog signal = 100 Hz.
$f_{nyquist} = 2 \cdot f_{max} = 2 \cdot 100\ Hz = 200$ samples/second.

Another problem often encountered is determining what happens if a signal is not sampled at a rate high enough to produce an accurate representation of the signal. A direct result of the sampling theorem is that all frequencies of the form $[f - kf_s]$, where $-\infty \leq k \leq \infty$ and $f_s = 1/T$, look the same once they are sampled.

EXAMPLE PROBLEM 11.5

A 360 Hz signal is sampled at 200 samples/second. What frequency will the "aliased" digital signal contain?

Solution

According to the preceding formula, $f_s = 200$, and the pertinent set of frequencies that look alike is in the form of $[360 - k\ 200] = [\ldots\ 360\ 160\ -40\ -240\ \ldots.]$. The only signal in this group that will be accurately sampled is 40 Hz, since the sampling rate is more than twice this value. Note that for real signals −40 Hz and +40 Hz are equivalent—that is, $\cos(-\omega t) = \cos(\omega t)$ and $\sin(-\omega t) = -\sin(\omega t)$. Thus, the sampled signal will exhibit a period of 40 Hz. The process is shown in Figure 11.6.

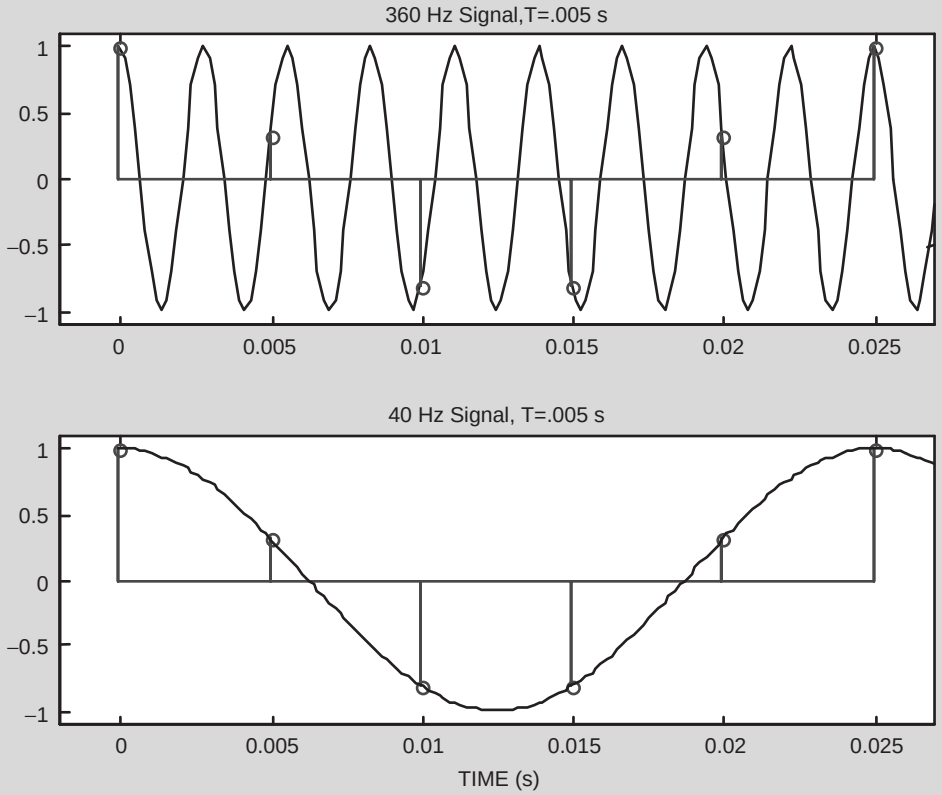


FIGURE 11.6 A 360 Hz sine wave is sampled every 5 ms—that is, at 200 samples/s. This sampling rate will adequately sample a 40 Hz sine wave but not a 360 Hz sine wave.

11.5 FREQUENCY DOMAIN REPRESENTATION OF BIOLOGICAL SIGNALS

In the early nineteenth century, Joseph Fourier laid out one of the most important theories on the field of function approximation. At the time, his result was applied toward the problem of heat transfer in solids, but it has since gained a much broader appeal. Today, Fourier's findings provide a general theory for approximating complex waveforms with simpler functions that has numerous applications in mathematics, physics, and engineering. This section summarizes the Fourier transform and variants of this technique that play an important role in the analysis and interpretation of biological signals.

11.5.1 Periodic Signal Representation: The Trigonometric Fourier Series

As an artist mixes oil paints on a canvas, a scenic landscape is meticulously recreated by combining various colors on a palate. It is well known that all shades of the color spectrum

can be recreated by simply mixing primary colors (red, green, and blue, or RGB) in the correct proportions. Television and computer displays often transmit signals as RGB, and these signals are collated together to create colors much as a master painter would on a canvas. In fact, the human visual system takes exactly the opposite approach. The retina decomposes images and scenery from the outside world into purely red-green-blue signals that are independently analyzed and processed by our brains. Despite this, we perceive a multitude of colors and shades.

This simple color analogy is at the heart of Fourier's theory, which states that a complex waveform can be approximated to any degree of accuracy with simpler functions. In 1807, Fourier showed that an arbitrary periodic signal of period, T, can be represented mathematically as a sum of trigonometric functions. Conceptually, this is achieved by summing or mixing sinusoids while simultaneously adjusting their amplitudes and frequency, as shown for a square wave function in Figure 11.7.

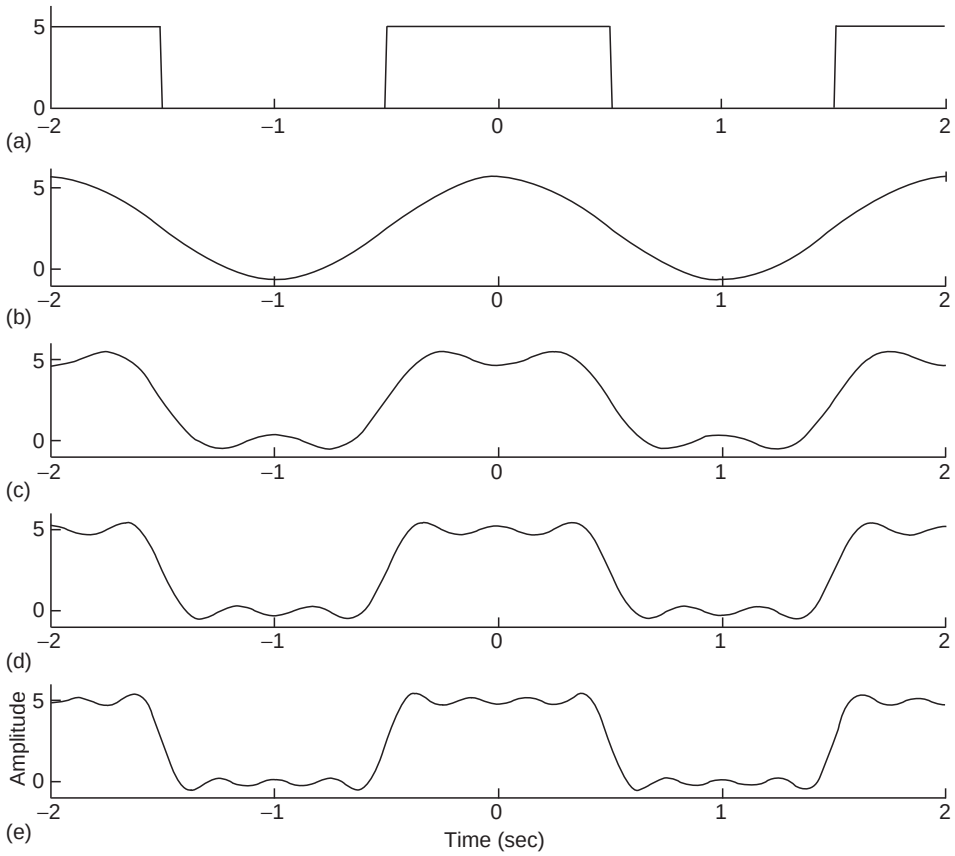


FIGURE 11.7 A square wave signal (a) is approximated by adding sinusoids (B–E). (b) 1 sinusoid, (c) 2 sinusoids, (d) 3 sinusoids, (e) 4 sinusoids. Increasing the number of sinusoids improves the quality of the approximation.

If the amplitudes and frequencies are chosen appropriately, the trigonometric signals add constructively, thus recreating an arbitrary periodic signal. This is akin to combining prime colors in precise ratios to recreate an arbitrary color and shade. RGB are the building blocks for more elaborate colors, much as sinusoids of different frequencies serve as the building blocks for more complex signals. All of these elements (the color and the required proportions; the frequencies and their amplitudes) have to be precisely adjusted to achieve a desired result. For example, a first-order approximation of the square wave is achieved by fitting the square wave to a single sinusoid of appropriate frequency and amplitude. Successive improvements in the approximation are obtained by adding higher-frequency sinusoid components, or *harmonics*, to the first-order approximation. If this procedure is repeated indefinitely, it is possible to approximate the square wave signal with infinite accuracy.

The Fourier series summarizes this result as

$$x(t) = a_0 + \sum_{m=1}^{\infty} (a_m \cos m\omega_o t + b_m \sin m\omega_o t) \tag{11.3a}$$

where $x(t)$ is the periodic signal to be approximated, $\omega_0 = 2\pi/T$ is the fundamental frequency of $x(t)$ in units of radians/s, and the coefficients a_m and b_m determine the amplitude of each cosine and sine term at a specified frequency $\omega_m = m\omega_0$. Equation (11.3a) tells us that the periodic signal, $x(t)$, is precisely replicated by summing an infinite number of sinusoids. The frequencies of the sinusoid functions always occur at integer multiples of ω_o and are referred to as "harmonics" of the fundamental frequency. If we know the coefficients a_m and b_m for each of the corresponding sine or cosine terms, we can completely recover the signal $x(t)$ by evaluating the Fourier series. How do we determine a_m and b_m for an arbitrary signal?

The coefficients of the Fourier series correspond to the amplitude of each sine and cosine. These are determined as

$$a_0 = \frac{1}{\mathrm{T}} \int_{\mathrm{T}} x(t)dt \tag{11.3b}$$

$$a_m = \frac{2}{\mathrm{T}} \int_{\mathrm{T}} x(t) \cos(m\omega_o t)dt \tag{11.3c}$$

$$b_m = \frac{2}{\mathrm{T}} \int_{\mathrm{T}} x(t) \sin(m\omega_o t)dt \tag{11.3d}$$

where the integrals are evaluated over a single period, T, of the waveform.

EXAMPLE PROBLEM 11.6

Find the trigonometric Fourier series of the square wave signal shown in Figure 11.7A, and implement the result in MATLAB for the first ten components. Plot the time waveform and the Fourier coefficients.

Continued

Solution

First note that

$$T = 2 \quad \text{and} \quad \omega_0 = \frac{2\pi}{T} = \pi$$

To simplify the analysis, integration for a_m and b_m is carried out over the first period of the waveform (from –1 to 1)

$$a_0 = \frac{1}{\mathrm{T}}\int_{-1}^{1} x(t)dt = \frac{1}{2}\int_{-1/2}^{1/2} 5dt = \frac{5}{2}$$

$$a_m = \frac{2}{\mathrm{T}}\int_{-1}^{1} x(t)\cos(m\omega_o t)dt = \int_{-1/2}^{1/2} 5 \cdot \cos(m\pi t)dt$$

$$= -5\frac{\sin(m\pi t)}{m\pi}\bigg|_{-1/2}^{1/2} = -5\frac{\sin(m\pi/2)}{m\pi/2} = 5 \cdot \mathrm{sinc}(m\pi/2)$$

$$b_m = \frac{2}{\mathrm{T}}\int_{-1}^{1} x(t)\sin(m\omega_o t)dt = \int_{-1/2}^{1/2} 5 \cdot \sin(m\pi t)dt = -5 \cdot \frac{\cos(m\pi t)}{m\pi}\bigg|_{-1/2}^{1/2} = 0$$

where by definition $\mathrm{sinc}(x) = \sin(x)/x$. Substituting the values for a_0, a_m, and b_m into Eq. (11.3a) gives

$$x(t) = \frac{5}{2} + 5 \cdot \sum_{m=1}^{\infty} \frac{\sin(m\pi/2)}{m\pi/2}\cos(m\pi t)$$

MATLAB implementation:

```
%Plotting Fourier Series Approximation
subplot(211)
time=-2:0.01:2; %Time Axis
x=5/2; %Initializing Signal
for m=1:10
  x=x+5*sin(m*pi/2)/m/pi*2*cos(m*pi*time);
end
plot(time,x,'k') %Plotting and Labels
xlabel('Time (sec)')
ylabel('Amplitude')
set(gca,'Xtick',[-2:2])
set(gca,'Ytick',[0 5])
set(gca,'Box','off')

%Plotting Fourier Magnitudes
subplot(212)
m=1:10;
```

```
Am=[5/2 5*sin(m*pi/2)./m/pi*2]; %Fourier Magnitudes
Faxis=(0:10)*.5; %Frequency Axis
plot(Faxis,Am,'k.') %Plotting
axis([0 5 -2 4])
set(gca,'Box','off')
xlabel('Frequency (Hz)')
ylabel('Fourier Amplitudes')
```

Note that the approximation of summing the first ten harmonics (Figure 11.8a) closely resembles the desired square wave. The Fourier coefficients, a_m, for the first ten harmonics are shown as a function of the harmonic frequency in Figure 11.8b. To fully replicate the sharp transitions of the square wave, an infinite number of harmonics are required.

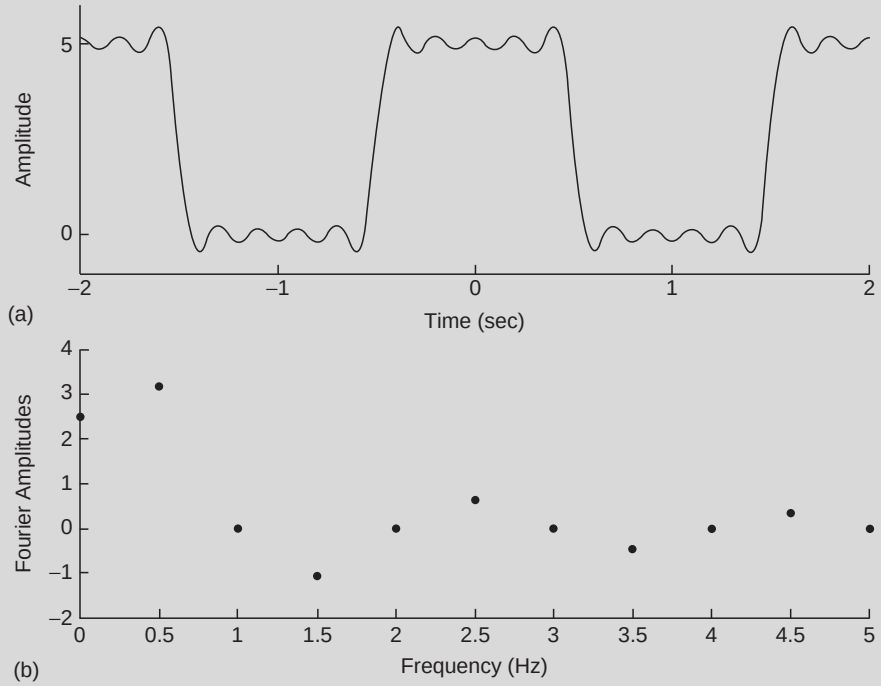


FIGURE 11.8 (a) MATLAB result showing the first ten terms of Fourier series approximation for the square wave. (b) The Fourier coefficients are shown as a function of the harmonic frequency.

11.5.2 Compact Fourier Series

The trigonometric Fourier series provides a direct approach for fitting and analyzing various types of biological signals, such as the repetitive beating of a heart or the cyclic oscillations produced by the vocal folds as one speaks. Despite its utility, alternate forms of the Fourier series are sometimes more appealing because they are easier to work with

and because signal measurements can often be interpreted more readily. The most widely used counterparts for approximating and modeling biological signals are the *exponential* and *compact* Fourier series.

The compact Fourier series is a close cousin of the standard Fourier series. This version of the Fourier series is obtained by noting that the sum of sinusoids and cosines can be rewritten by a single cosine term with the addition of a phase constant $a_m \cos m\omega_o t + b_m \sin m\omega_o t = A_m \cos(m\omega_o t + \phi_m)$, which leads to the compact form of the Fourier series:

$$x(t) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos(m\omega_o t + \phi_m). \tag{11.4a}$$

The amplitude for each cosine, A_m, is related to the Fourier coefficients through

$$A_m = \sqrt{a_m^2 + b_m^2} \tag{11.4b}$$

and the cosine phase is obtained from a_m and b_m as

$$\phi_m = \tan^{-1}\left(\frac{-b_m}{a_m}\right). \tag{11.4c}$$

EXAMPLE PROBLEM 11.7

Convert the standard Fourier series for the square pulse function of Example Problem 11.5 to compact form and implement in MATLAB.

Solution

We first need to determine the magnitude, A_m, and phase, ϕ_m, for the compact Fourier series. The magnitude is obtained as

$$A_m = \sqrt{a_m^2 + b_m^2} = \sqrt{(5 \cdot \operatorname{sinc}(m\pi/2))^2 + (0)^2} = 5\frac{|\sin(m\pi/2)|}{\pi m/2}$$

Since

$$|\sin(m\pi/2)| = \begin{cases} 1 & m = odd \\ 0 & m = even \end{cases}$$

we have

$$A_m = \begin{cases} 10/m\pi & m = odd \\ 0 & m = even \end{cases}.$$

Unlike a_m or b_m in the standard Fourier series, note that A_m is strictly a positive quantity for all m. The phase term is determined as

$$\phi_m = \tan^{-1}\left(\frac{-b_m}{a_m}\right) = \tan^{-1}\left(\frac{0}{5 \cdot \operatorname{sinc}(m\pi/2)}\right) = \begin{cases} 0 & \text{for } m = 0, 1, 4, 5, 8, 9 \ldots \\ \pi & \text{for } m = 2, 3, 6, 7, 10, 11 \ldots \end{cases}$$

Combining results

$$x(t) = \frac{5}{2} + \sum_{m=1}^{\infty} \frac{10}{m\pi} \cos(m\omega_o t + \phi_m)$$

where ϕ_m is as just defined. An interesting point regards the similarity of standard and compact versions of the Fourier series for this square wave example. In the standard form, the coefficient a_m alternates between positive and negative values, while for the compact form the Fourier coefficient, A_m, is identical in magnitude to a_m, but it is always a positive quantity. The sign (+ or –) of the standard Fourier coefficient is now consumed in the phase term, which alternates between 0 and π. This forces the cosine to alternate in its external sign because $-\cos(x) = \cos(x + \pi)$. The two equations are therefore mathematically identical, differing only in the way that the trigonometric functions are written out.

MATLAB implementation:

```
%Plotting Fourier Series Approximation
time=-2:0.01:2; %Time Axis
x=5/2; %Initializing Signal
m=1:10;
A=(10*sin(m*pi/2)./m/pi); %Fourier Coefficients
P=angle(A); %Phase Angle
A=abs(A); %Fourier Magnitude
for m=1:10
   x=x+A(m)*cos(m*pi*time+P(m));
end
subplot(211)
plot(time,x,'k') %Plotting and Labels
xlabel('Time (sec)')
ylabel('Amplitude')
set(gca,'Xtick',[-2:2])
set(gca,'Ytick',[0 5])
set(gca,'Box','off')

%Plotting Fourier Magnitudes
subplot(212)
m=1:10;
A=[5/2 A]; %Fourier Magnitudes
Faxis=(0:10)*.5; %Frequency Axis
plot(Faxis,A,'k.') %Plotting
axis([0 5 -2 4])
set(gca,'Box','off')
xlabel('Frequency (Hz)')
ylabel('Fourier Amplitudes')
```

The results are identical to those shown in Figures 11.8a and b.

11.5.3 Exponential Fourier Series

The main result from the Fourier series analysis is that an arbitrary periodic signal can approximate by summing individual cosine terms with specified amplitudes and phases. This result serves as much of the conceptual and theoretical framework for the field of signal analysis. In practice, the Fourier series is a useful tool for modeling various types of quasi-periodic signals.

An alternative and somewhat more convenient form of this result is obtained by noting that complex exponential functions are directly related to sinusoids and cosines through Euler's identities: $\cos(\theta) = \left(e^{j\theta} + e^{-j\theta}\right)/2$ and $\sin(\theta) = \left(e^{j\theta} - e^{-j\theta}\right)/2j$, where $j = \sqrt{-1}$. By applying Euler's identity to the compact trigonometric Fourier series, an arbitrary periodic signal can be expressed as a sum of complex exponential functions:

$$x(t) = \sum_{m=-\infty}^{+\infty} c_m e^{jk\omega_o t} \tag{11.5a}$$

This equation represents the exponential Fourier series of a periodic signal. The coefficients c_m are complex numbers that are related to the trigonometric Fourier coefficients

$$c_m = \frac{a_m - jb_m}{2} = \frac{A_m}{2} e^{j\phi_m} \tag{11.5b}$$

The proof for this result is beyond the scope of this text, but it is important to realize that the trigonometric and exponential Fourier series are intimately related, as can be seen by comparing their coefficients. The exponential coefficients can also be obtained directly by integrating $x(t)$,

$$c_m = \frac{1}{\mathrm{T}} \int_{\mathrm{T}} x(t) e^{-jm\omega_o t} dt \tag{11.5c}$$

over one cycle of the periodic signal. As for the trigonometric Fourier series, the exponential form allows us to approximate a periodic signal to any degree of accuracy by adding a sufficient number of complex exponential functions. A distinct advantage of the exponential Fourier series, however, is that it requires only a single integral (Eq. (11.5c)), compared to the trigonometric form, which requires three separate integrations.

EXAMPLE PROBLEM 11.8

Find the exponential Fourier series for the square wave of Figure 11.7a and implement in MATLAB for the first ten terms. Plot the time waveform and the Fourier series coefficients.

Solution

Like Example Problem 11.6, the Fourier coefficients are obtained by integrating from -1 to 1. Because a single cycle of the square wave signal has nonzero values between $-1/2$ and $+1/2$, the integral can be simplified by evaluating it between these limits:

$$c_m = \frac{1}{T}\int_T x(t)e^{-jm\omega_o t}dt = \frac{1}{2}\int_{-1/2}^{1/2} 5e^{-jm\pi t}dt = \frac{5}{2}\cdot\frac{e^{-jm\pi t}}{-jm\pi}\Bigg|_{-1/2}^{1/2}$$

$$= \frac{5}{2}\cdot\frac{e^{+jm\pi/2} - e^{-jm\pi/2}}{jm\pi} = \frac{5}{2}\cdot\frac{\sin(m\pi/2)}{m\pi/2}.$$

Therefore,

$$x(t) = \sum_{m=-\infty}^{+\infty} c_m e^{jk\omega_o t} = \sum_{m=-\infty}^{\infty}\frac{5}{2}\cdot\frac{\sin(m\pi/2)}{m\pi/2}\cdot e^{jm\pi t}$$

MATLAB implementation:

```
%Plotting Fourier Series Approximation
subplot(211)
time=-2:0.01:2; %Time Axis
x=0; %Initialize Signal
for m=-10:10
   if m==0
      x=x+5/2; %Term for m=0
   else
      x=x+5/2*sin(m*pi/2)/m/pi*2*exp(j*m*pi*time);
   end
end
plot(time,x,'k') %Plotting and Labels
xlabel('Time (sec)')
ylabel('Amplitude')
set(gca,'Xtick',[-2:2])
set(gca,'Ytick',[0 5])
set(gca,'Box','off')

%Plotting Fourier Magnitudes
subplot(212)
m=(-10:10)+1E-10;
A=[5/2*sin(m*pi/2)./m/pi*2]; %Fourier Magnitudes
Faxis=(-10:10)*.5; %Frequency Axis
plot(Faxis,A,'k.') %Plotting
axis([-5 5 -2 4])
set(gca,'Box','off')
xlabel('Frequency (Hz)')
ylabel('Fourier Amplitudes')
```

Note that we now require positive and negative frequencies in the approximation. Results showing the MATLAB output are shown in Figure 11.9.

Continued

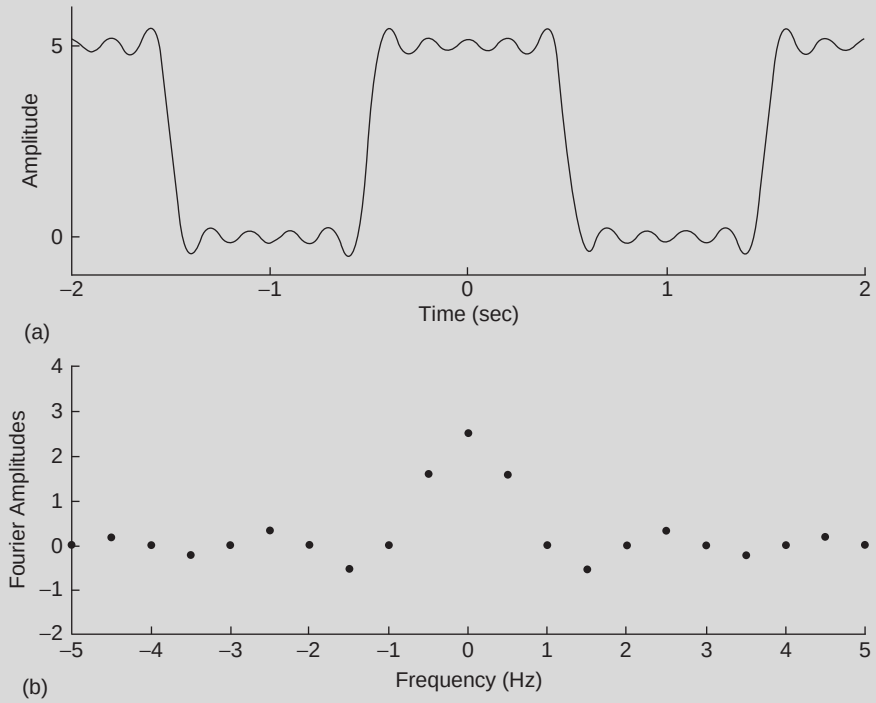


FIGURE 11.9 (a) MATLAB result showing the first ten terms of exponential Fourier series approximation for the square wave. (b) The compact Fourier coefficients are shown as a function of the harmonic frequency. Note that both negative and positive frequencies are now necessary to approximate the square wave signal.

In practice, many periodic or quasi-periodic biological signals can be accurately approximated with only a few harmonic components. Figures 11.10 and 11.11 illustrate a harmonic reconstruction of an aortic pressure waveform obtained by applying a Fourier series approximation. Figure 11.10 plots the coefficients for the cosine series representation as a function of the harmonic number. Note that the low-frequency coefficients are large in amplitude, whereas the high-frequency coefficients contain little energy and do not contribute substantially to the reconstruction. The amplitude coefficients, A_m, are plotted on a $\log_{10}$ scale so the smaller values are magnified and are therefore visible. Figure 11.11 shows several levels of harmonic reconstruction. The mean plus the first and second harmonics provide the basis for the general systolic and diastolic shape, since the amplitudes of these harmonics are large and contribute substantially to the reconstructed waveform. Additional harmonics add fine details but do not contribute significantly to the raw waveform.

11.5.4 Fourier Transform

In many instances, conceptualizing a signal in terms of its contributing cosine or sine functions has various advantages. The concept of frequency domain is an abstraction that

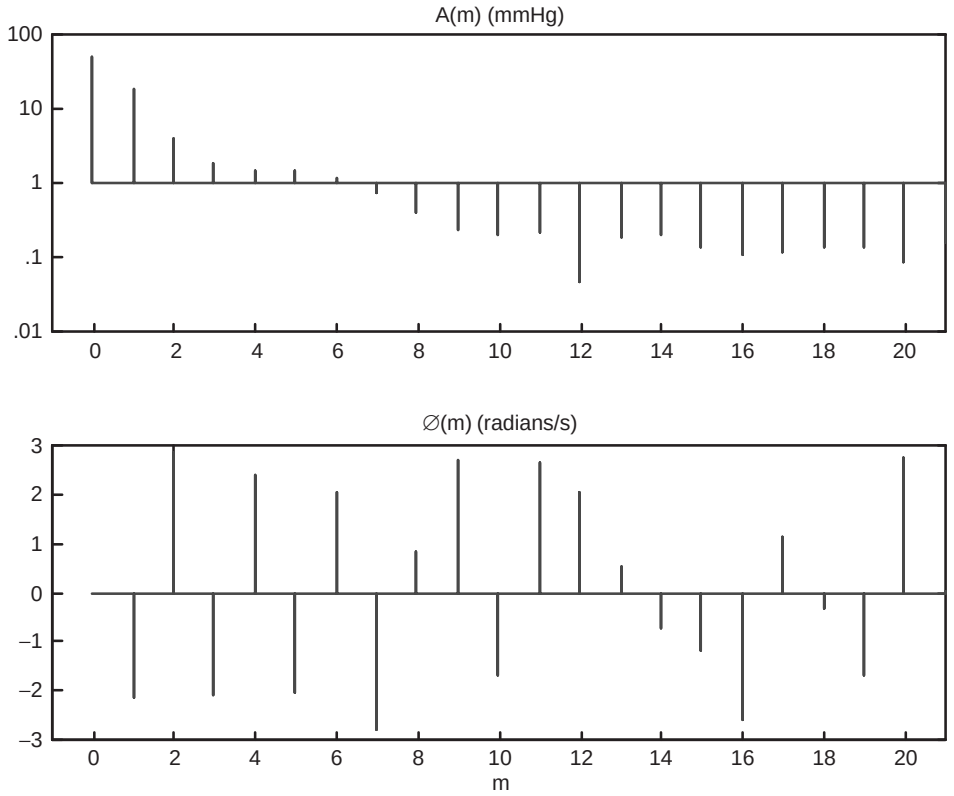


FIGURE 11.10 Harmonic coefficients of the aortic pressure waveform shown in Figure 11.2.

is borne out of the Fourier series representation for a periodic signal. A signal can be expressed either in the "time-domain" by the signal's time function, $x(t)$, or alternatively in the "frequency-domain" by specifying the Fourier coefficient and phase, A_m and ϕ_m, as a function of the signal's harmonic frequencies, $\omega_m = m\omega_o$. Thus, if we know the Fourier coefficients and the frequency components that make up the signal, we can fully recover the periodic signal $x(t)$.

One of the disadvantages of the Fourier series is that it applies only to periodic signals, and many biological signals are not periodic. In fact, a broad class of biological signals includes signals that are continuous functions of time but that never repeat in time. Luckily, the concept of Fourier series can also be extended for signals that are not periodic. The Fourier integral, also referred to as the Fourier transform, is used to decompose a continuous aperiodic signal into its constituent frequency components

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{11.6}$$

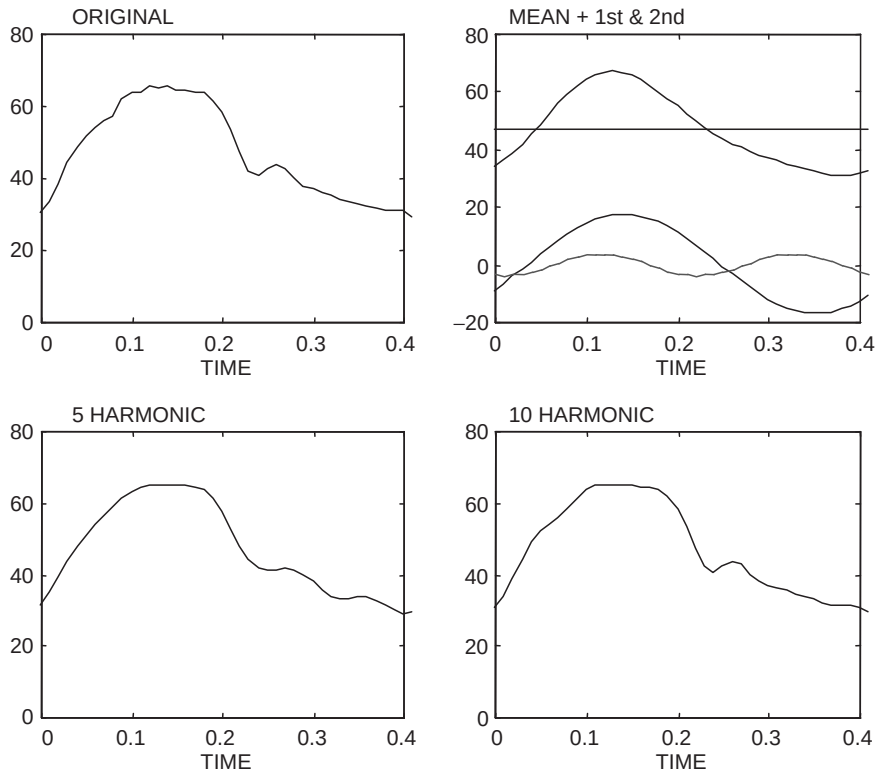


FIGURE 11.11 Harmonic reconstruction of the aortic pressure waveform shown in Figure 11.2.

much as the Fourier series decomposes a periodic signal into its corresponding trigonometric components. $X(\omega)$ is a complex valued function of the continuous frequency, ω, and is analogous to the coefficients of the complex Fourier series, c_m. A rigorous proof for this relationship is beyond the scope of this text, but it is useful to note that the Fourier integral is derived directly from the exponential Fourier series by allowing the period, T, to approach infinity. The coefficients c_m of the trigonometric series approach $X(\omega)$ as $T \to \infty$. Conceptually, a function that repeats at infinity can be considered as aperiodic, since you will never observe it repeating. Tables of Fourier transforms for many common signals can be found in most signals and systems or signal processing textbooks.

As for the Fourier series, a procedure for converting the frequency-domain version of the signal, $X(\omega)$, to its time-domain expression is desired. The time-domain signal, $x(t)$, can be completely recovered from the Fourier transform with the inverse Fourier transform (IFT)

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega. \tag{11.7}$$

These two representations of a signal are interchangeable, meaning that we can always go back and forth between the time-domain version of the signal, $x(t)$, and the frequency-domain

version obtained with the Fourier transform, $X(\omega)$. The frequency domain expression therefore provides all of the necessary information for the signal and allows one to analyze and manipulate biological signals from a different perspective.

EXAMPLE PROBLEM 11.9

Find the Fourier Transform (FT) of the rectangular pulse signal

$$x(t) = 1, |t| < a$$
$$0, |t| > a$$

Solution

Equation (11.6) is used.

$$X(\omega) = \int_{-a}^{a} e^{-j\omega t} dt = \frac{e^{-j\omega t}}{-j\omega}\Bigg|_{-a}^{a} = \frac{2\sin \omega a}{\omega}$$

As for the Fourier series representation of a signal, the magnitude and the phase are important attributes of the Fourier transform. As stated previously, $X(\omega)$ is a complex valued function, meaning that it has a real, $\mathrm{Re}\{X(\omega)\}$, and imaginary, $\mathrm{Im}\{X(\omega)\}$, component and can be expressed as

$$X(\omega) = \mathrm{Re}\{X(\omega)\} + j\,\mathrm{Im}\{X(\omega)\}. \tag{11.8}$$

As for the Fourier series, the magnitude determines the amplitude of each complex exponential function (or equivalent cosine) required to reconstruct the desired signal, $x(t)$, from its Fourier transform

$$|X(\omega)| = \sqrt{\mathrm{Re}\{X(\omega)\}^2 + \mathrm{Im}\{X(\omega)\}^2} \tag{11.9}$$

In contrast, the phase determines the time shift of each cosine signal relative to a reference of time zero. It is determined as

$$\theta(\omega) = \tan^{-1}\left(\frac{\mathrm{Im}\{X(\omega)\}}{\mathrm{Re}\{X(\omega)\}}\right). \tag{11.10}$$

Note the close similarity for determining the magnitude and phase from the trigonometric and compact forms of the Fourier series (Eqs. (11.4a, b, c)). The magnitude of the Fourier transform, $|X(\omega)|$, is analogous to A_m, whereas a_m and b_m are analogous to $\mathrm{Re}\{X(\omega)\}$ and $\mathrm{Im}\{X(\omega)\}$, respectively. The equations are identical in all other respects.

EXAMPLE PROBLEM 11.10

Find the magnitude and phase of the signal with the Fourier transform

$$X(\omega) = \frac{1}{1 + j\omega}$$

Continued

Solution

The signal has to be put in a recognizable form similar to Eq. (11.8). To achieve this,

$$X(\omega) = \frac{1}{1+j\omega} \cdot \frac{1-j\omega}{1-j\omega} = \frac{1-j\omega}{1+\omega^2} = \frac{1}{1+\omega^2} - j\frac{\omega}{1+\omega^2}.$$

Therefore

$$\mathrm{Re}\{X(\omega)\} = \frac{1}{1+\omega^2} \quad \text{and} \quad \mathrm{Im}\{X(\omega)\} = -\frac{\omega}{1+\omega^2}.$$

Using Eqs. (11.9) and (11.10), the magnitude is

$$|X(\omega)| = \frac{1}{1+\omega^2}$$

and the phase

$$\theta(\omega) = \tan^{-1}(-\omega) = -\tan^{-1}(\omega).$$

11.5.5 Properties of the Fourier Transform

In practice, computing Fourier transforms for complex signals may be somewhat tedious and time consuming. When working with real-world problems, it is therefore useful to have tools available that help simplify calculations. The FT has several properties that help simplify frequency domain transformations. Some of these are summarized following.

Let $x_1(t)$ and $x_2(t)$ be two signals in the time domain. The FTs of $x_1(t)$ and $x_2(t)$ are represented as $X_1(\omega) = F\{x_1(t)\}$ and $X_2(\omega) = F\{x_2(t)\}$.

Linearity

The Fourier transform is a linear operator. Therefore, for any constants a_1 and a_2,

$$F\{a_1x_1(t) + a_2x_2(t)\} = a_1X_1(\omega) + a_2X_2(\omega) \tag{11.11}$$

This result demonstrates that the scaling and superposition properties defined for a liner system also hold for the Fourier transform.

Time Shifting/Delay

If $x_1(t - t_0)$ is a signal in the time domain that is shifted in time, the Fourier transform can be represented as

$$F\{x_1(t - t_0)\} = X(\omega) \cdot \mathrm{e}^{-j\omega \cdot t_0} \tag{11.12}$$

In other words, shifting a signal in time corresponds to multiplying its Fourier transform by a phase factor, $\mathrm{e}^{-\mathrm{j}\omega\mathrm{t}_0}$.

Frequency Shifting

If $X_1(\omega-\omega_0)$ is the Fourier transform of a signal, shifted in frequency, the inverse Fourier transform is

$$F^{-1}\{X_1(\omega - \omega_0)\} = x(t) \cdot \mathrm{e}^{-j\omega_0 t} \tag{11.13}$$

Convolution Theorem

The convolution between two signals, $x_1(t)$ and $x_2(t)$, in the time domain is defined as

$$c(t) = \int_{-\infty}^{\infty} x_1(\tau)x_2(t-\tau)d\tau = x_1(t) * x_2(t) \tag{11.14}$$

where * is shorthand for the convolution operator. The convolution has an equivalent expression in the frequency domain

$$C(\omega) = F\{c(t)\} = F\{x_1(t) * x_2(t)\} = X_1(\omega) \cdot X_2(\omega). \tag{11.15}$$

Convolution in the time domain, which is relatively difficult to compute, is a straightforward multiplication in the frequency domain.

Next, consider the convolution of two signals, $X_1(\omega)$ and $X_2(\omega)$, in the frequency domain. The convolution integral in the frequency domain is expressed as

$$X(\omega) = \int_{-\infty}^{\infty} X_1(v)X_2(\omega - v)dv = X_1(\omega) * X_2(\omega) \tag{11.16}$$

It can be shown that the inverse Fourier transform (IFT) of $X(\omega)$ is

$$x(t) = F^{-1}\{X(\omega)\} = F^{-1}\{X_1(\omega) * X_2(\omega)\} = 2\pi \cdot x_1(t) \cdot x_2(t) \tag{11.17}$$

Consequently, the convolution of two signals in the frequency domain is 2π times the product of the two signals in the time domain. As we will see subsequently for linear systems, convolution is an important mathematical operator that fully describes the relationship between the input and output of a linear system.

EXAMPLE PROBLEM 11.11

What is the FT of $3 \sin(25t) + 4 \cos(50t)$? Express your answer only in a symbolic equation. Do not evaluate the result.

Solution

$$F\{3 \sin(25t) + 4 \cos(50t)\} = 3F\{\sin(25t)\} + 4F\{\cos(50t)\}$$

11.5.6 Discrete Fourier Transform

In digital signal applications, continuous biological signals are first sampled by an analog-to-digital converter (see Figure 11.4) and then transferred to a computer, where they can be further analyzed and processed. Since the Fourier transform applies only to continuous signals of time, analyzing discrete signals in the frequency domain requires that we first modify the Fourier transform equations so they are structurally compatible with the digital samples of a continuous signal.

The discrete Fourier transform (DFT)

$$X(m) = \sum_{k=0}^{N-1} x(k) e^{-j\frac{2\pi mk}{N}}; m = 0, 1, \ldots, N/2 \tag{11.18}$$

provides the tool necessary to analyze and represent discrete signals in the frequency domain. The DFT is essentially the digital version of the Fourier transform. The index m represents the digital frequency index, $x(k)$ is the sampled approximation of $x(t)$, k is the discrete time variable, N is an even number that represents the number of samples for $x(k)$, and $X(m)$ is the DFT of $x(k)$.

The inverse discrete Fourier transform (IDFT) is the discrete-time version of the inverse Fourier transform. The inverse discrete Fourier transform (IDFT) is represented as

$$x(k) = \frac{1}{N} \sum_{m=0}^{N-1} X(m) e^{j\frac{2\pi mk}{N}}; k = 0, 1, \ldots, N-1 \tag{11.19}$$

As for the FT and IFT, the DFT and IFT represent a Fourier transform pair in the discrete domain. The DFT allows one to convert a set of digital time samples to its frequency domain representation. In contrast, the IDFT can be used to invert the DFT samples, allowing one to reconstruct the signal samples $x(k)$ directly from its frequency domain form, $X(m)$. These two equations are thus interchangeable, since either conveys all of the signal information.

EXAMPLE PROBLEM 11.12

Find the discrete Fourier transform of the signal $x(k) = 0.25^k$ for $k = 0{:}15$

$$X(m) = \sum_{k=0}^{N-1} x(k) e^{-j\frac{2\pi mk}{N}} = \sum_{k=0}^{15} 0.25^k e^{-j\frac{2\pi mk}{16}} = \sum_{k=0}^{15} \left(0.25 \cdot e^{-j\frac{2\pi m}{N}}\right)^k = \sum_{k=0}^{15} a^k$$

Note that the preceding is a geometric sum in which $a = 0.25 \cdot e^{-j\frac{2\pi m}{N}}$. Since for a geometric sum

$$\sum_{k=M}^{N} a^k = \frac{a^{N+1} - a^M}{a - 1}$$

we obtain

$$X(m) = \frac{a^{16} - a^0}{a - 1} = \frac{0.25^{16} e^{-j\frac{32m\pi}{N}} - 1}{0.25 e^{-j\frac{2m\pi}{N}} - 1}$$

An efficient computer algorithm for calculating the DFT is the fast Fourier transform (FFT). The output of the FFT and DFT algorithms are the same, but the FFT has a much faster execution time than the DFT (proportional to $N \cdot \log_2(N)$ versus N^2 operations). The ratio of computing time for the DFT and FFT is therefore

$$\frac{DFT\ computing\ time}{FFT\ computing\ time} = \frac{N^2}{N \cdot \log_2 N} = \frac{N}{\log_2 N} \tag{11.20}$$

In order for the FFT to be efficient, the number of data samples, N, must be a power of two. If $N = 1024$ signal samples, the FFT algorithm is approximately $1024/\log_2(1024) = 10$ times faster than the direct DFT implementation. If N is not a power of two, alternate DFT algorithms are usually used.

Figures 11.12 and 11.13 show two signals and the corresponding DFT, which was calculated using the FFT algorithm. The signal shown in Figure 11.12a is a sine wave

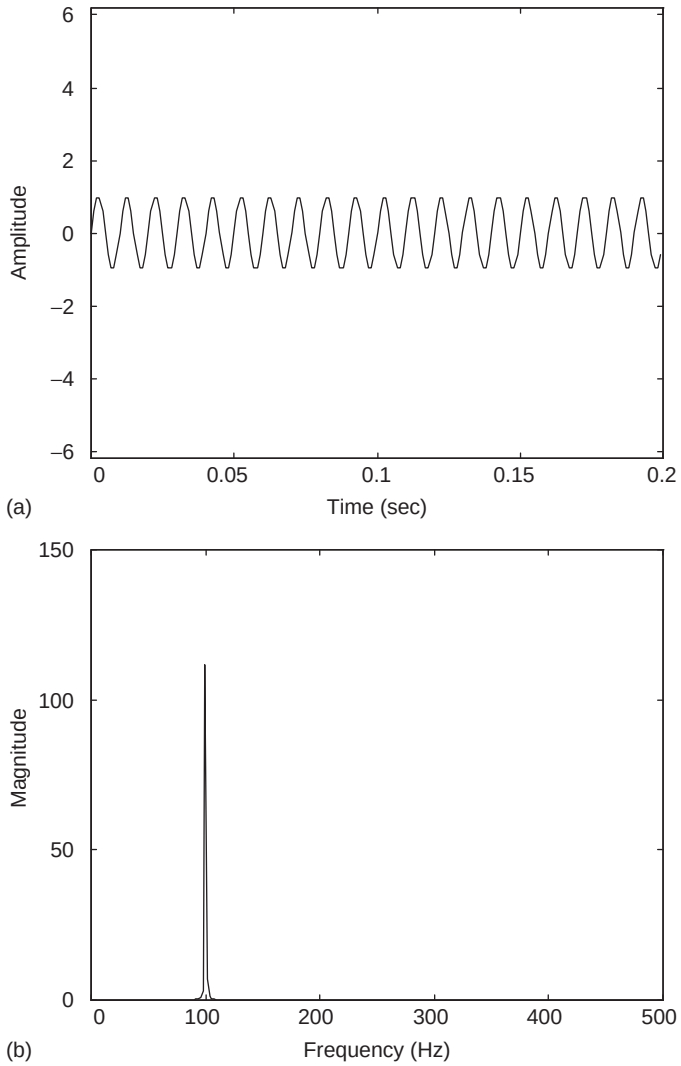


FIGURE 11.12 (a) 100 Hz sine wave. (b) Fast Fourier transform (FFT) of 100 Hz sine wave.

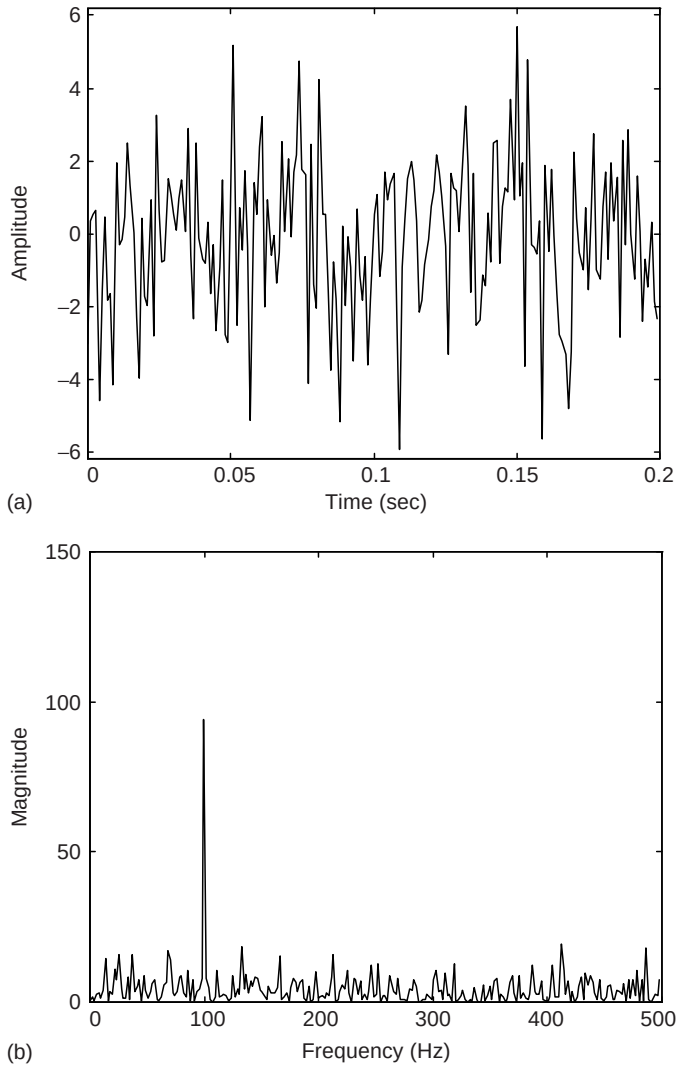


FIGURE 11.13 (a) 100 Hz sine wave corrupted with noise. (b) Fast Fourier transform (FFT) of noisy 100 Hz sine wave.

of frequency of 100 Hz. Figure 11.12b shows the FFT of the 100 Hz sine wave. Notice that the peak of the FFT occurs at 100 Hz frequency, indicating that all of the energy is confined to this frequency. Figure 11.13a shows a 100 Hz sine wave corrupted with random noise that was added to the waveform. The frequency of the signal is not distinct in the time domain. After transforming this signal to the frequency domain, the signal (Figure 11.13b) reveals a definite component at 100 Hz frequency, which is marked by the large peak in the FFT.

EXAMPLE PROBLEM 11.13

Find and plot the magnitude of the discrete Fourier transform of the signal $x(n) = \sin(\pi/4 \cdot n) + 2 \cdot \cos(\pi/3 \cdot n)$ in MATLAB.

Solution

```
n=1:1024; %Discrete Time Axis
x=sin(pi/4*n)+2*cos(pi/3*n); %Generating the signal
X=fft(x,1024*16)/1024; %Computing 16k point Fast Fourier Transform
Freq=(1:1024*16)/(1024*16)*2*pi; %Normalizing Frequencies
between 0-2*pi
plot(Freq,abs(X),'k') %Plotting
axis([0.7 1.15 0 1.2])
xlabel('Frequency (rad/s)')
ylabel('Fourier Magnitude')
```

Results are shown in Figure 11.14.

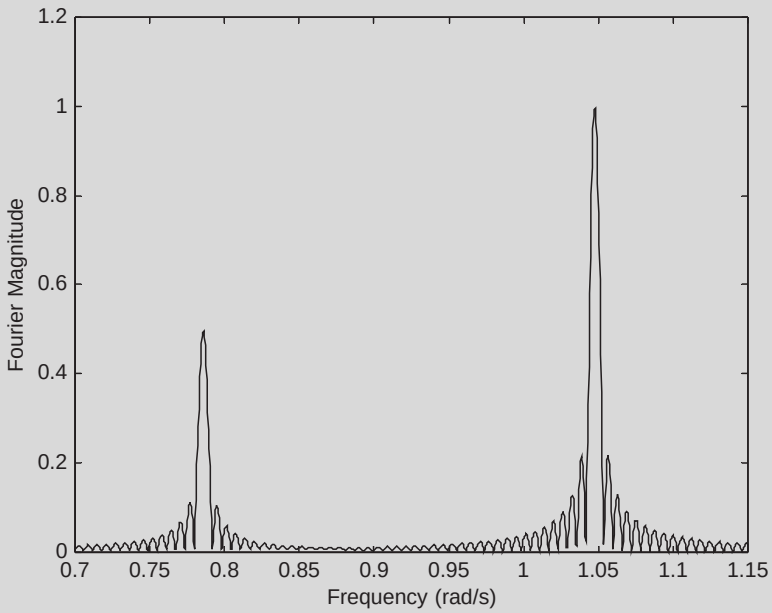


FIGURE 11.14 Fast Fourier transform magnitude for the sum of two sinusoids. Dominant energy peaks are located at the signal frequencies $\pi/3$ and $\pi/4$ rad/s.

11.5.7 The *z*-Transform

The *z*-transform provides an alternative tool for analyzing discrete signals in the frequency domain. This transform is essentially a variant of the DFT, where we allow $z = e^{-j\frac{2\pi m}{N}}$.

In most applications, the z-transform is somewhat easier to work with than the DFT because it does not require the use of complex numbers directly. The z-transform plays a similar role for digital signals as the Laplace transform does for the analysis of continuous signals.

If a discrete sequence $x(k)$ is represented by x_k, the (one-sided) z-transform of the discrete sequence is expressed by

$$X(z) = \sum_{k=0}^{\infty} x_k z^{-k} = x_0 + x_1 z^{-1} + x_2 z^{-2} + \mathrm{K} + x_k z^{-k} \tag{11.21}$$

Note that the z-transform can be obtained directly from the DFT by allowing $N \to \infty$ and replacing $z = e^{-j\frac{2\pi m}{N}}$ in Eq. (11.18). In most practical applications, sampled biological signals are represented by a data sequence with N samples so the z-transform is estimated for $k = 0 \ldots N\text{-}1$ only. Tables of common z-transforms and their inverse transforms can be found in most digital signal processing textbooks.

After a continuous signal has been sampled into a discrete sequence, its z-transform is found quite easily. Since the data sequence of a sampled signal is represented as

$$\boldsymbol{x} = [x(0), x(T), x(2T), \ldots, x(kT)] \tag{11.22}$$

its z-transform is obtained by applying Eq. (11.21) to its samples

$$X(z) = x(0) + x(T)z^{-1} + x(2T)z^{-2} + \ldots + x(kT)z^{-k} \tag{11.23}$$

where T is the sampling period or sampling interval.

A sampled signal is a data sequence with each sample separated from its neighboring samples by precisely one sampling period. In the z-transform, the value of the multiplier, $x(kT)$, is the value of the data sample. The terms z^{-k} have an intuitive graphical explanation. The power, k, corresponds to the number of sampling periods following the start of the sampling process at time zero; z^{-k} can therefore be thought of as a "shift operator" that delays the sample by exactly k sampling periods or kT. The variable z^{-1}, for instance, represents a time separation of one period, T, following the start of the signal at time zero. In Eq. (11.18), $z(0)$ is the value of the sampled data at $t = 0$, and $x(T)$ is the value of the sampled data that was obtained after the first sampling period. The z-transform is an important method for describing the sampling process of an analog signal.

EXAMPLE PROBLEM 11.14

The discrete unit impulse function is represented as the sequence $\boldsymbol{x} = [1, 0, 0, 0, \ldots, 0]$. Find the z-transform of this sequence.

Solution

$$\mathrm{X}(z) = 1 + 0z^{-1} + 0z^{-2} + \ldots + 0z^{-k} = 1 + 0 + 0 + \ldots + 0 = 1$$

EXAMPLE PROBLEM 11.15

An A/D converter is used to convert a recorded signal of the electrical activity inside a nerve into a digital signal. The first five samples of the biological signal are [−60.0, −49.0, −36.0, −23.0, −14.0] mV. What is the z-transform of this data sequence? How many sample periods after the start of the sampling process was the data sample –23.0 recorded?

Solution

$$Y(z) = -60.0 - 49.0z^{-1} - 36.0z^{-2} - 23.0z^{-3} - 14.0z^{-4}$$

The value of the negative exponent of the −23.0 mV z-term is 3. Therefore, the data sample with the value of −23.0 was recorded 3 sampling periods after the start of sampling.

11.5.8 Properties of the z-Transform

The z-transform obeys many of the same rules and properties that we've already shown for the Fourier transform. These properties can significantly simplify the process of evaluating z-transforms for complex signals. The following are some of the properties of the z-transform. Note the close similarity to the properties for Eqs. (11.11), (11.12), and (11.14). Let $x_1(k)$ and $x_2(k)$ be two digital signals with corresponding z-transforms $X_1(z)$ and $X_2(z)$.

Linearity:
The z-transform is a linear operator. For any constants a_1 and a_2,

$$Z\{a_1x_1(k) + a_2x_2(k)\} = \sum_{k=0}^{\infty}[a_1x_1(k) + a_2x_2(k)]z^{-k} = a_1X_1(z) + a_2X_2(z) \tag{11.24}$$

Delay:
Let $x_1(k - n)$ be the original signal that is delayed by n samples. The z-transform of the delayed signal is

$$Z\{x_1(k-n)\} = \sum_{k=0}^{\infty} x_1(k-n)z^{-k} = \sum_{k=0}^{\infty} x_1(k)z^{-(k+n)} = z^{-n}X_1(z) \tag{11.25}$$

As described previously, note that the operator z^{-n} represents a shift of n samples or precisely nT seconds.

Convolution:
Let $x(k)$ be the discrete convolution between $x_1(k)$ and $x_2(k)$,

$$x(k) = x_1(k) * x_2(k)$$

$X(z)$, the z-transform of $x(k)$, is calculated as

$$X(z) = Z\{x(k)\} = X_1(z)X_2(z) \tag{11.26}$$

As with the Fourier transform, this result demonstrates that convolution between two sequences is performed by simple multiplication in the z-domain.

11.6 LINEAR SYSTEMS

A system is a process, machine, or device that takes a signal as an *input* and manipulates it to produce an *output* that is related to, but is distinctly different from, its input. Figure 11.15 shows a system block diagram.

Biological systems and organs are very often modeled as systems. The heart, for instance, is a large-scale system that takes oxygen-deficient blood from the veins (the *input*) and pumps it through the lungs. This produces a blood *output* via the main arteries of the heart that is rich in oxygen content. Neurons in the brain can also be thought of as a simple microscopic system that takes electrical nerve impulses from various neurons as the input and sums these impulses to produce action potentials: the output. Linear systems are a special class of systems with a unique set of properties that make them easy to analyze.

11.6.1 Linear System Properties

While biological systems are not per se linear, very often they can be approximated by a linear system model. This is desired because it makes their analysis and the subsequent interpretation more tractable.

All linear systems are characterized by the principles of *superposition* (or *additivity*) and *scaling*. The superposition property states that the sum of two independent inputs produces an output that is the sum or superposition of the outputs for each individual input. The scaling property tells us that a change in the size of the input produces a comparable change at the output. Mathematically, if we know the outputs for two separate inputs—that is,

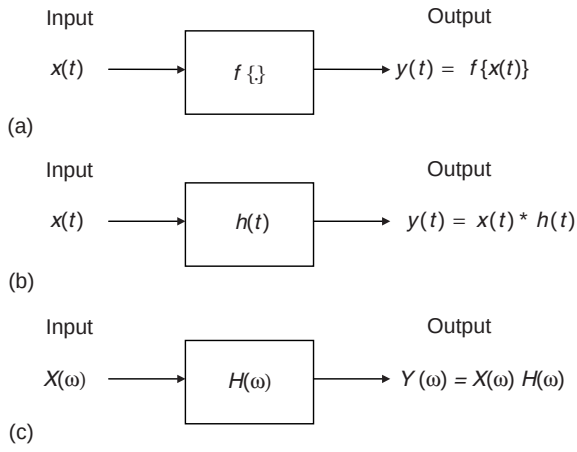


FIGURE 11.15 (a) Block diagram representation of a system. The input signal, $x(t)$, passes through the system transformation $f\{\cdot\}$ to produce an output, $y(t)$. (b) Time domain representation of a linear system. The output of the linear system is represented by the convolution of the input and impulse response. (c) Frequency domain representation of a linear system. The output corresponds to the product of the input and the system transfer function.

$$\begin{array}{ccc} \text{Input} & & \text{Output} \\ x_1(t) & \rightarrow & y_1(t) \\ x_2(t) & \rightarrow & y_2(t), \end{array}$$

we can easily determine the output of a linear system to any arbitrary combination of these inputs. More generally, a linear superposition and scaling of the input signals produces a linear superposition and scaling of the output signals

$$\begin{array}{ccc} \text{Input} & & \text{Output} \\ k_1 \cdot x_1(t) + k_2 \cdot x_2(t) & \rightarrow & k_1 \cdot y_1(t) + k_2 \cdot y_2(t) \end{array} \tag{11.27}$$

where k_1 and k_2 are arbitrary amplitude scaling constants. These constants scale the input amplitudes by making them larger ($k > 1$) or smaller ($k < 1$). This produces a comparable change in the net outputs, which are likewise scaled by the same constants.

EXAMPLE PROBLEM 11.16

The following information is given for a linear system

$$\begin{array}{ccc} \text{Input} & & \text{Output} \\ x_1(t) = \cos(t) & \rightarrow & y_1(t) = \cos(t + \pi/2) \\ x_2(t) = \cos(t) + \sin(2t) & \rightarrow & y_2(t) = \cos(t + \pi/2) + 5\sin(2t) \\ x_3(t) = \cos(3t) & \rightarrow & y_3(t) = 2\cos(3t) \end{array}$$

Find the output if the input is: $x(t) = 3\sin(2t) + 1/2\cos(3t)$

Solution

The input is represented as a superposition of x_1, x_2, and x_3

$$x(t) = 3(x_2(t) - x_1(t)) + 1/2x_3(t) = 3x_2(t) - 3x_1(t) + 1/2x_3(t)$$

Applying the superposition and scaling properties produces an output

$$y(t) = 3y_2(t) - 3y_1(t) + 1/2y_3(t) = 3(\cos(t + \pi/2) + 5\sin(2t)) - 3(\cos(t + \pi/2)) + 1/2(2\cos(3t))$$
$$= 15 \cdot \sin(2t) + \cos(3t)$$

EXAMPLE PROBLEM 11.17

Consider the system given by the expression

$$y(t) = f\{x(t)\} = A \cdot x(t) + B.$$

Determine if this is a linear system.

Solution

To solve this problem, consider a superposition of two separate inputs, $x_1(t)$ and $x_2(t)$, that independently produce outputs $y_1(t)$ and $y_2(t)$. Apply the input $x(t) = k_1 \cdot x_1(t) + k_2 \cdot x_2(t)$. If the system is linear, the output obeys

Continued

$$\begin{aligned} y_{Lin}(t) &= k_1 \cdot y_1(t) + k_2 \cdot y_2(t) = k_1 \cdot (A \cdot x_1(t) + B) + k_2 \cdot (A \cdot x_2(t) + B) \\ &= A(k_1 \cdot x_1(t) + k_2 \cdot x_2(t)) + (k_1 + k_2)B \end{aligned}$$

The true system output, however, is determined as

$$y(t) = f\{x(t)\} = f\{k_1 \cdot x_1(t) + k_2 \cdot x_2(t)\} = A \cdot (k_1 \cdot x_1(t) + k_2 \cdot x_2(t)) + B$$

We need to compare our expected linear system output, $y_{lin}(t)$, with the true system output, $y(t)$. Note that $y(t) \neq y_{Lin}(t)$, and therefore the system is not linear.

The superposition principle takes special meaning when applied to periodic signals. Because periodic signals are expressed as a sum of cosine or complex exponential functions with the Fourier series, their output must also be expressed as a sum of cosine or exponential functions. Thus, if a linear system is stimulated with a periodic signal, its output is also a periodic signal with identical harmonic frequencies. The output, $y(t)$, of a linear system to a periodic input, $x(t)$, is related by

$$\overbrace{x(t) = \frac{A_0}{2} + \sum_{m=-\infty}^{+\infty} A_m \cos(m\omega_0 t + \phi_m)}^{\text{Input}} \quad \Rightarrow \quad \overbrace{x(t) = \frac{B_0}{2} + \sum_{m=-\infty}^{+\infty} B_m \cos(m\omega_0 t + \theta_m)}^{\text{Output}} \tag{11.28}$$

meaning the input and output contain cosines with identical frequencies, $m\omega_0$, and are expressed by equations with similar form. A similar form of this expression is also obtained for the exponential Fourier series:

$$\overbrace{x(t) = \sum_{m=-\infty}^{+\infty} c_m e^{jm\omega_0 t}}^{\text{Input}} \quad \Rightarrow \quad \overbrace{y(t) = \sum_{m=-\infty}^{+\infty} b_m e^{jm\omega_0 t}}^{\text{Output}} \tag{11.29}$$

where the input and output coefficients, c_m and b_m, are explicitly related to A_m and B_m via Eq. (11.5b).

From Eqs. (11.28) and (11.29) the input and output of a linear system to a periodic input differ in two distinct ways. First the amplitudes of each cosine are selectively scaled by different constants, A_m for the input and B_m for the output. These constants are uniquely determined by the linear system properties. Similarly, the phases angle of the input components, ϕ_m, are different from the output components, θ_m, meaning that the input and output components are shifted in time in relationship to each other. As for the amplitudes, the phase difference between the input and output is a function of the linear system. Thus, if we know the mathematical relationship of how the input components are amplitude scaled and phase shifted between the input and output, we can fully describe the linear system. This relationship is described by the system *transfer function*, H_m. The transfer function fully describes how the linear system manipulates the amplitude and phases of the input to produce a specific output. This transformation is described by two separate components: the *magnitude* and the *phase*.

The magnitude of H_m is given by the ratio of the output to the input for the *m-th* component

$$|H_m| = \frac{B_m}{A_m} \tag{11.30}$$

Note that if we know the input magnitudes, we can determine the output Fourier coefficients by multiplying the transfer function magnitude by the input Fourier coefficients: $B_m = |H_m| \cdot A_m$. The phase angle of the transfer function describes the phase relationship between the input and output for the *m-th* frequency component

$$\angle H_m = \theta_m - \phi_m. \tag{11.31}$$

If we know the input phase, the output phase is determined as $\theta_m = \angle H_m + \phi_m$. Equations (11.30) and (11.31) are the two critical pieces of information that are necessary to fully describe a linear system. If these two properties of the transformation are known, it is possible to determine the output for any arbitrary input.

11.6.2 Time Domain Representation of Linear Systems

The relationship between the input and output of a linear system can be described by studying its behavior in the time domain (Figure 11.15b). The *impulse response* function, $h(t)$, is a mathematical description of the linear system that fully characterizes its behavior. As we will see subsequently, the impulse response of a linear system is directly related to the system transfer function as outlined for the periodic signal. If one knows $h(t)$, one can readily compute the output, $y(t)$, to any arbitrary input, $x(t)$, using the *convolution integral*

$$y(t) = h(t) * x(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau. \tag{11.32}$$

The symbol * is shorthand for the convolution between the input and the system impulse response. Integration is performed with respect to the dummy integration variable τ. For the discrete case, the output of a discrete linear system is determined with the convolution sum

$$y(k) = h(k) * x(k) = \sum_{m=-\infty}^{\infty} h(m)x(k-m) \tag{11.33}$$

where $h(m)$ is the impulse response of the discrete system. A detailed treatment of the convolution integral is found in many signal processing textbooks and is beyond the scope of this text. As shown in a subsequent section, a simpler treatment of the input-output relationship of a linear system is obtained by analyzing it in the "frequency-domain."

EXAMPLE PROBLEM 11.18

A cytoplasmic current injection $i(t) = u(t)$ to a cell membrane produces an intracellular change in the membrane voltage, $v(t)$. The membrane of a cell is modeled as a linear system with impulse response $h(t) = A \cdot e^{-t/\tau} \cdot u(t)$, where A is a constant in units V/s/A and τ is the cell membrane time constant (units: seconds). Find the cell membrane voltage output in closed form. Simulate in MATLAB, for A = 100 and τ = 0.01 sec. Compare the closed form solution to the simulated results.

Continued

Solution

The input, $i(t)$, and output, $v(t)$, are related by the convolution integral (Eq. (11.32)):

$$v(t) = h(t) * i(t) = \int_{-\infty}^{\infty} h(\varsigma) i(t-\varsigma) d\varsigma = \int_{-\infty}^{\infty} A \cdot e^{-\varsigma/\tau} u(\varsigma) u(t-\varsigma) d\varsigma$$

where we use a dummy integration variable, ς, to distinguish it from the cell time constant, τ. The unit step functions inside the integral take values of one or zero—in which case they do not contribute to the integral. $u(\varsigma) = 1$ if $\varsigma > 0$ and $u(t-\varsigma) = 1$ if $t - \varsigma > 0$. Combining these two inequalities, we have that $u(t-\varsigma) \cdot u(\varsigma) = 1$ whenever $0 < \varsigma < t$, and we can therefore change the limits of integration and replace the unit step function with 1:

$$v(t) = \int_{0}^{t} A \cdot e^{-\varsigma/\tau} d\varsigma = A \cdot \tau \cdot \left(1 - e^{-t/\tau}\right). \quad .$$

MATLAB Solution

```
%Theoretical Result
dt=0.0001; %Sampling Resolution
Fs=1/dt; %Sampling Rate
time=0:dt:5; %5-second time-axis
Tau=0.1; %Cell Membrane Time Constant
A=10;
y=A*Tau*(1-exp(-time/Tau)); %Closed Form Output Equantion
plot(time,y,'color',[.75 .75 .75],'linewidth',3) %Plotting
Theoretical Output
hold on

%Simulated Output
h=A*exp(-time/Tau); %Impulse Response
x=[zeros(1,Fs) ones(1,Fs*4)]; %Step Input
y=conv(h,x)*dt; %Step Response, obtained by convolving: y=h*x

%Plotting Simulated Results
time=(0:length(x)-1)*dt-1;
plot(time,x,'k-.') %Plotting Input
hold on
time=(0:length(y)-1)*dt-1;
plot(time,y,'k') %Plotting Output
axis([-1 1 0 1.2])
xlabel('time (sec)','fontsize',14)
ylabel('Amplitude','fontsize',14)
set(gca,'XTick',[-1 -0.5 0 0.5 1 ])
set(gca,'YTick',[0:.2:1.2])
```

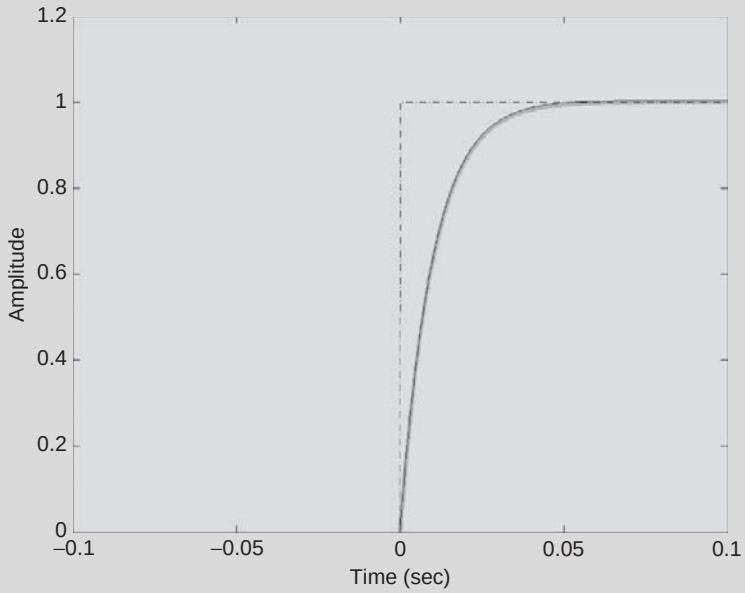


FIGURE 11.16 MATLAB results showing the step response of a linear cell membrane. Dotted lines represent the step input. Continuous gray line shows the theoretical step response. Black line shows the simulated step response. Note that the theoretical solution and simulated results are superimposed.

The results for the cell membrane step response are shown in Figure 11.16. Note that the simulated (black) and theoretical (gray) outputs are precisely matched.

11.6.3 Frequency Domain Representation of Linear Systems

We have already considered the special case of linear systems in the frequency domain for periodic inputs. Recall that the system output of a linear system to a periodic stimulus is fully described by the system transfer function. The output of a linear system is also expressed in the time domain by the convolution integral (Figure 11.15b). The impulse response is the mathematical model that describes the linear system in the time domain. These two descriptions for the input-output relationship of a linear system are mutually related. Notably, for the aperiodic signal case, the transfer function is the Fourier transform of the impulse response

$$\begin{array}{c}\text{Time Domain}\\ h(t)\end{array} \quad \Leftrightarrow \quad \begin{array}{c}\text{Frequency Domain}\\ H(\omega)\end{array} \tag{11.34}$$

where $H(\omega)$ is the system transfer function. Since the impulse response is a complete model of a linear system and since the Fourier transform is invertible (we can always go back and forth between $h(t)$ and $H(\omega)$), the transfer function contains all the necessary information to fully describe the system. The advantage of the transfer function comes in its simplicity of use. Rather than performing a convolution integral, which can be quite intricate in many

applications, the output of a linear system in the frequency domain is expressed as a product of the system input and its transfer function: $Y(\omega) = X(\omega)H(\omega)$. This result is reminiscent of the result for the Fourier series (Eqs. (11.30) and (11.31)). Specifically, the *convolution property* of the Fourier transform states that a convolution in time corresponds to a multiplication in the frequency domain

$$\begin{array}{cc} \text{Time Domain} & \text{Frequency Domain} \\ y(t) = x(t) * h(t) \quad \Leftrightarrow & Y(\omega) = X(\omega)H(\omega) \end{array} \tag{11.35}$$

and thus the output of a linear system, $Y(\omega)$, is expressed in the frequency domain by the product of $X(\omega)$ and $H(\omega)$ (Figure 11.15c). Note that this result is essentially the convolution theorem (Eqs. (11.14) and (11.15); for proof, see Example Problem 11.19) applied to the output of a linear system (Eq. (11.32)). In many instances, this is significantly easier to compute than a direct convolution in the time domain.

EXAMPLE PROBLEM 11.19

Prove the convolution property of the Fourier transform.

Solution

$$Y(\omega) = FT\{y(t)\} = \int y(t)e^{-j\omega t}dt = \int\int x(\tau)h(t-\tau)d\tau e^{-j\omega t}dt$$

Make a change of variables, $u = t - \tau$, $du = dt$

$$= \int\int x(\tau)h(u)d\tau e^{-j\omega(u+\tau)}du = \int x(\tau)e^{-j\omega\tau}d\tau \cdot \int h(u)e^{-j\omega u}du = X(\omega)H(\omega)$$

11.6.4 Analog Filters

Filters are a special class of linear systems that are widely used to manipulate the properties of a biological signal. Conceptually, a filter allows the user to selectively remove an undesired signal component while preserving or enhancing some component of interest. Although most of us are unaware of this, various types of filters are commonplace in everyday settings. Sunblock, for instance, is a type of filter that removes unwanted ultraviolet light from the sun in order to minimize the likelihood of sunburn and potentially reduce the risk of skin cancer. Filters are also found in many audio applications. Treble and bass controls in an audio system are a special class of filter that the user selectively adjusts to boost or suppress the amount of high-frequency (treble) and low-frequency (bass) sound to a desired level and quality.

Filters play an important role in the analysis of biological signals, in part because signal measurements in clinical settings are often confounded by undesirable noise. Such noise distorts the signal waveform of interest, making it difficult to obtain a reliable diagnosis. Ideally, if one could completely remove unwanted noise, one could significantly improve the quality of a signal and thus minimize the likelihood of an incorrect diagnosis.

Practically, most filters can be subdivided into three broad classes, according to how they modify the frequency spectrum of the desired signal. These broad classes include *low-pass*, *high-pass*, and *band-pass* filters. Low-pass filters work by removing high frequencies from a signal while selectively keeping the low frequencies (Figure 11.17a). This allows the low frequencies of the signal to pass through the filter uninterrupted, hence the name low-pass. In some instances, low frequencies could be accentuated further by magnifying them while selectively removing the high frequencies. High-pass filters perform exactly the opposite function of a low-pass filter (Figure 11.17b). They selectively pass the high frequencies but remove the low frequencies of the signal. The treble control in an audio system is a form of high-pass filter that accentuates the high frequencies, thus producing crisp and rich sound. In contrast, the bass control is a form of low-pass filter that selectively enhances low frequencies or the "bass," creating a "warmer" sound quality. Band-pass filters fall somewhere in between the low-pass and high-pass filter. Rather than simply removing the low or high frequencies, band-pass filters remove both high and low frequencies but selectively keep a small "band" of frequencies (Figure 11.17c), hence the name. The function of a band-pass filter could be achieved by simply combining a low-pass and high-pass filter, as we will see subsequently.

Since filters are linear systems, the output of a filter is expressed by the convolution between the input and the filter's impulse response (Eq. (11.32)). Conversely, if the output is determined in the frequency domain, the output corresponds to the product of the filter

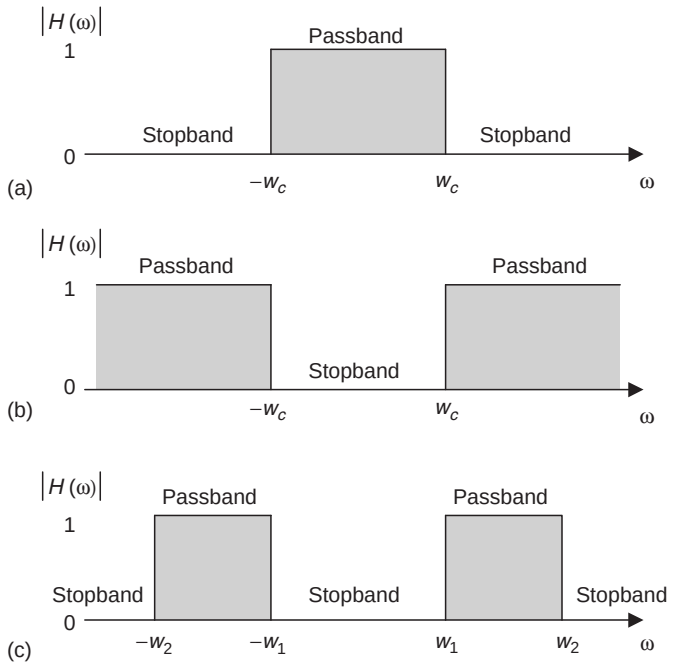


FIGURE 11.17 Frequency domain magnitude response plot, $|H(\omega)|$, of the ideal (a) low-pass filter, (b) high-pass filter, and (c) band-pass filter. Signals in the shaded region, the passband, are preserved at the output, whereas signals in the stopband are selectively removed from the output.

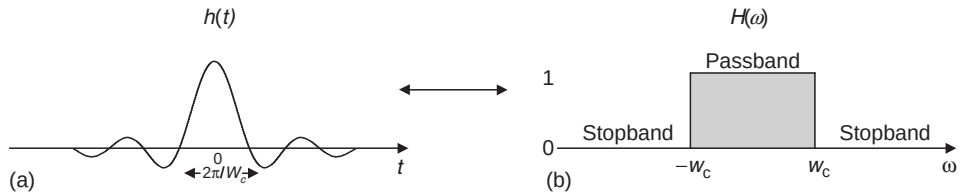


FIGURE 11.18 Time and frequency domain representation of the ideal low-pass filter. (a) The impulse response of the ideal low-pass filter, $h(t)$. (b) Transfer function of the ideal low-pass filter, $H(\omega)$.

transfer function and the input Fourier transform (Eq. (11.35)). The impulse response and transfer function of the ideal low-pass filter are shown in Figure 11.18. The transfer function of this filter takes a value of one within the filter passband and zero within the stopband.

Since the output of a linear system in the frequency domain is given as the product of the input Fourier transform and the signal transfer function, $Y(\omega) = H(\omega)X(\omega)$, any signal presented to this filter within its passband will pass through to the output uninterrupted because the frequency components are multiplied by one. In contrast, signals in the stopband are removed at the output of the filter, since the frequency components are multiplied by zero. The impulse response of the ideal low-pass analog filter is

$$h_{LP}(t) = \frac{W_c}{\pi}\,\mathrm{sinc}(W_c t). \tag{11.36}$$

where $W_c = 2\pi f_c$ is the filter cutoff frequency. In the frequency domain, the ideal low-pass filter transfer function is

$$H_{LP}(\omega) = \begin{cases} 1 & |\omega| < W_c \\ 0 & |\omega| > W_c \end{cases}. \tag{11.37}$$

This dual time and frequency domain representation of the ideal low-pass filter is shown in Figure 11.18. Note that the transfer function takes a value of one only within the passband. At the cutoff frequency, the filter transfer function transitions from a value of one in the passband to a value of zero in the stopband.

In the frequency domain, the ideal high-pass filter performs the exact opposite function of the low-pass filter

$$H_{HP}(\omega) = \begin{cases} 0 & |\omega| < W_c \\ 1 & |\omega| > W_c \end{cases}, \tag{11.38}$$

that is, the passband exists for frequencies above the cutoff frequency, whereas the stopband exists for frequencies below the filter's cutoff. This filter therefore preserves high-frequency signal components (above the cutoff frequency) and selectively removes low-frequency signals. The ideal high-pass filter transfer function can be easily derived from the ideal low-pass filter as

$$H_{HP}(\omega) = 1 - H_{LP}(\omega) \tag{11.39}$$

In the time domain, the ideal high-pass filter impulse response is obtained as

$$h_{HP}(t) = \delta(t) - h_{LP}(t) = \delta(t) - \frac{W_c}{\pi}\,\mathrm{sinc}(W_c t) \tag{11.40}$$

A schematic depiction of the ideal high-pass filter transfer function is shown in Figure 11.17b.

The final class of filter we will consider is the band-pass filter. The prototypical band-pass filter is somewhat more complex than the low-pass and high-pass filters, since it requires the definition of a lower and upper cutoff frequency, W_1 and W_2. Figure 11.17c illustrates the magnitude response of the ideal band-pass filter transfer function and impulse response. Only signals between the two cutoff frequencies are allowed to pass through to the output. All other signals are rejected. The transfer function of the ideal band-pass filter is given by

$$H_{HP}(\omega) = \begin{cases} 0 & W_1 < |\omega| < W_2, \\ 1 & \textit{otherwise} \end{cases} \tag{11.41}$$

In the frequency domain, the ideal band-pass filter can be obtained by combining a high-pass filter with cutoff W_1 and a low-pass filter with cutoff W_2. The band-pass filter transfer function can therefore be expressed as the product of transfer functions for a low-pass and high-pass filter:

$$H_{BP}(\omega) = H_{HP}(\omega) \cdot H_{LP}(\omega). \tag{11.42}$$

In the time domain, the band-pass filter impulse response is obtained by the inverse Fourier transform of the filter transfer function

$$h_{BP}(t) = h_{HP}(t) * h_{LP}(t). \tag{11.43}$$

This is done by applying the convolution theorem (Eqs. (11.14) and (11.15)) to Eq. (11.42).

EXAMPLE PROBLEM 11.20

Consider the cell membrane cytoplasmic current injection for Example Problem 11.18. Find the cell's output voltage in the Fourier domain. Also, compute and plot in MATLAB the transfer function magnitude of the cell membrane. What type of filter is this?

Solution

The Fourier transform of the step current input is

$$I(\omega) = \int i(t)e^{-j\omega t}dt = \int_0^\infty 1 \cdot e^{-\omega t}dt = \frac{e^{-\omega t}}{-j\omega}\bigg|_0^\infty = \frac{1}{j\omega}$$

The transfer function is determined as the Fourier transform of the impulse response

$$H(\omega) = \mathrm{FT}\{h(t)\} = \int A \cdot e^{-t/\tau} \cdot u(t)dt = \int_0^\infty A \cdot e^{-t/\tau}e^{-j\omega t}dt = \frac{A}{j\omega + 1/\tau}$$

Continued

The cell's voltage output in the frequency domain is determined as

$$V(\omega) = H(\omega)I(\omega) = \frac{A}{j\omega + 1/\tau} \cdot \frac{1}{j\omega} = \frac{A}{j\omega} - \frac{A}{j\omega + 1/\tau}$$

Next, using the derived transfer function, the transfer function magnitude is obtained as

$$|H(\omega)| = \sqrt{H(\omega) \cdot H(\omega)^*} = \sqrt{\frac{A}{j\omega + 1/\tau} \cdot \frac{A}{-j\omega + 1/\tau}} = \frac{A}{\sqrt{\omega^2 + 1/\tau^2}}$$

where $H(\omega)^*$ is the complex conjugate transfer function. In MATLAB the transfer function magnitude is plotted as follows:

```
%Cell Membrane Magnitude Response
tau=0.01;
A=1/tau^2;
w=0:0.1:500;
H=A./(w.^2+1/tau^2);
plot(w/2/pi,H,'k')
axis([0 500/2/pi 0 1.2])
xlabel('Frequency (Hz)')
ylabel('Magnitude')
```

The results are shown in Figure 11.19. Note that the transfer function tends to preserve the low frequencies while rejecting the high frequencies. Thus, the cell membrane behaves like a low-pass filter.

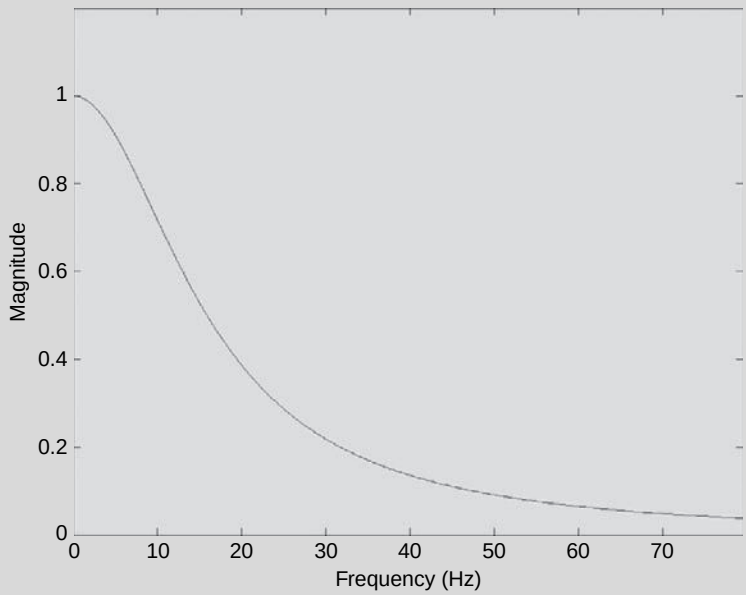


FIGURE 11.19 Cell membrane transfer function magnitude. Note that the cell behaves like a low-pass filter.

EXAMPLE PROBLEM 11.21

An electromyographic (EMG) signal contains energy within the frequencies 25 and 100 Hz. Design a filter to remove unwanted noise.

Solution

We need to design a band-pass filter with passband frequencies 25 and 100 Hz. First determine the cutoff frequencies in rad/s. Since $W_c = 2\pi f_c$,

$$W_1 = 50\pi$$

$$W_2 = 200\pi$$

Next, we find the impulse response of the corresponding low-pass and high-pass filters:

$$h_{HP}(t) = \delta(t) - \frac{W_1}{\pi}\,\text{sinc}\,(W_1 t) = \delta(t) - 50\,\text{sinc}\,(50\pi t)$$

$$h_{LP}(t) = \frac{W_2}{\pi}\,\text{sinc}\,(W_2 t) = 200\,\text{sinc}\,(200\pi t)$$

The band-pass filter impulse response is

$$h_{BP}(t) = h_{BP}(t) * h_{LP}(t) = [\delta(t) - 50\,\text{sinc}\,(50\pi t)] * 200\,\text{sinc}\,(200\pi t)$$

The described ideal analog filters provide a conceptual reference for various filter design applications. In practice, real analog filters cannot be implemented to achieve the strict specifications of the ideal filter because the impulse response of ideal filters is of infinite duration (extends from $-\infty$ to $+\infty$). Thus, the ideal filters require that one integrate over an infinite amount of time to produce an output. Typically, most analog filters are designed with simple electronic circuits. Various approximations to the ideal low-pass, high-pass, and band-pass filter can be derived that are well suited for a variety of applications, including signal analysis of biomedical signals.

11.6.5 Digital Filters

Digital systems are described by difference equations, just like analog systems are described by differential equations. Difference equations are essentially discretized differential equations that have been sampled at a particular sampling rate. The general form of a real-time digital filter/difference equation is

$$y(k) = \sum_{m=0}^{M} b_m x(k-m) - \sum_{m=1}^{N} a_m y(k-m) \tag{11.44}$$

where the discrete sequence $x(k)$ corresponds to the input and $y(k)$ represents the output sequence of the discrete system. For instance, if $M = 2$ and $N = 2$, then

$$y(k) = b_0 x(k) + b_1 x(k-1) + b_2 x(k-2) - a_1 y(k-1) - a_2 y(k-2)$$

where $x(k)$ and $y(k)$ represent the input and output at time k, $x(k-1)$ and $y(k-1)$ represent the input and output one sample into the past, and similarly, $x(k-2)$ and $x(k-2)$ correspond to the input and output two samples into the past.

Digital systems, like analog systems, can also be defined by their impulse responses, $h(k)$, and the convolution sum (Eq. (11.33)). If the response has a finite number of nonzero points, the filter is called an FIR filter or a "finite impulse response filter." If the response has an infinite number of nonzero points, the filter is called an IIR or "infinite impulse response filter." One positive quality of digital filters is the ease with which the output for any input can be calculated.

EXAMPLE PROBLEM 11.22

Find the impulse response for the digital filter

$$y(k) = \frac{1}{2}x(k) + \frac{1}{2}y(k-1)$$

Solution

Assume the system is at rest before input begins—that is, $y(n) = 0$ for $n < 0$.

$$y(-2) = \frac{1}{2}\delta(-2) + \frac{1}{2}y(-3) = 0 + 0 = 0$$

$$y(-1) = \frac{1}{2}\delta(-1) + \frac{1}{2}y(-2) = 0 + 0 = 0$$

$$y(0) = \frac{1}{2}\delta(0) + \frac{1}{2}y(-1) = \frac{1}{2} + 0 = \frac{1}{2}$$

$$y(1) = \frac{1}{2}\delta(1) + \frac{1}{2}y(0) = 0 + \frac{1}{2}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^2$$

$$y(2) = \frac{1}{2}\delta(2) + \frac{1}{2}y(1) = 0 + \frac{1}{2}\left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^3$$

$$y(3) = \frac{1}{2}\delta(3) + \frac{1}{2}y(2) = 0 + \frac{1}{2}\left(\frac{1}{2}\right)^3 = \left(\frac{1}{2}\right)^4$$

$$\ldots.$$

$$y(k) = \left(\frac{1}{2}\right)^{k+1} u(k)$$

The impulse response for the filter is an exponential sequence. This is an IIR filter because the impulse response is of infinite duration.

EXAMPLE PROBLEM 11.23

Find the impulse response for the digital filter

$$y(k) = \frac{1}{3}x(k) + \frac{1}{3}x(k-1) + \frac{1}{3}x(k-2)$$

Solution

Assume the system is at rest before input begins—that is, $y(n) = 0$ for $n < 0$.

$$y(-2) = \frac{1}{3}\delta(-2) + \frac{1}{3}\delta(-3) + \frac{1}{3}\delta(-4) = 0 + 0 + 0 = 0$$

$$y(-1) = \frac{1}{3}\delta(-1) + \frac{1}{3}\delta(-2) + \frac{1}{3}\delta(-3) = 0 + 0 + 0 = 0$$

$$y(0) = \frac{1}{3}\delta(0) + \frac{1}{3}\delta(-1) + \frac{1}{3}\delta(-2) = \frac{1}{3} + 0 + 0 = \frac{1}{3}$$

$$y(1) = \frac{1}{3}\delta(1) + \frac{1}{3}\delta(0) + \frac{1}{3}\delta(-1) = 0 + \frac{1}{3} + 0 = \frac{1}{3}$$

$$y(2) = \frac{1}{3}\delta(2) + \frac{1}{3}\delta(1) + \frac{1}{3}\delta(0) = 0 + 0 + \frac{1}{3} = \frac{1}{3}$$

$$y(3) = \frac{1}{3}\delta(3) + \frac{1}{3}\delta(2) + \frac{1}{3}\delta(1) = 0 + 0 + 0 = 0$$

$$y(4) = \frac{1}{3}\delta(4) + \frac{1}{3}\delta(3) + \frac{1}{3}\delta(2) = 0 + 0 + 0 = 0$$

$$\cdots$$

$$y(k) = 0; k \geq 3$$

This is an FIR filter with only three nonzero coefficients.

IIR filters are particularly useful for simulating analog systems. The main advantage of an IIR filter is that the desired job can usually be accomplished with fewer filter coefficients than for an FIR filter; in other words, IIR filters tend to be more efficient. The main disadvantage of an IIR filter is that signals may be distorted in an undesirable way. FIR filters can be designed with symmetry to prevent undesired signal distortion. Methods for dealing with the distortion problem in FIR filters are outside our discussions here.

Digital filters, as the name implies, are most often designed to perform specific "filtering" operations: low-pass filters, high-pass filters, band-pass filters, bandstop filters, notch filters, and so on. However, digital filters can be used to simulate most analog systems—for example, to differentiate and to integrate. Many textbooks have been written on digital filter design. The key components of the process are illustrated following.

From Digital Filter to Transfer Function

The transfer function for the digital system, $H(z)$, can be obtained by rearranging the difference equation (Eq. (11.23)) and applying Eq. (11.21). $H(z)$ is the quotient of the z-transform of the output, $Y(z)$, divided by the z-transform of the input, $X(z)$.

$$y(k) + a_1 y(k-1) + a_2 y(k-2) \ldots + a_N y(k-N) = b_0 x(k) + b_1 x(k-1) + \ldots + b_M x(k-M)$$

$$Y(z) + a_1 z^{-1} \mathrm{Y(z)} + a_2 z^{-2} Y(z) \ldots + a_N z^{-N} Y(z) = b_0 X(z) + b_1 z^{-1} X(z) + b_2 z^{-2} X(z) \ldots . + b_M z^{-M} X(z)$$

$$Y(z)(1 + a_1 z^{-1} + a_2 z^{-2} \ldots + a_N z^{-N}) = X(z)(b_0 + b_1 z^{-1} + b_2 z^{-2} \ldots . + b_M z^{-M})$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-1} \ldots + b_M z^{-M}}{1 + a_1 z^{-1} + a_2 z^{-1} \ldots + a_N z^{-N}} \tag{11.45}$$

From Transfer Function to Frequency Response

The frequency response ($H'(\Omega)$) of a digital system can be calculated directly from $H(z)$, where Ω is in radians. If the data are samples of an analog signal as previously described, the relationship between ω and Ω is $\Omega = \omega \mathrm{T}$:

$$H'(\Omega) = H(z)|_{z=e^{j\Omega}} \tag{11.46}$$

For a linear system, an input sequence of the form

$$\mathrm{x(k)} = \mathrm{A}\sin(\Omega_0 \mathrm{k} + \Phi)$$

will generate an output whose steady-state sequence will fit into the following form:

$$y(k) = \mathrm{B}\sin(\Omega_0 k + \emptyset)$$

Values for B and $\emptyset$ can be calculated directly:

$$\mathrm{B} = A|H'(\Omega_0)|$$

$$\emptyset = \Phi + \mathrm{angle}(H'(\Omega_0))$$

EXAMPLE PROBLEM 11.24

The input sequence for the digital filter used in Example Problem 11.23 is

$$x(k) = 100\sin\left(\frac{\pi}{2}k\right)$$

What is the steady-state form of the output?

Solution

$$y(k) - \frac{1}{2}y(k-1) = \frac{1}{2}x(k)$$

The difference equation is first converted into the z-domain:

$$Y(z) - \frac{1}{2}Y(z)z^{-1} = Y(z)\left[1 - \frac{1}{2}z^{-1}\right] = \frac{1}{2}X(z).$$

Solving for $H(z)$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\frac{1}{2}}{1 - \frac{1}{2}z^{-1}}$$

gives the filter transfer function. To determine the output, the transfer function is evaluated at the frequency of the input sinusoid $(z = e^{j\frac{\pi}{2}})$

$$H'\left(\frac{\pi}{2}\right) = H\left(e^{j\frac{\pi}{2}}\right) = \frac{\frac{1}{2}}{1 - \frac{1}{2}e^{-j\frac{\pi}{2}}} = \frac{\frac{1}{2}}{1 + \frac{1}{2}j} = 0.4 - j0.2 = 0.45e^{-j0.15\pi}.$$

This transfer function tells us that the output is obtained by scaling the input magnitude by 0.45 and shifting the signal by a phase factor of 0.15π rads. Therefore, the output is

$$y(k) = 45\sin\left(\frac{\pi}{2}k - .15\pi\right)$$

Filter design problems begin with identifying the frequencies that are to be kept versus the frequencies that are to be removed from the signal. For ideal filters, $|H'(\Omega_{\text{keep}})| = 1$ and $|H'(\Omega_{\text{remove}})| = 0$. The filters in Example Problems 11.23 and 11.24 can both be considered as low-pass filters. However, their frequency responses (Figure 11.20) show that neither is a particularly good low-pass filter. An ideal low-pass filter that has a cutoff frequency of $\pi/4$ with $|H'(\Omega)| = 1$ for $\Omega < \pi/4$ and $|H'(\Omega)| = 0$ for $\pi/4$ and $|\Omega| < \pi$ is superimposed for comparison.

Windowed FIR Filter Design

The ideal filters in Section 11.6.4 provide a general framework from which to build a variety of filter functions to meet specific design criteria for both analog and discrete systems. Unfortunately, the ideal low-pass filter is not physically realizable, as we will see following. Using the ideal low-pass filter impulse response as a starting reference, we will develop a modified filter function that overcomes this limitation. The design of a windowed FIR filter is illustrated following for the case of a low-pass filter, but the same procedures and concepts apply for high-pass and band-pass filters.

There are two practical limitations associated with the ideal low-pass filter. First, note that the impulse response of the ideal low-pass filter has infinite duration, extending from $-\infty$ to $+\infty$. Thus, implementing an ideal low-pass filter requires an infinite amount of time. The simplest way to overcome this limitation is to truncate the impulse response over a finite time interval from –T to T (Figure 11.21a; shown for T = 0.1 sec). However, truncation leads to a second undesired effect. The sharp transitions in the impulse

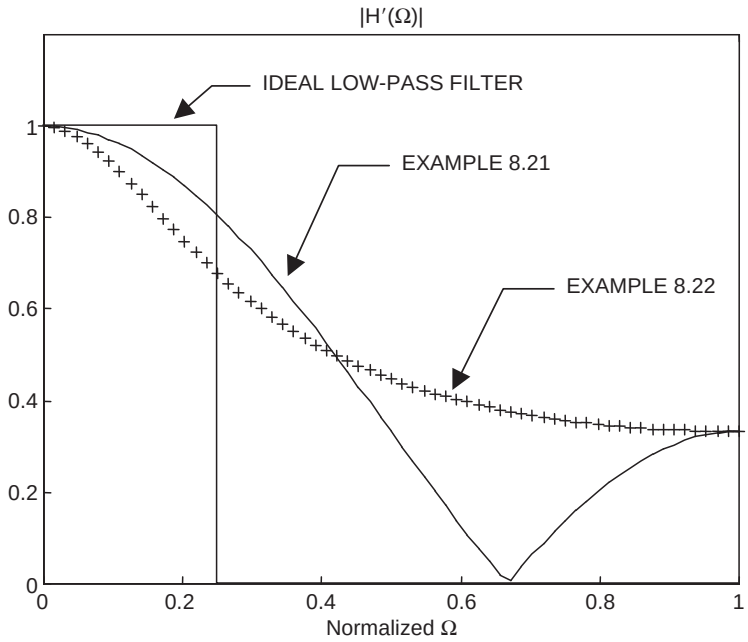


FIGURE 11.20 A frequency domain comparison of low-pass filters described in Example Problems 11.23 and 11.24. An ideal low-pass filter with a cutoff frequency at $\pi/4$ rads or 0.25 when normalized by π radians is superimposed for comparison. The cutoff frequency of a low-pass filter is usually defined as the frequency at which the amplitude is equal to $1/\sqrt{2}$ or approximately 0.71, which matches Example Problem 11.12. Both digital filters have the same amplitude at f_{max}—that is, where normalized $\Omega = 1$.

response to the selected interval do not allow for convergence of the Fourier integral (Gibbs Phenomena). In the frequency domain, the effects of the truncation can be seen in the transfer function magnitude of the truncated low-pass filter (Figure 11.21b). The truncated filter exhibits errors in the passband and stopband regions. These errors are undesirable because they will distort the signal in the passpand while allowing signals in the stopband from passing through.

One way to overcome the limitation is to gradually truncate the filter impulse response with a smooth window function, $w(t)$. The modified low-pass filter is expressed as the product of the ideal filter and the window function:

$$h(t) = h_{LP}(t)w(t) = \frac{W_c}{\pi}\text{sinc}(W_c t)w(t) \tag{11.47}$$

where $w(t)$ is restricted to the interval –T to T (Figure 11.21c). The windowing procedure is illustrated in Figure 11.21d, which shows the product of the window with the ideal filter impulse response. The window allows for a smooth truncation of the impulse response, thus allowing for convergence of the Fourier integral. As can be seen, the resulting stopband and passband errors of the windowed filter (Figure 11.21e) are substantially smaller than for the truncated ideal filter (Figure 11.21b). Details of the design of window functions

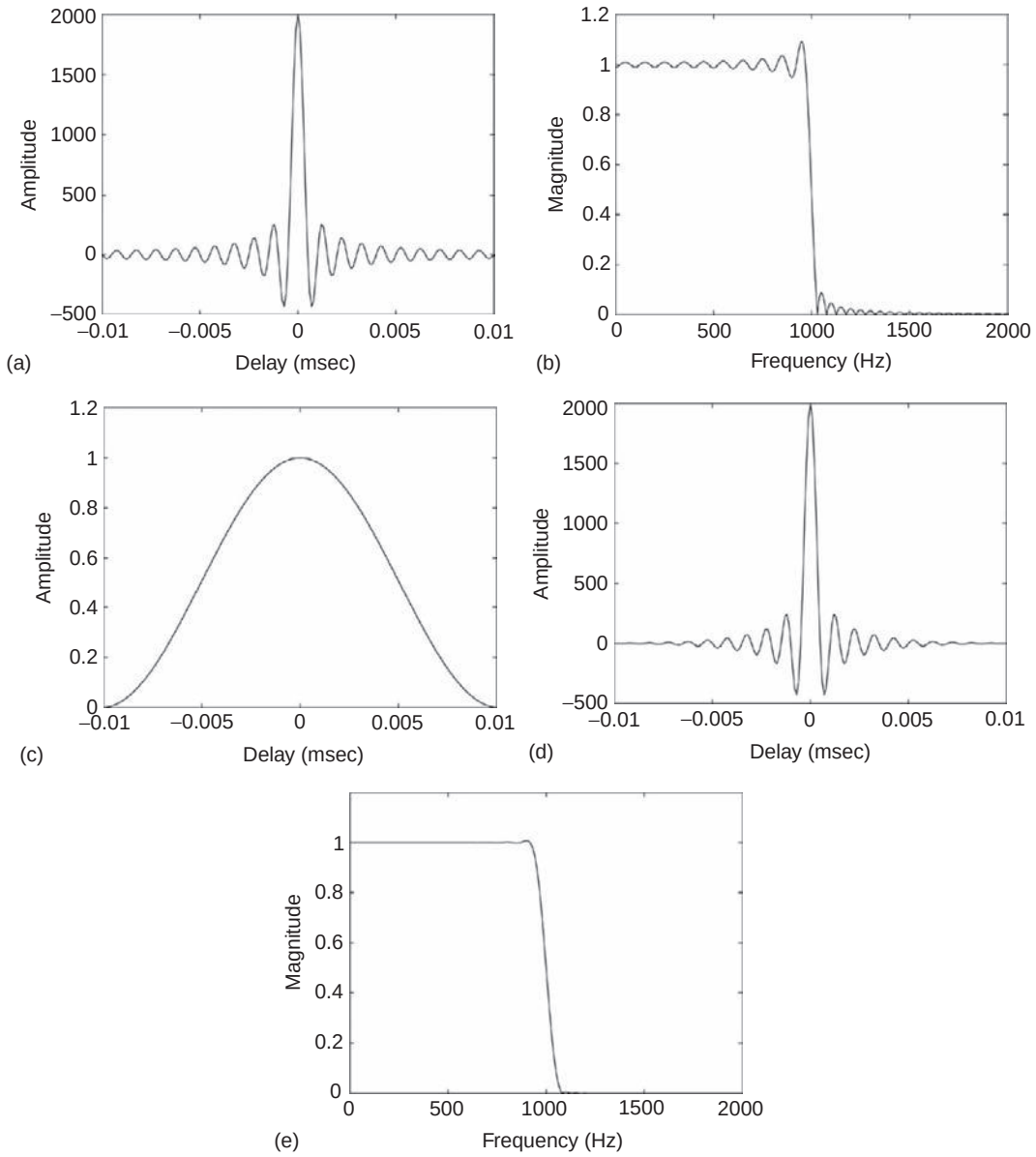


FIGURE 11.21 Filter design by windowing. (a) Truncated ideal low-pass filter impulse response. (b) Magnitude response for the ideal truncated filter. Note the errors in the passband and stopband regions. (c) Window function. (d) Windowed impulse response. (e) Magnitude response of the windowed filter exhibits substantially smaller passband and stopband errors.

are beyond the scope of this text and can be found in many signal processing textbooks. However, numerous window functions have been developed for a variety of applications. Some of the most celebrated window functions include the Kaiser, Hanning, and Hamming windows.

So far, we have illustrated the windowing method for a continuous filter, but we would like to apply this procedure to develop a discrete FIR filter. To achieve this, the continuous filter impulse response is sampled by allowing $t = k/F_s$, for integer k. The resulting discrete impulse response is expressed as

$$h[k] = \frac{W_c}{\pi}\text{sinc}\left(\frac{W_c}{F_s}k\right)w(k) = \frac{W_c}{\pi}\text{sinc}(\Omega_c k)w(k) \tag{11.48}$$

where $\Omega_c = W_c/F_s$ is the digital cutoff frequency of the filter. Note that according to the Nyquist sampling theorem, $0 < W_c < F_s/2$, and thus $0 < \Omega_c < \pi$.

EXAMPLE PROBLEM 11.25

In MATLAB, implement a 201 sample digital low-pass and high-pass filter using a Hanning window. The cutoff frequency of both filters is 1000 Hz. Plot the impulse response and transfer function magnitude.

Solution

```
Wc=2*pi*1000; %Cutoff Frequency in Radians / Sec
Fs=10000; %Sampling Rate in Hz
T=1/Fs;

%Ideal Filters
N=100; %Filter Order
t=(-N:N)/Fs; %Time Axis Sampled at Fs
h_lp=Wc/pi*sinc(1/pi*Wc*t); %Sampled Ideal Low-pass Filter
Impulse Response
delta=[zeros(1,N) Fs zeros(1,N)]; %Discrete Diract Impulse
Function
h_hp=(delta-h_lp); %Sampled Ideal High-pass Filter Impulse
Response

%Hanning Filters
W=hanning(2*N+1)'; %Hanning Window
h_lp=h_lp.*W; %Hanning Low-pass Filter Impulse Response
h_hp=h_hp.*W; %Hanning High-pass Filter Impulse Response
NFFT=1024*8; %FFT number of samples
faxis=(0:NFFT-1)/NFFT*Fs; %Frequency Axis
H_lp=abs(fft(h_lp,NFFT))/Fs; %Hanning Low-pass Filter Transfer
Function Magnitude
H_hp=abs(fft(h_hp,NFFT))/Fs; %Hanning High-pass Filter Transfer
Function Magnitude
```

```
%Plotting Results
subplot(221)
plot(t,h_lp,'k')
xlabel('Time (sec)')
ylabel('Amplitude')

subplot(222)
plot(faxis,H_lp,'k');
axis([0 Fs/2 0 1.2])
xlabel('Frequency (Hz)')
ylabel('Magnitude')

subplot(223)
plot(t,h_hp,'k')
xlabel('Time (sec)')
ylabel('Amplitude')

subplot(224)
plot(faxis,H_hp,'k');
axis([0 Fs/2 0 1.2])
xlabel('Frequency (Hz)')
ylabel('Magnitude')
```

The results are shown in Figure 11.22.

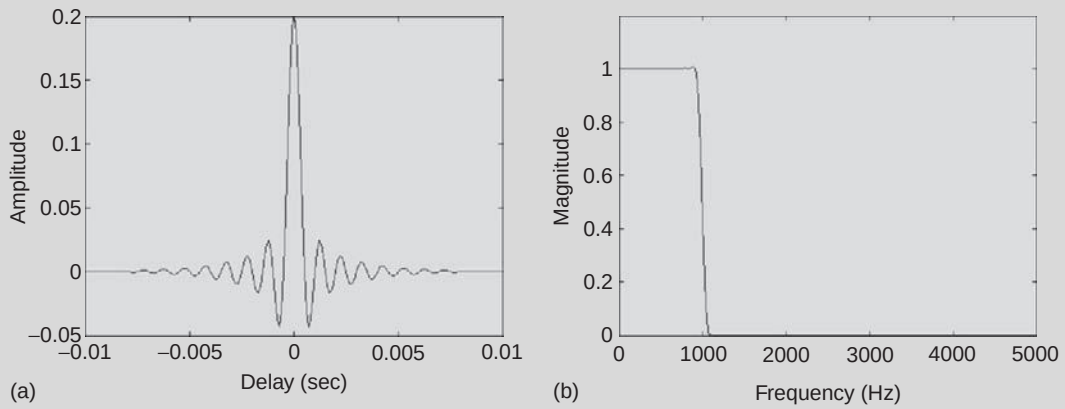


FIGURE 11.22 Low-pass and high-pass FIR filter MATLAB simulation. (a) Hanning low-pass filter impulse response and magnitude response (b).

Continued

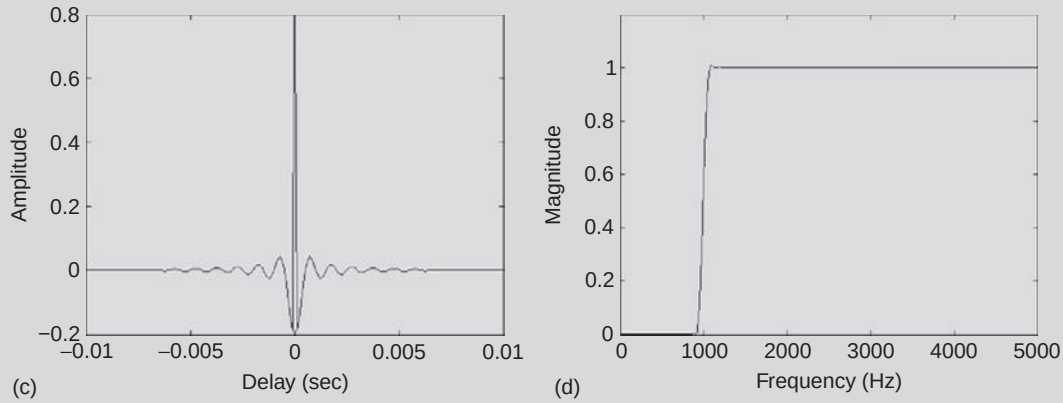


FIGURE 11.22, cont'd (c) Hanning high-pass filter impulse response and magnitude response (d).

EXAMPLE PROBLEM 11.26

Using the digital low-pass and high-pass filters of Example Problem 11.25, filter a white noise (flat spectrum) signal. Plot the magnitude spectrum of the input and output signals.

Solution

```
%Filtering the Random Noise Signal
X=randn(1,Fs); %1 second of Random Noise
Y_lp=conv(X,h_lp); %Low-pass Filtered Noise
Y_hp=conv(X,h_hp); %High-pass Filtered Noise

%Plotting Results
subplot(221)
psd(X,1024,Fs); %Input Power spectrum Magnitude
hold on
psd(Y_lp,1024,Fs); %Low-pass Output Power spectrum Magnitude
ch=get(gca,'children')
set(ch(1),'color','k')
set(ch(2),'color',[0.5 0.5 0.5])

subplot(222)
psd(X,1024,Fs); %Input Power spectrum Magnitude
hold on
psd(Y_hp,1024,Fs); %High-pass Output Power spectrum Magnitude
ch=get(gca,'children')
ch=get(gca,'children')
set(ch(1),'color','k')
set(ch(2),'color',[0.5 0.5 0.5])
```

The results are shown in Figure 11.23.

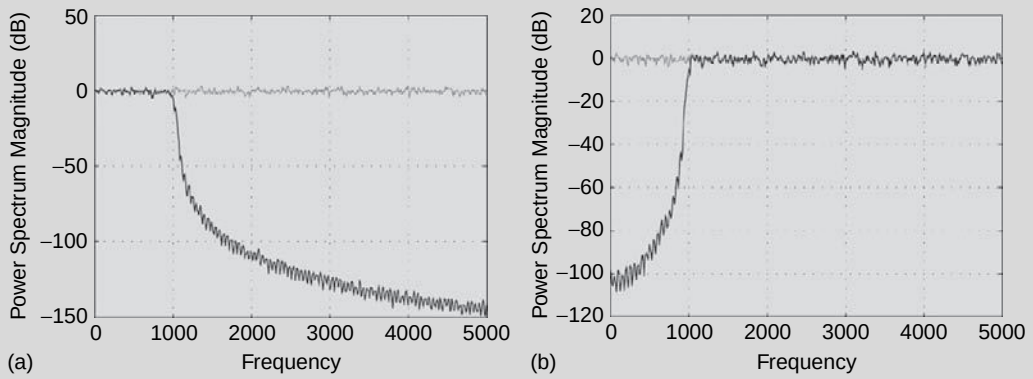


FIGURE 11.23 Low-pass and high-pass filtered noise signal. (a and b) Gray lines show the power spectrum of the input signal. (a) Low-pass filtered noise power spectrum is shown as black. (b) High-pass filtered noise power spectrum is shown as black.

11.7 SIGNAL AVERAGING

Biological signal measurements are often confounded by measurement noise. Variability in the measurement of a signal often makes it difficult to determine the signal characteristics, making it nearly impossible to obtain a reliable clinical diagnosis.

Many classes of biological signals are modeled as the sum of an ideal noiseless signal component, $x(t)$, and separate independent noise term, $n(t)$:

$$x_i(t) = x(t) + n_i(t). \tag{11.49}$$

The signal, $x_i(t)$, corresponds to the "measured" i-th trial or i-th measurement of the signal. Note that the i-th measurement contains both a deterministic component, $x(t)$, and a random or stochastic noise term, $n_i(t)$. Although the deterministic component of the signal is fixed from trial to trial, the noise term represents intrinsic variability, which may arise from a number of separate sources. The i-th measurement can therefore exhibit significant trial-to-trial variability because the random component, $n_i(t)$, is different across consecutive trials. As an example, a measurement ECG (electrocardiogram) electrode can pick up extraneous signals from the muscles, lungs, and even the internal electronics of the recording devices (e.g., 60 cycle noise from the power supply). The activity of these signals is unrelated to the activity of the beating heart, and it therefore shows up in the signal measurement as noise. Other unpredictable changes in the activity of the heartbeat, such as from the caffeine jolt after drinking an espresso, could also show up in a measurement and be interpreted as noise to an uninformed observer.

We have already examined one possible way to separate out the signal term from the noise term by filtering the signal with an appropriately designed filter. Appropriate filtering allows one to clean up the signal, thus improving the quality of signal and the diagnostic reliability in clinical settings. If the spectrum of the noise and signal components do not overlap in the frequency domain, one can simply design a filter that keeps or enhances the desired signal term, $x(t)$, and discards the unwanted noise term, $n_i(t)$. While this is a simple and useful way of cleaning up a signal, this approach does not work in many instances because the biological signal and noise spectrums overlap.

Many biological signals are approximately periodic in nature. Signals associated with the beating heart—blood pressure, blood velocity, electrocardiogram (ECG)—fall into this category. However, due to intrinsic natural variability, noise, and/or the influence of other functions, such as respiration, beat-to-beat differences are to be expected. Figure 11.2 is an example of a blood pressure signal that has all of the described variability.

Blood pressure signals have many features that clinicians and researchers use to determine a patient's health. Some variables that are often measured include the peak pressure while the heart is ejecting blood (systolic phase), the minimum pressure achieved while the aortic valve is closed (diastolic phase), the peak derivative (dP/dt) during the early part of the systolic phase (considered an indication of the strength of the heart), and the time constant of the exponential decay during diastole (a function of the resistance and compliance of the blood vessels). One way to determine variables of interest is to calculate the variables or parameters for each beat in a series of beats and then report the means. This is often not possible because noise from individual measurements makes it very difficult to accurately determine the relevant biological parameters. An alternate approach is to first average the signal measurements from separate trials,

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t), \tag{11.50}$$

such that a representative beat is obtained. If the signal is discrete, this average is represented by

$$\bar{x}(k) = \frac{1}{N}\sum_{i=1}^{N} x_i(k) \tag{11.51}$$

Here, $x_i(t)$, or $x_i(k)$ for the discrete case, represents the i-th measured heartbeat signal out of a total of N measurements. The signal $\bar{x}(t)$, $\bar{x}(k)$ for the discrete case, represents the mean or average waveform obtained following the averaging procedure. Substituting Eq. (11.50) into (11.49) leads to

$$\bar{x}(t) = x(t) + \frac{1}{N}\sum_{i=1}^{N} n_i(t) = x(t) + \varepsilon(t). \tag{11.52}$$

If the noise term, $n_i(t)$, is random and independent from trial-to-trial, it can be shown that the measurement error term in Eq. (11.52), $\varepsilon(t)$, which contains the influence of the noise, approaches 0 as $N \to \infty$. Thus, $\bar{x}(t) \approx x(t)$ for very large N, where $\varepsilon(t)$ tends to be small. This is a very powerful result! It tells us that we can effectively remove the noise by simply averaging measurements from many trials. Essentially, if we average a sufficiently large number of signal trials, the averaged signal closely approximates the true noiseless signal waveform.

Many biological acquisition systems are designed to calculate signal averages (Eqs. (11.50) and (11.51)) as data are collected. The summation process is triggered by a signal or a signal-related feature. The ECG signal, which has many sharp features, is often used for heartbeat-related data. Figure 11.24 shows a signal-averaged blood pressure waveform for the data shown in Figure 11.2. Figure 11.25 shows the signal-averaging procedure for an auditory brainstem response (ABR) EEG measurement.

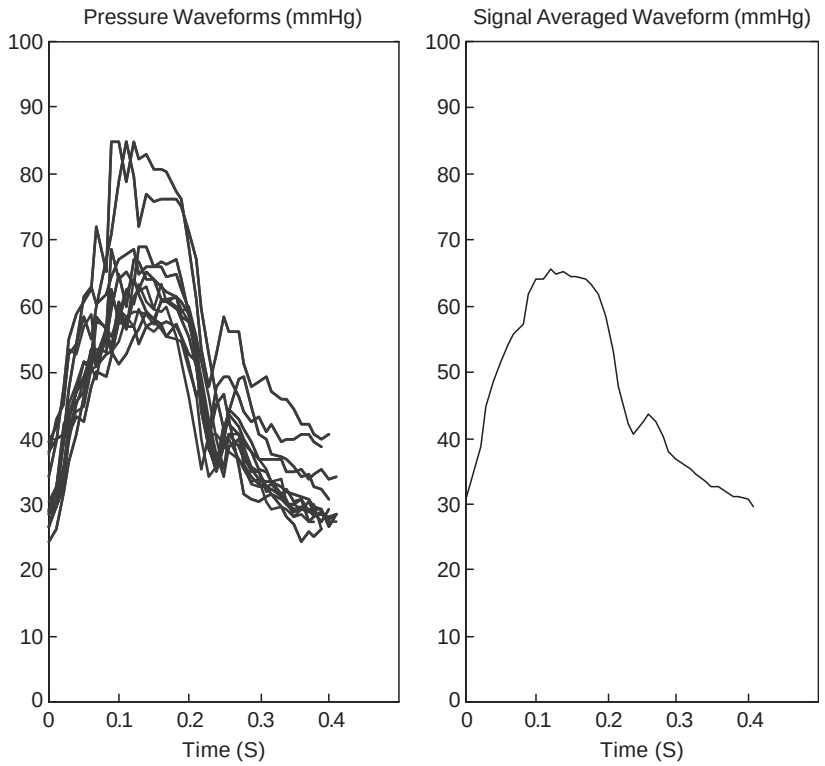


FIGURE 11.24 A signal-averaged pressure waveform for the data shown in Figure 11.2.

In both cases, note that the averaging helps to preserve the relevant signal features and remove undesirable noise disturbances.

The blood pressure and ABR examples illustrate signal averaging in the time domain. For signals that are random in nature, signal averaging in the frequency domain is sometimes preferable. Figure 11.26 shows an EEG signal sampled over the occipital lobe of a patient. The sampling rate was 16 kHz. EEG analysis is usually done in the frequency domain, since the presence of different frequencies is indicative of different brain states, such as sleeping, resting, alertness, and so on. The power at each frequency estimate, which can be approximated by the square of the Fourier transform, is the measurement of choice.

If a DFT is performed on the data to estimate the power of the frequencies in the signal, the expected noise in the measurement is of the same size as the measurement itself. To reduce the noise variance, a statistical approach must be undertaken. One popular approach is known as the "Welch" or "periodogram averaging" method. The signal is broken into L sections (disjoint if possible) of N points each. A DFT is performed on each of the L sections. The final result for the N frequencies is then the average at each frequency for the L sections.

The N data points in the i-th segment are denoted as

$$x_i(k) = x(k + (i-1)N) \quad 0 \le k \le N-1, 1 \le i \le L$$

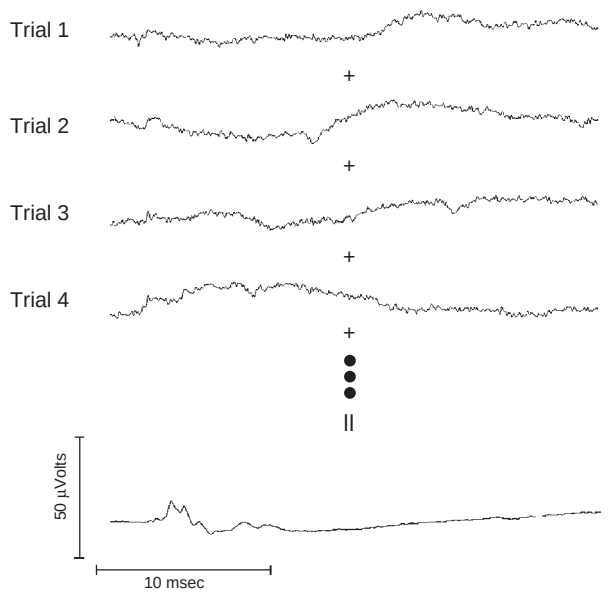


FIGURE 11.25 Single trials from an auditory evoked response to a brief sound pulse (at time zero) were measured on the temporal lobe. The auditory response from individual trials is obscured by random noise (shown first 4 out of 1,000). Averaged response of 1,000 trials reveals the auditory response component (bottom trace).

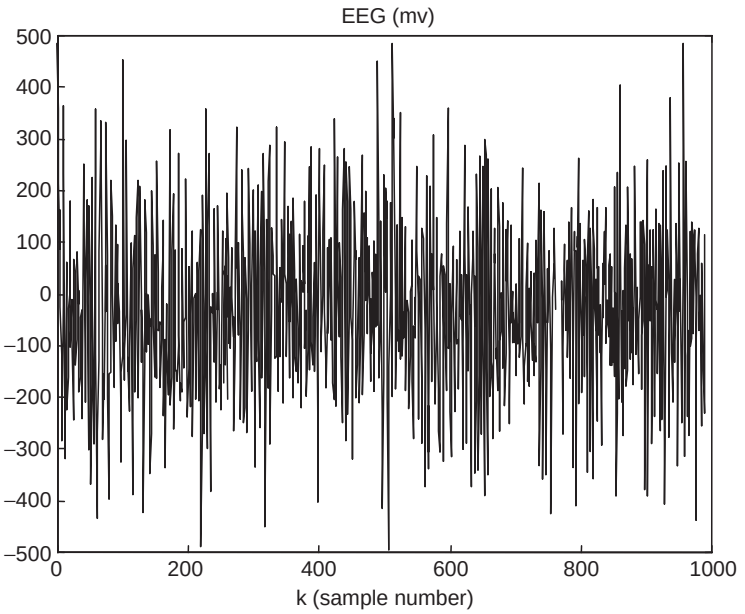


FIGURE 11.26 An EEG signal containing 1,000 samples sampled at 16 kHz from the occipital.

if the segments are consecutive and disjoint. The power estimate based on the DFT of an individual segment "i" is

$$\hat{P}_i(m) = \frac{1}{N}\left|\sum_{k=0}^{N-1} x_i(k)e^{-j\frac{2\pi mk}{N}}\right|^2 \text{ for } 0 \le m \le N-1 \tag{11.53}$$

where m is associated with the power at a frequency of $\Omega = 2\pi m/N$ radians. The averaged signal spectrum is calculated by taking the mean at each frequency

$$\hat{P}(m) = \frac{1}{L}\sum_{i=1}^{L} \hat{P}_i(m) \tag{11.54}$$

The selection of N is very important, since N determines the resolution in the frequency domain. For example, if data are sampled at 500 samples/s and the resolution is desired at the 1 Hz level, at least 1 second or 500 samples ($N = 500$) should be included in each of the L sections. If resolution at the 10 Hz level is sufficient, only 0.1 seconds or 50 data points need to be included in each section. Note that the reduction in the frequency resolution leads to improvement in the certainty of the measurement because one can now average more signal segments. The averaging procedure decreases the variance by a factor of $1/L$. This averaging process is demonstrated for the EEG data in Figure 11.27. Modifications to

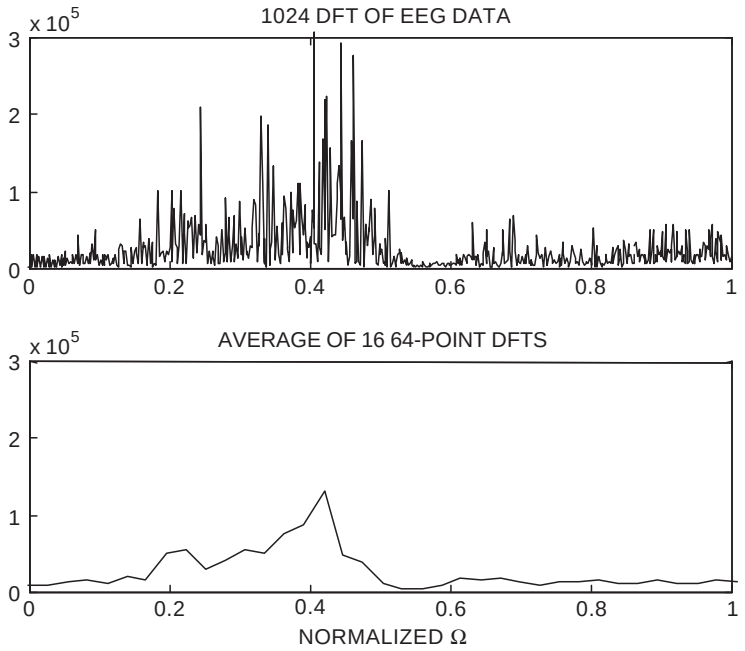


FIGURE 11.27 DFT averaging of an EEG. The top trace shows the raw DFT. The bottom trace shows the periodogram-averaged DFT obtained with 16 64-point segments of the data.

the procedure may include using overlapping segments if a larger value for L is needed and the number of available data points is not sufficient and/or multiplying each section by a window that forces continuity at the end points of the segments.

EXAMPLE PROBLEM 11.27

Consider the sinusoid signal

$$x(k) = \sin(\pi/4k) + n(k)$$

that is corrupted by random noise, $n(k)$. Using MATLAB, show that averaging the signal removes the noise component and reveals the deterministic component. Show results for 1, 10, and 100 averages.

Solution

```
k=1:64; %Discrete Time Axis
for i=1:100 %Generating 100 signal Trials
    x(i,:)=sin(pi/4*k)+randn(1,64); %i-th trial
end
X1=x(1,:); %1 Averages
X10=mean(x(1:10,:)); %10 Averages
X100=mean(x); %100 Averages

subplot(311) %Plotting Results, 1 Average
plot(k,X1,'k')
axis([1 64 -3 3])
title('1 Average')
ylabel('Amplitude')

subplot(312) %Plotting Results, 10 Averages
plot(k,X10,'k')
axis([1 64 -3 3])
title('10 Averages')
ylabel('Amplitude')

subplot(313) %Plotting Results, 100 Averages
plot(k,X100,'k')
axis([1 64 -3 3])
title('100 Averages')
xlabel('Discrete Time')
ylabel('Amplitude')
```

The results are shown in Figure 11.28.

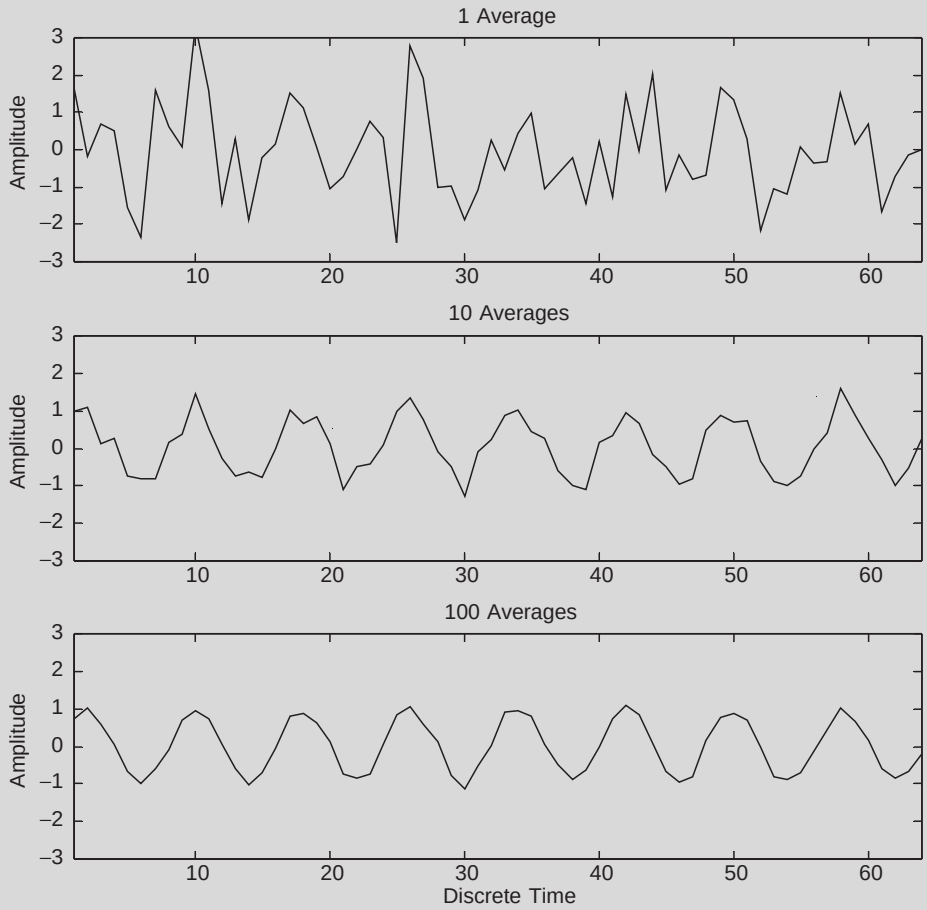


FIGURE 11.28 MATLAB results showing noise removal by averaging a noisy sinusoid signal. Shown for 1, 10, and 100 averages.

11.8 THE WAVELET TRANSFORM AND THE SHORT-TIME FOURIER TRANSFORM

The Fourier transform (Eq. (11.6)) is a well-known signal processing tool for breaking a signal into constituent sinusoidal waveforms of different frequencies. For many applications, particularly those that change little over time, knowledge of the overall frequency content may be all that is desired. The Fourier transform, however, does not delineate how a signal changes over time.

The short-time Fourier transform (STFT) and the wavelet transform (WT) have been designed to help preserve the time-domain information. The STFT approach is to perform a Fourier transform on only a small section (window) of data at a time, thus mapping

the signal into a 2D function of time and frequency. The transform is described mathematically as

$$X(\omega, a) = \int_{-\infty}^{\infty} x(t)g(t-a)e^{-j\omega t}dt \tag{11.55}$$

where $g(t)$ may define a simple box or pulse function. The inverse of the STFT is given as

$$X(\omega, a) = K_g \int\int X(\omega, a)g(t-a)e^{j\omega t}dtda \tag{11.56}$$

where K_g is a function of the window used.

To avoid the "boxcar" or "rippling" effect associated with a sharp window, the window may be modified to have more gradually tapered sides. Both designs are shown in Figure 11.29. The windows are superimposed on a totally periodic aortic pressure signal. For clarity, the windows have been multiplied by a factor of 100.

The STFT amplitudes for three box window sizes, ½ period, 1 period, and 2 periods, are illustrated in Figure 11.30. The vertical lines in the top figure are indicative of longer periodicities than the window. The solid-colored horizontal lines in the bottom two figures indicate that the frequency content is totally independent of time at that window size. This is expected, since the window includes either one or two perfect periods. The dark (little or no frequency content) horizontal lines interspersed with the light lines in the bottom figure indicate that multiple periods exist within the window.

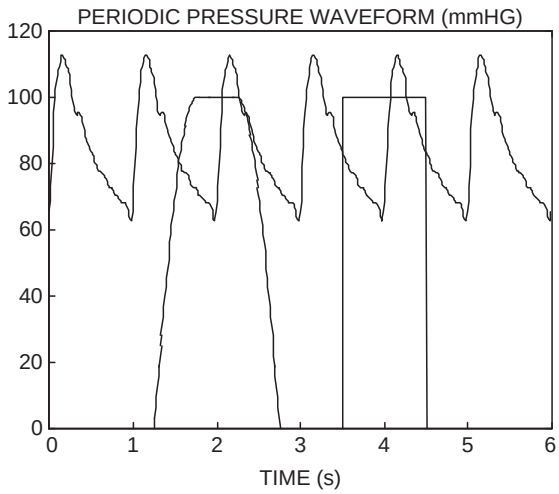


FIGURE 11.29 An example of two windows that might be used to perform an STFT on a perfectly periodic aortic pressure waveform. Each window approximates the width of one pulse. The tapered window on the left can help avoid the "boxcar" or "rippling" effect associated with the sharp window on the right. For clarity, the windows have been multiplied by a factor of 100.

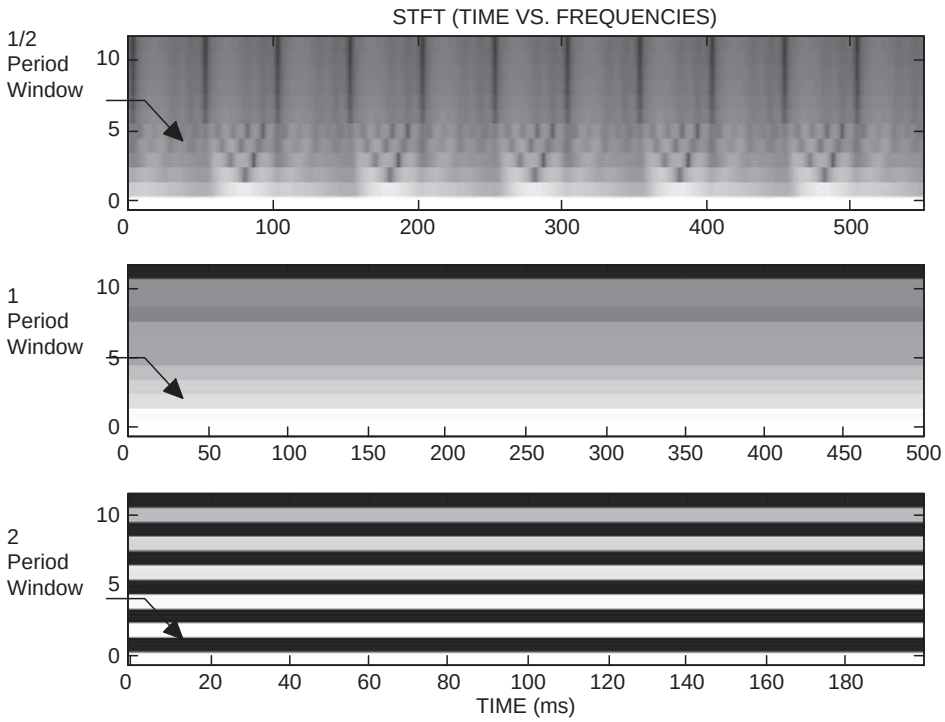


FIGURE 11.30 A two-dimensional rendering of the STFT amplitude coefficients for three box window sizes—½ period, 1 period, and 2 periods—applied to the perfectly periodic data shown in Figure 11.17. The lighter the color, the larger the amplitude. For example, the 0th row corresponds to the mean term of the transform, which is the largest in all cases. Higher rows correspond to harmonics of the data, which, in general, decrease with frequency. The vertical lines in the top figure are indicative of longer periodicities than the window. The solid-colored horizontal lines in the bottom two figures indicate that the frequency content is totally independent of time at that window size. The dark (little or no frequency content) horizontal lines interspersed with the light lines in the bottom figure indicate that multiple periods exist within the window.

In contrast, Figure 11.31 shows an amplitude STFT spectrum for the aperiodic pressure waveform shown in Figure 11.2, with the window size matched as closely as possible to the heart rate. The mean has been removed from the signal so the variation in the lowest frequencies (frequency level 0) reflects changes with respiration. The level of the heart rate (level 1) is most consistent across time, and the variability increases with frequency.

The main disadvantage of the STFT is that the width of the window remains fixed throughout the analysis. Wavelet analysis represents a change from both the FT and STFT in that the constituent signals are no longer required to be sinusoidal and the windows are no longer of fixed length. In wavelet analysis, the signals are broken up into shifted and scaled versions of the original or "mother" wavelet, $\psi(t)$. Figure 11.32 shows examples of two wavelets: the Haar on the left and one from the Daubechies (db2) series on the right. Conceptually, these mother wavelet functions are analogous to the impulse response of a band-pass filter. The sharp corners enable the transform to match up with local details that are not possible to observe using a Fourier transform.

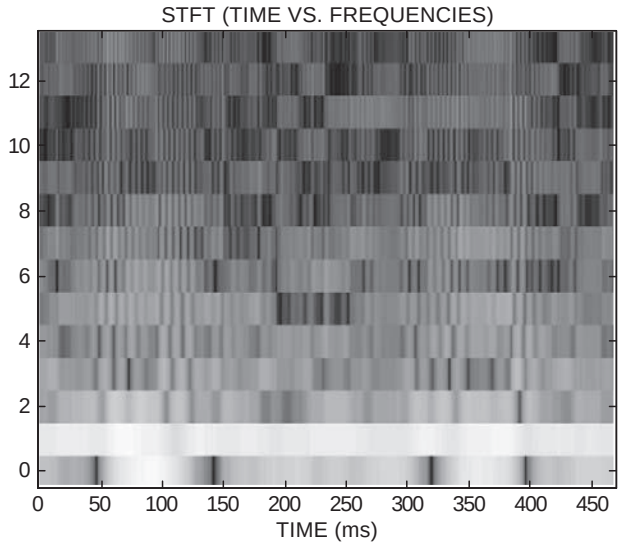


FIGURE 11.31 A two-dimensional rendering of the STFT of the aperiodic aortic pressure tracing shown in Figure 11.2. The window size was matched as closely as possible to the heart rate. The mean was removed from the signal so the variation in the lowest frequencies—that is, frequency level 0—reflects changes with respiration.

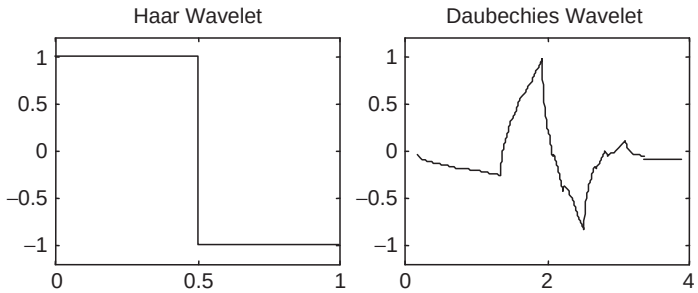


FIGURE 11.32 The general shape of two wavelets commonly used in wavelet analysis. The sharp corners enable the transform to match up with local details not possible to observe when using a Fourier transform that matches only sinusoidal shapes.

The notation for the 2D WT is

$$C(a,s) = \int_{-\infty}^{\infty} x(t)\varphi(a,s,t)dt \tag{11.57}$$

where a = scale factor and s = the position factor. C can be interpreted as the correlation coefficient between the scaled, shifted wavelet and the data. Figure 11.33 shows the db2 (φ(t)) wavelet at different scales and positions—for example, φ (2,–100,t) = φ (2t-100). The inverse wavelet transform

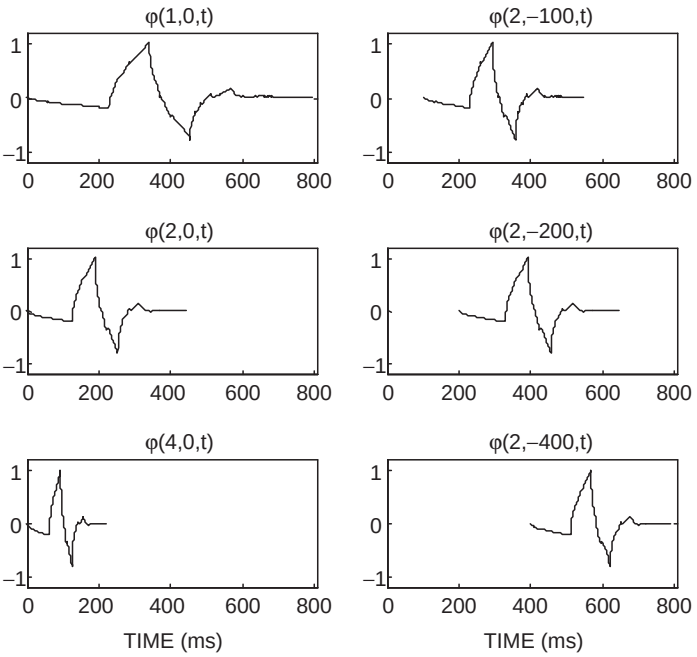


FIGURE 11.33 Illustrations of the db2 wavelet at several scales and positions. The upper left-hand corner illustrates the basic waveform φ (t). The notation for the illustrations is given in the form φ (scale, delay, t). Thus, φ (t) = φ (1,0,t), φ (2t-100) = φ (2,–100,t), and so on.

$$x(t) = K_{\varphi} \int\int C(a,s)\varphi(a,s,t)dtds \tag{11.58}$$

can be used to recover the original signal, $x(t)$, from the wavelet coefficients, $C(a, s)$. $K\varphi$ is a function of the wavelet used.

In practice, wavelet analysis is performed on digitized signals using a subset of scales and positions (see MATLAB's Wavelet Toolbox). One computational process is to recursively break the signal into low-frequency ("high-scale" or "approximation") and high-frequency ("low-scale" or "detail") components using digital low-pass and high-pass filters that are functions of the mother wavelet. The output of each filter will have the same number of points as the input. In order to keep the total number of data points the same at each level, every other data point of the output sequences is discarded. This process is known as "downsampling." Using "upsampling" and a second set of digital filters, called reconstruction filters, the process can be reversed, and the original data set is reconstructed. Remarkably, the inverse discrete wavelet transform does exist!

While this process will rapidly yield wavelet transform coefficients, the power of discrete wavelet analysis lies in its ability to examine waveform shapes at different resolutions and to selectively reconstruct waveforms using only the levels of approximation and detail that are desired. Applications include detecting discontinuities and breakdown points, detecting long-term evolution, detecting self-similarity (e.g., fractal trees), identifying pure

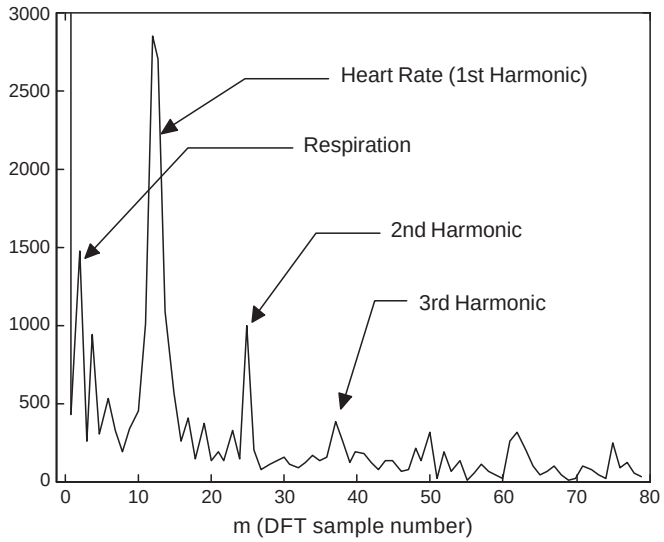


FIGURE 11.34 DFT of pressure data from Figure 11.2. The first, second, and third harmonics of the heart rate are clearly visible.

frequencies (similar to Fourier transform), and suppressing, denoising, and/or compressing signals.

For comparison purposes, discrete Fourier transforms and discrete wavelet transforms are illustrated for the pressure waveforms shown in Figure 11.2. Figure 11.34 shows details of the DFT on the entire record of data. The beat-to-beat differences are reflected by the widened and irregular values around the harmonics of the heart rate. The respiration influence is apparent at the very low frequencies.

Finally, an example from the MATLAB Wavelet Toolbox is shown that uses the same pressure waveform. Figure 11.35 is a 2D rendering of the wavelet transform coefficients. The *x*-axis shows the positions and the *y*-axis shows the scales, with the low scales on the bottom and the high scales on the top. The top scale clearly shows the two respiratory cycles in the signal. More informative than the transform coefficients, however, is a selective sample of the signal details and approximations. As the scale is changed from a1 to a7, the approximation goes from emphasizing the heart rate components to representing the respiration components. The details show that the noise at the heart rate levels is fairly random at the lower scales but moves to being quite regular as the heart rate data become the noise (Figure 11.36).

11.9 ARTIFICIAL INTELLIGENCE TECHNIQUES

Artificial intelligence (AI) is a broad field that focuses on the application of computer systems that exhibit intelligent capabilities. AI systems can be built from a number of separate technologies, including fuzzy logic, neural networks, and expert systems. The principal aim of AI is to create intelligent machines that can function under adverse and unpredictable circumstances. The term *intelligent* as it applies to machines indicates

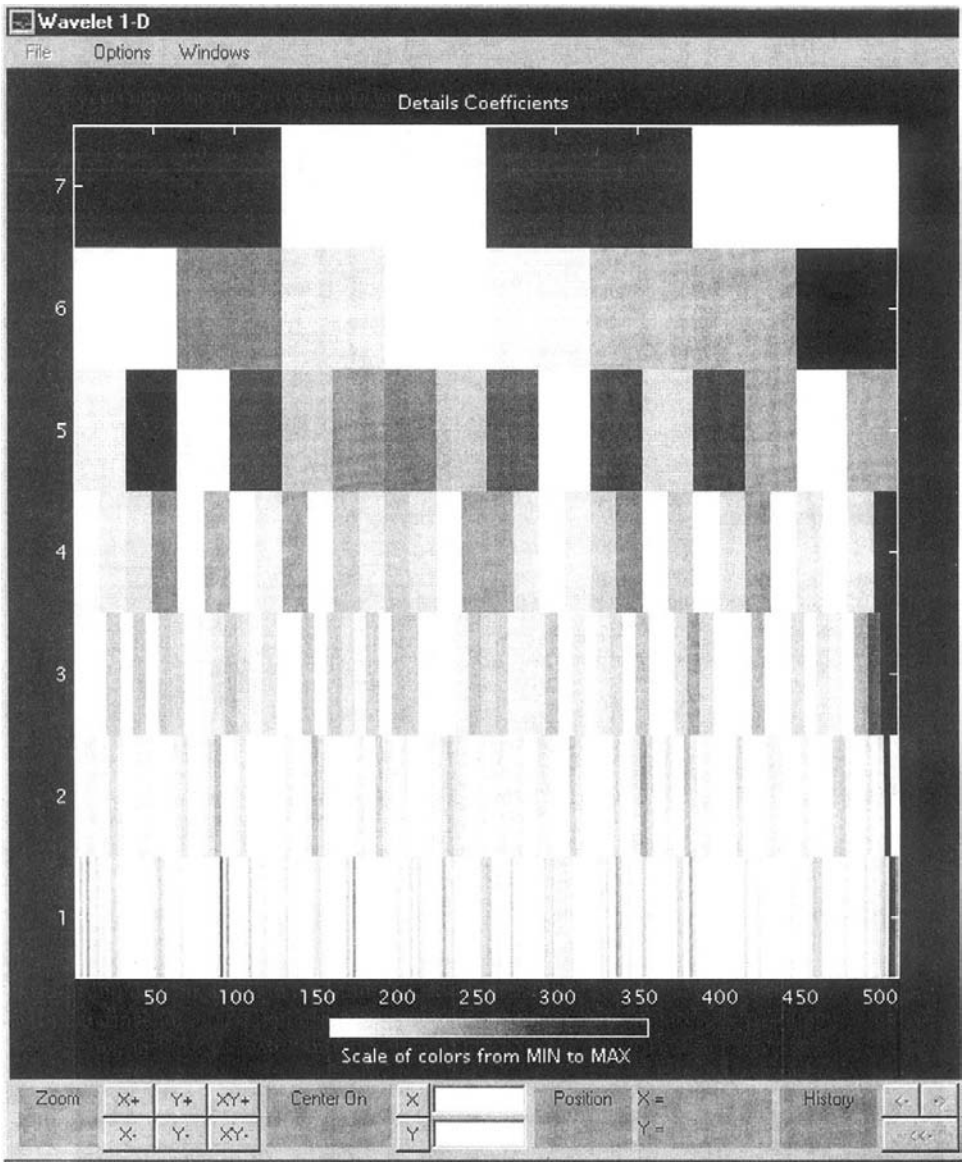


FIGURE 11.35 MATLAB was used to produce a two-dimensional rendering of the wavelet transform coefficients, with the Daubechies wavelet applied to the aortic pressure tracing in Figure 11.2. The *x*-axis shows the positions and the *y*-axis shows the scales with the low scales on the bottom and the high scales on the top. The associate waveforms at selected levels of these scales are shown in Figure 11.33.

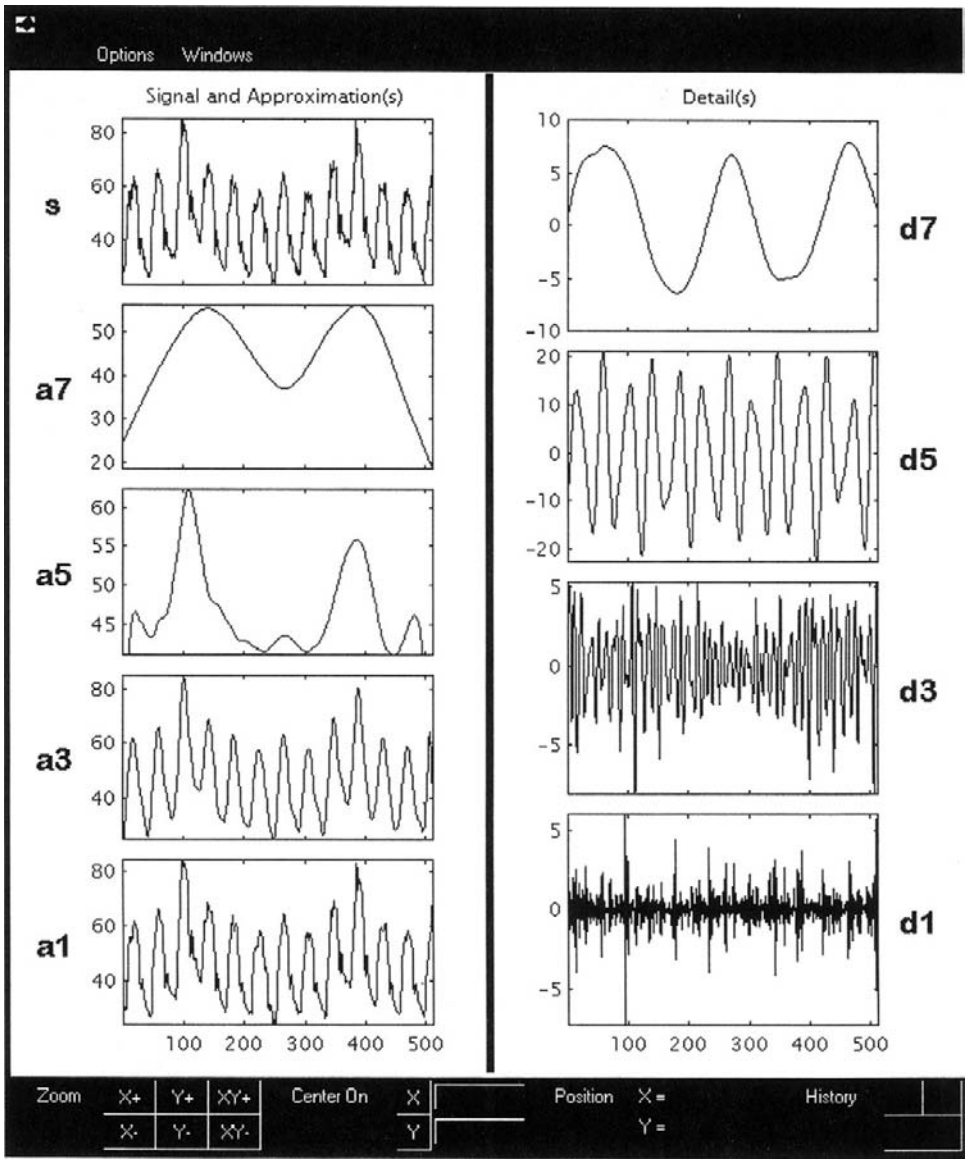


FIGURE 11.36 A selective sample of the signal details and approximations generated by MATLAB as part of the wavelet transform process.

computer-based systems that can interact with their environment and adapt to changes in the environment. The adaptation is accomplished through self-awareness and perceived models of the environment that are based on qualitative and quantitative information. In other words, the basic goal of AI techniques is to produce machines that are more capable of human-like reasoning, decision making, and adaptation.

The machine intelligence quotient (MIQ) is a measure of the intelligence level of machines. The higher a machine's MIQ, the higher the capacity of the machine for automatic reasoning and decision making. The MIQ of a wide variety of machines has risen significantly during the past few years. Many computer-based consumer products, industrial machinery, and biomedical instruments and systems are using more sophisticated artificial intelligence techniques. Advancements in the development of fuzzy logic, neural networks, and other soft computing techniques have contributed significantly to the improvement of the MIQ of many machines.

Soft computing is an alliance of complementary computing methodologies. These methodologies include fuzzy logic, neural networks, probabilistic reasoning, and genetic algorithms. Various types of soft computing often can be used synergistically to produce superior intelligent systems. The primary aim of soft computing is to allow for imprecision, since many of the parameters that machines must evaluate do not have precise numeric values. Parameters of biological systems can be especially difficult to measure and evaluate precisely.

11.9.1 Fuzzy Logic

Fuzzy logic is based on the concept of using words rather than numbers for computing, since words tend to be much less precise than numbers. Computing has traditionally involved calculations that use precise numerical values, while human reasoning generally uses words. Fuzzy logic attempts to approximate human reasoning by using linguistic variables. Linguistic variables are words that are used to describe a parameter. For body temperature, linguistic variables that might be used are high fever, above normal, normal, below normal, and frozen. The linguistic variables are more ambiguous than the number of degrees Fahrenheit, such as 105.0, 98.9, 98.6, 97.0, and 27.5.

In classical mathematics, numeric sets called crisp sets are defined, while the basic elements of fuzzy systems are fuzzy sets. An example of a crisp set is A = [0, 20]. Crisp sets have precisely defined, numeric boundaries. Fuzzy sets do not have sharply defined bounds. Consider the categorization of people by age. Using crisp sets, the age groups could be divided as A = [0, 20], B = [30, 50], and C = [60, 80]. Figure 11.37a shows the characteristic function for the sets A, B, and C. The value of the function is either 0 or 1, depending on whether the age of a person is within the bounds of set A, B, or C. The scheme using crisp sets lacks flexibility. If a person is 25 years old or 37 years old, he or she is not categorized.

If the age groups were instead divided into fuzzy sets, the precise divisions between the age groups would no longer exist. Linguistic variables, such as young, middle-aged, and old, could be used to classify the individuals. Figure 11.37b shows the fuzzy sets for age categorization. Note the overlap between the categories. The words are basic descriptors, not precise measurements. A 30-year-old woman may seem old to a 6-year-old boy but quite young to an 80-year-old man. For the fuzzy sets, a value of 1 represents a 100 percent degree of membership to a set. A value of 0 indicates that there is no membership in the set. All numbers between 0 and 1 show the degree of membership to a group. A 35-year-old person, for instance, belongs 50 percent to the young set and 50 percent to the middle-aged set.

As with crisp sets from classical mathematics, operations are also defined for fuzzy sets. The fuzzy set operation of intersection is shown in Figure 11.38a. Figure 11.38b shows the

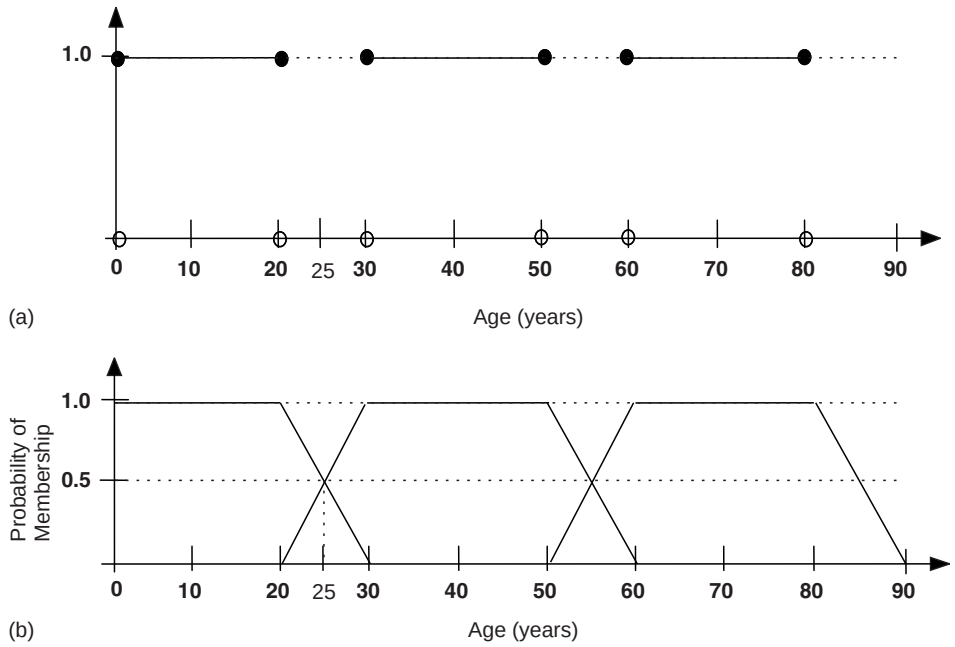


FIGURE 11.37 (a) Crisp sets for the classification of people by age. (b) Fuzzy sets for the classification of people by age.

fuzzy union operator, and Figure 11.38c shows the negation operator for fuzzy sets. The solid line indicates the result of the operator in each figure.

Although it is easy to form fuzzy sets for a simple example such as age classification, fuzzy sets for more sophisticated applications are derived by using sophisticated calibration techniques. The linguistic variables are formulated mathematically and then can be processed by computers. Once the fuzzy sets have been established, rules are constructed. Fuzzy logic is a rule-based logic. Fuzzy systems are constructed by using a large number of rules. Most rules used in fuzzy logic computing are if-then statements that use linguistic variables. Two simple rules that use the fuzzy sets for age classification might be the following:

If the patient is YOUNG, then use TREATMENT A.
If the patient is MIDDLE-AGED or OLD, then use TREATMENT B.

The degree of membership in a group helps to determine which rule will be used and, consequently, the type of action that will be taken or, in the preceding example, the sort of treatment that will be used. Defuzzification methods are used to determine which rules will be used to produce the final output of the fuzzy system.

For many applications, fuzzy logic has significant advantages over traditional numeric computing methods. Fuzzy logic is particularly useful when information is too limited or too complex to allow for numeric precision, since it tolerates imprecision. If an accurate mathematical model cannot be constructed, fuzzy logic may prove valuable. However, if

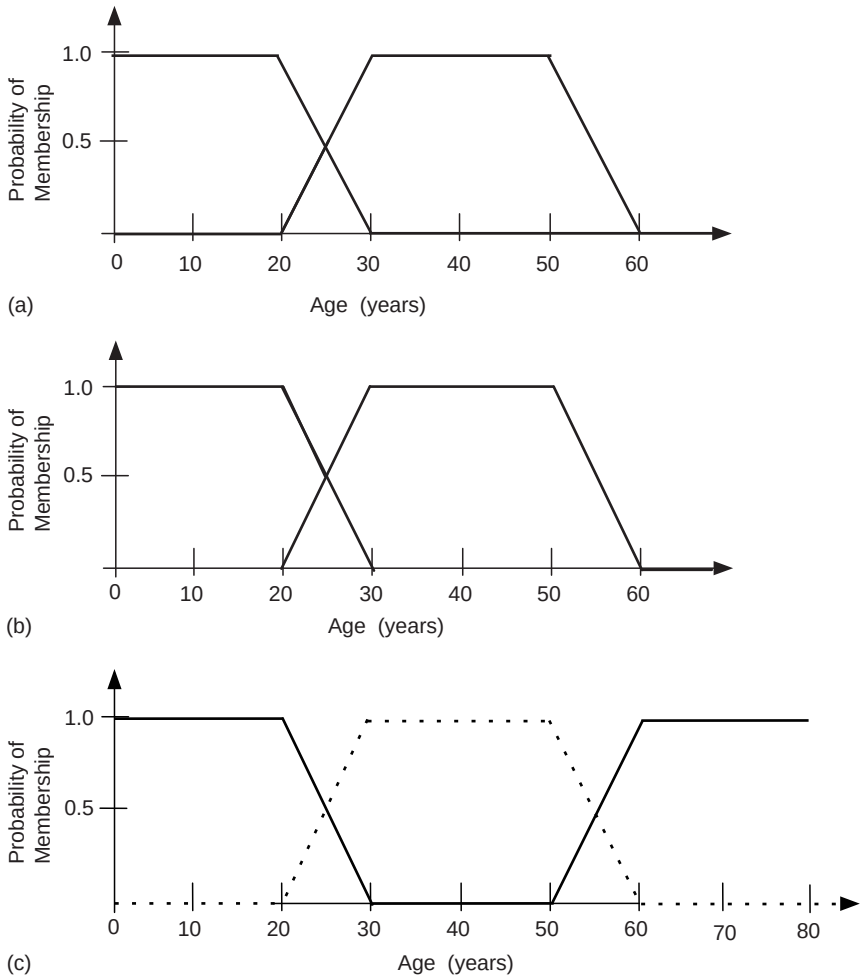


FIGURE 11.38 (a) Intersection of fuzzy sets: YOUNG AND MIDDLE-AGED. (b) Union of fuzzy sets: MIDDLE-AGED OR OLD. (c) Negation of fuzzy sets: NOT OLD.

a process can be described or modeled mathematically, then fuzzy logic will not generally perform better than traditional methods.

Biomedical engineering applications, which involve the analysis and evaluation of biosignals, often have attributes that confound traditional computing methods but are well suited to fuzzy logic. Biological phenomena are often not precisely understood and can be extremely complex. Biological systems also vary significantly from one individual to the next individual. In addition, many key quantities in biological systems cannot be measured precisely due to limitations in existing sensors and other biomedical measuring devices. Sensors may have the capability to measure biological quantities intermittently or in combination with other parameters but not independently. Blood glucose sensors,

for example, are sensitive not only to blood glucose but also to urea and other elements in the blood. Fuzzy logic can be used to help compensate for the limitations of sensors.

Fuzzy logic is being used in a variety of biomedical engineering applications. Closed loop drug delivery systems, which are used to automatically administer drugs to patients, have been developed by using fuzzy logic. In particular, fuzzy logic may prove valuable in the development of drug delivery systems for anesthetic administration, since it is difficult to precisely measure the amount of anesthetic that should be delivered to an individual patient by using conventional computing methods. Fuzzy logic is also being used to develop improved neuroprosthetics for paraplegics. Neuroprosthetics for locomotion use sensors controlled by fuzzy logic systems to electrically stimulate necessary leg muscles and will ideally enable the paraplegic patient to walk.

EXAMPLE PROBLEM 11.28

A fuzzy system is used to categorize people by heart rates. The system is used to help determine which patients have normal resting heart rates, bradycardia, or tachycardia. Bradycardia is a cardiac arrhythmia in which the resting heart rate is less than 60 beats per minute, while tachycardia is defined as a cardiac arrhythmia in which the resting heart rate is greater than 100 beats per minute. A normal heart rate is considered to be in the range of 70–80 beats per minute. What are three linguistic variables that might be used to describe the resting heart rates of the individuals?

Solution

A variety of linguistic variables may be used. The names are important only in that they offer a good description of the categories and problem. Slow, normal, and fast might be used. Another possibility is simply bradycardia, normal, and tachycardia.

11.9.2 Artificial Neural Networks

Artificial neural networks (ANN) are the theoretical counterpart of real biological neural networks. The human brain is a highly sophisticated biological neural network, consisting of billions of brain cells (i.e., neurons) that are highly interconnected among one another. Such a highly interconnected architecture of neurons allows for immense computational power, typically far beyond our most sophisticated computers. The brains of humans, mammals, and even simple invertebrate organism (e.g., a fly) can easily learn from experience, recognize relevant sensory signals (e.g., sounds and images), and react to changes in the organism's environment. Artificial neuronal networks are designed to mimic and attempt to replicate the function of real brains.

ANNs are simpler than biological neural networks. A sophisticated ANN contains only a few thousand neurons with several hundred connections. Although simpler than biological neural networks, the aim of ANNs is to build computer systems that have learning, generalized processing, and adaptive capabilities resembling those seen in real brains. Artificial neural networks can learn to recognize certain inputs and to produce a particular output for a given input. Therefore, artificial neural networks are commonly used for pattern detection and classification of biosignals.

ANNs consist of multiple interconnected neurons. Different types of neurons can be represented in an ANN. Neurons are arranged in a layer, and the different layers of neurons are connected to other neurons and layers. The manner in which the neurons are interconnected determines the architecture of the ANN. There are many different ANN architectures, some of which are best suited for specific applications. Figure 11.39 shows a schematic of a simple ANN with three layers of neurons and a total of six neurons. The first layer is called the *input layer* and has two neurons, which accept the input to the network. The middle layer contains three neurons and is where much of the processing occurs. The *output layer* has one neuron that provides the result of the ANN.

Mathematical equations are used to describe the connections between the neurons. The diagram in Figure 11.40 represents a single neuron and a mathematical method for determining the output of the neuron. The equation for calculating the total input to the neuron is

$$x = (\text{Input}_1 \bullet \text{Weight}_1) + (\text{Input}_2 \bullet \text{Weight}_2) + \text{Bias Weight} \tag{11.59}$$

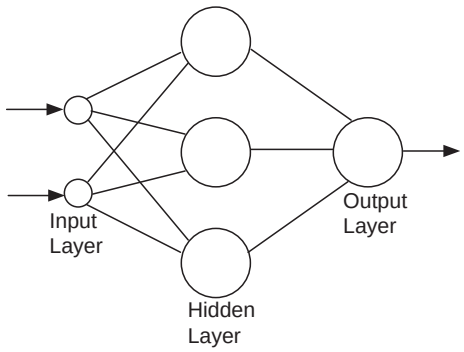


FIGURE 11.39 A simple artificial neural network (ANN) with six neurons and three layers.

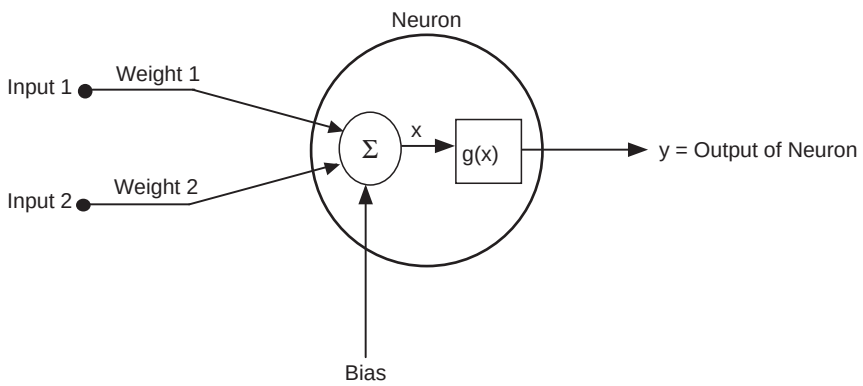


FIGURE 11.40 A single neuron showing mathematical input and output relationships.

The output for the neuron is determined by using a mathematical function, $g(x)$. Threshold functions and nonlinear sigmoid functions are commonly used. The output y of a neuron using the sigmoid function is calculated from the following simple equation:

$$y = 1/(1 + e^{-x}) \tag{11.60}$$

In biosignal processing applications, the inputs to the first layer or input layer of the ANN can be raw data, a preprocessed signal, or extracted features from a biosignal. Raw data are generally a sample from a digitized signal. Preprocessed signals are biosignals that have been transformed, filtered, or processed using some other method before being input to the neural network. Features can also be extracted from biosignals and used as inputs for the neural network. Extracted features might include thresholds; a particular, reoccurring waveshape; or the period between waveforms.

The ANN must learn to recognize the features or patterns in an input signal, but this is not the case initially. In order for the ANN to learn, a training process must occur in which the user of the ANN presents the neural network with many different examples of important input. Each example is given to the ANN many times. Over time, after the ANN has been presented with all of the input examples several times, the ANN learns to produce particular outputs for specific inputs.

There are a variety of types of learning paradigms for ANNs. Learning can be broadly divided into two categories: unsupervised learning and supervised learning. In unsupervised learning, the outputs for the given input examples are not known. The ANN must perform a sort of self-organization. During unsupervised learning, the ANN learns to recognize common features in the input examples and produces a specific output for each different type of input. Types of ANNs with unsupervised learning that have been used in biosignal processing include the Hopfield network and self-organizing feature maps networks.

In supervised learning, the desired output is known for the input examples. The output that the ANN produces for a particular input or inputs is compared against the desired output or output function. The desired output is known as the target. The difference between the target and the output of the ANN is calculated mathematically for each given input example. A common training method for supervised learning is backpropagation. The multilayered perceptron trained with backpropagation is a type of a network with supervised learning that has been used for biosignal processing.

Backpropagation is an algorithm that attempts to minimize the error of the ANN. The error of the ANN can be regarded as simply the difference between the output of the ANN for an input example and the target for that same input example. Backpropagation uses a gradient-descent method to minimize the network error. In other words, the network error is gradually decreased down an error slope that is in some respects similar to how a ball rolls down a hill. The name *backpropagation* refers to the way by which the ANN is changed to minimize the error. Each neuron in the network is "credited" with a portion of the network error. The relative error for each neuron is then determined, and the connection strengths between the neurons are changed to minimize the errors. The weights, such as those that were shown in Figure 11.40, represent the connection strengths between neurons. The calculations of the neuron errors and weight changes propagate backward through the ANN from the output neurons to the input neurons. Backpropagation is the method of finding the optimum weight values that produce the smallest network error.

ANNs are well suited for a variety of biosignal processing applications and may be used as a tool for nonlinear statistical analysis. They are often used for pattern recognition and classification. In addition, ANNs have been shown to perform faster and more accurately than conventional methods for signals that are highly complex or contain high levels of noise. ANNs also have the ability to solve problems that have no algorithmic solution—in other words, problems for which a conventional computer program cannot be written. Since ANNs learn, algorithms are not required to solve problems.

As advances are made in artificial intelligence techniques, ANNs are being used more extensively in biosignal processing and biomedical instrumentation. The viability of ANNs for applications ranging from the analysis of ECG and EEG signals to the interpretation of medical images and the diagnosis of a variety of diseases has been investigated. In neurology, research has been conducted by using ANNs to characterize brain defects that occur in disorders such as epilepsy, Parkinson's disease, and Alzheimer's disease. ANNs have also been used to characterize and classify ECG signals of cardiac arrhythmias. One study used an ANN in the emergency room to diagnose heart attacks. The results of the study showed that, overall, the ANN was better at diagnosing heart attacks than the emergency room physicians were. ANNs have the advantage of not being affected by fatigue, distractions, or emotional stress. As artificial intelligence technologies advance, ANNs may provide a superior tool for many biosignal processing tasks.

EXAMPLE PROBLEM 11.29

A neuron in a neural network has three inputs and uses a sigmoid function to calculate the output of the neuron. The three values of the inputs are 0.1, 0.9, and 0.1. The weights associated with these three inputs are 0.39, 0.72, and 0.26, and the bias weight is 0.48 after training. What is the output of the neuron?

Solution

Using Eq. (11.32) to calculate the relative sum of the inputs gives

$$\begin{aligned} x &= (\text{Input}_1 \bullet \text{Weight}_1) + (\text{Input}_2 \bullet \text{Weight}_2) + (\text{Input}_3 \bullet \text{Weight}_3) + \text{Bias Weight} \\ &= (0.1)\,0.39 + (0.9)\,0.72 + (0.1)\,0.24 + 0.48 \\ &= 1.19 \end{aligned}$$

The output of the neuron is calculated using Eq. (11.60):

$$y = 1/(1 + e^{-x}) = 1/(1 + e^{-1.19}) = 0.77$$

11.10 EXERCISES

1. What types of biosignals would the nerves in your legs produce during a sprint across the street?
2. What types of biosignals can be recorded with an EEG? Describe in terms of both origins and characteristics of the signal.

Continued

3. Describe the biosignal that the electrical activity of a normal heart would generate during a bicycle race.
4. A 16-bit A/D converter is used to convert an analog biosignal with a minimum voltage of –30 mV and a maximum voltage of 90 mV. What is the sensitivity?
5. An EMG recording of skeletal muscle activity has been sampled at 200–250 Hz and correctly digitized. What is the highest frequency of interest in the original EMG signal?
6. Two signals, $x_1(t)$ and $x_2(t)$, have the magnitude spectrum shown in Figure 11.41. Find the Nyquist rate for:
   (a) $x_1(t)$
   (b) $x_2(t)$
   (c) $x(t) = x_1(t) * x_2(t)$
   (Hint: Apply the convolution theorem.)

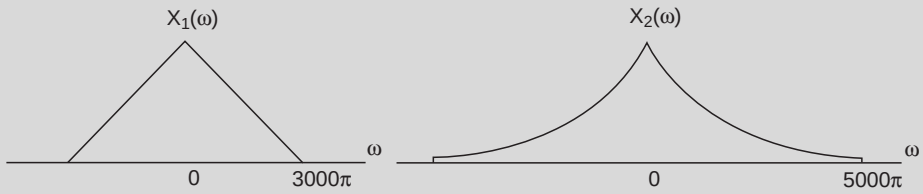


FIGURE 11.41

7. Consider the signal

$$x(t) = 3 + \sin(2\pi\, 100t) + \cos(2\pi\, 250t + \pi/3)$$

   Find the Nyquist frequency.
8. A sinusoid with the frequency of 125 kHz is sampled at 70,000 samples per second. What is the apparent frequency of the sampled signal?
9. An electroencephalographic (EEG) signal has a maximum frequency of 300 Hz. The signal is sampled and quantized into a binary sequence by an A/D converter.
   (a) Determine the sampling rate if the signal is sampled at a rate 50 percent higher than the Nyquist rate.
   (b) The samples are quantized into 2,048 levels. How many binary bits are required for each sample?
10. Find the exponential Fourier series for the signal shown in Figure 11.42a.
11. Find the exponential Fourier series for the signal shown in Figure 11.42b.
12. $f(t)$ is a periodic signal shown in Figure 11.43. Find its trigonometric Fourier series.
13. Consider the following trigonometric Fourier series:

$$f(t) = 3 + 3\cos(t) + 2\cos(2t) + 4\sin(2t) - 4\left(\frac{e^{j4t} + e^{-j4t}}{2}\right).$$

    Write $f(t)$ in its compact trigonometric Fourier series form.
14. Explain why the exponential Fourier series requires negative frequencies.

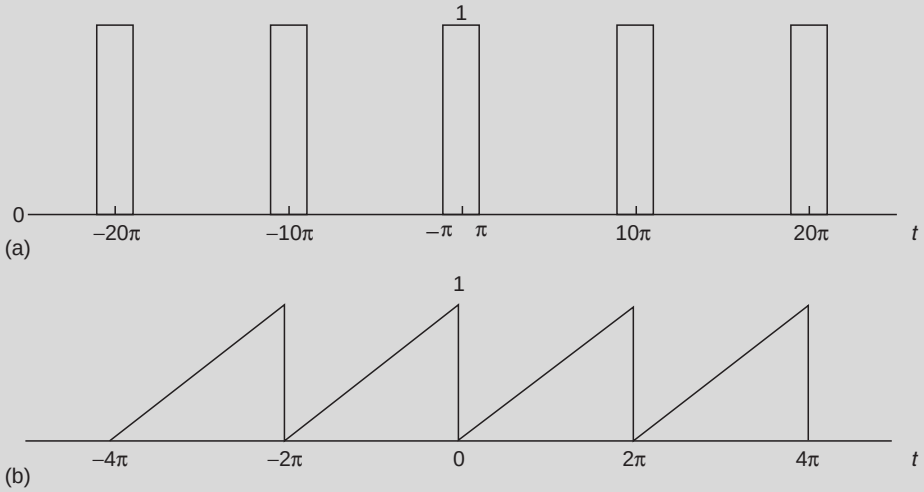


FIGURE 11.42

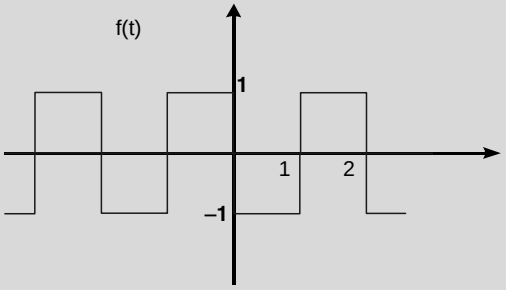


FIGURE 11.43

15. Find the Fourier transform of
(a) $u(t)$
(b) $e^{-at}u(t)$
(c) $\cos(at)u(t)$

16. Find the Fourier transform of $f_1(t) = e^{-3t}u(t)$.

17. Find the Fourier transform of $f(t) = e^{-3|t|}$ and sketch its time and frequency domain representations. (Hint: Find a few points on the curve by substituting values for the variable.)

18. Prove the shift property of the Fourier transform.

19. Given $x(t) = e^{-at}u(t)$ and $h(t) = e^{-bt}u(t)$, where a and b are constants greater than zero, explain why it would be easier to evaluate the convolution $x(t) * h(t)$ in the frequency domain.

20. A brief current pulse of duration 50 ms and amplitude 1 mA is presented to a cell membrane with time constant 10 ms. Find the cell membrane voltage output.

Continued

21. The ion exchange process of a cell is estimated to have the following impulse response: $h(t) = e^{-4t}u(t)$.

(a) Explain what type of general information would be available to the researcher if this estimation of $h(t)$ were accurate.

(b) If sodium ions are injected into the system for two seconds in the form of a brief pulse approximated by the equation $x(t) = 3u(t) - 3u(t-2)$, how would the cell respond to (e.g., pump out ions) such input? Find the answer using time domain procedures.

(Hint: Convolve the input and the impulse response.)

22. An ECG recording of the electrical activity of the heart during ventricular fibrillation is digitized, and the signal begins with the data sequence [–90.0, 10.0, –12.0, –63.0, 7.0, –22.0]. The units of the data sequence are given in mV. What is the z-transform of this data sequence of the biosignal?

23. For the systems described by the following equations, determine which of the systems is linear and which is not.

(a) $\frac{dy}{dt} + 2y(t) = f^2(t)$

(b) $\frac{dy}{dt} + 3ty(t) = t^2 f(t)$

(c) $\frac{dy}{dt} + 2y(t) = f(t)\frac{df}{dt}$

(d) $y(t) = \int_{-\infty}^{t} f(\tau)d\tau$

24. Examine the characteristics of the digital filter

$$y(k) = \frac{1}{4}x(k) + \frac{1}{4}x(k-1) + \frac{1}{2}y(k-1).$$

Find the impulse response, $H(z)$ and $H'(\Omega)$. Use MATLAB to calculate and plot $|H'(\Omega)|$ for $0 < \Omega < \pi$. Observe the difference between this filter and the filter in Example Problem 10.12. Why is this a better low-pass filter? What is the output if the input sequence is $x(k) = 100\sin\left(\frac{\pi}{2}k + \frac{\pi}{8}\right)$? What is the output if the input sequence is $x(k) = 100u(k)$?

25. Find the z-transform of

(a) $x(k) = u(k)$

(b) $x(k) = a^k u(k)$

(c) $x(k) = \cos(b \cdot k)u(k)$

26. Find the z-transform of the following:

(a) $x[k] = \left(\frac{1}{2}\right)^k u(k)$

(b) $x[k] = (\cos \Omega k)u[k]$

27. Find the first four outputs of the discrete system

$$y[k] - 3y[k-1] + 2y[k-2] = f[k-1]$$
$$\text{if } y[-1] = 2,\ \ y[-2] = 3, \text{ and } f[k] = 3^k u[k].$$

28. Find the first four outputs of the discrete system

$$y[k] - 2y[k-1] + 2y[k-2] = f[k-2]$$
$$\text{if } y[-1] = 1,\ \ y[-2] = 0, \text{ and } f[k] = u[k].$$

29. In MATLAB, design a routine to show that averaging random noise across many trials approaches zero as the number of trials increases.

30. Accurate measurements of blood glucose levels are needed for the proper treatment of diabetes. Glucose is a primary carbohydrate that circulates throughout the body and serves as an energy source for cells. In normal individuals the hormone insulin regulates the levels of glucose in the blood by promoting glucose transport out of the blood to skeletal muscle and fat tissues. Diabetics suffer from improper management of glucose levels, and the levels of glucose in the blood can become too high. Describe how fuzzy logic might be used in the control of a system for measuring blood glucose levels. What advantages would the fuzzy logic system have over a more conventional system?

31. Describe three different biosignal processing applications for which artificial neural networks might be used. Give at least two advantages of artificial neural networks over traditional biosignal processing methods for the applications you listed.

32. The fuzzy sets in Example Problem 11.28 have been calibrated so a person with a resting heart rate of 95 beats per minute has a 75 percent degree of membership in the normal category and a 25 percent degree of membership in the tachycardia category. A resting heart rate of 65 beats per minutes indicates a 95 percent degree of membership in the normal category. Draw a graph of the fuzzy sets.

Suggested Readings

M. Akay, Biomedical Signal Processing, Academic Press, Inc., San Diego, CA, 1994.

M. Akay (Ed.), Time Frequency and Wavelets in Biomedical Signal Processing, IEEE Press, New York, NY, 1998.

P. Bauer, S. Nouak, R. Winkler, A Brief Course in Fuzzy Logic and Fuzzy Control, Fuzzy Logic Laboratorium Linz-Hagenberg, Linz, Austria, 1996.http://www.flll.uni-linz.ac.at/fuzzy.

C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press Inc., New York, NY, 1995.

E.N. Bruce, Biomedical Signal Processing and Signal Modeling, Wiley-Interscience, New York, NY, 2000.

E.J. Ciaccio, S.M. Dunn, M. Akay, Biosignal Pattern Recognition and Interpretation Systems: Part 1 of 4: Fundamental Concepts, IEEE Eng. Med. Biol. Mag. 12 (1993) 810–897.

E.J. Ciaccio, S.M. Dunn, M. Akay, Biosignal Pattern Recognition and Interpretation Systems: Part 2 of 4: Methods for Feature Extraction and Selection, IEEE Eng. Med. Biol. Mag. 12 (1993) 106–113.

E.J. Ciaccio, S.M. Dunn, M. Akay, Biosignal Pattern Recognition and Interpretation Systems: Part 3 of 4: Methods of Classification, IEEE Eng. Med. Biol. Mag. 12 (1994) 269–279.

E.J. Ciaccio, S.M. Dunn, M. Akay, Biosignal Pattern Recognition and Interpretation Systems: Part 4 of 4: Review of Applications, IEEE Eng. Med. Biol. Mag. 13 (1994) 269–273.

A. Cohen, Biomedical Signal Processing: Volume I Time and Frequency Domain Analysis, CRC Press, Boca Raton, FL, 1986.

A. Cohen, Biomedical Signal Processing: Volume II Compression and Automatic Recognition, CRC Press, Boca Raton, FL, 1986.

J. Dempster, Computer Analysis of Electrophysiological Signals, Academic Press Inc., San Diego, CA, 1993.

S.R. Devasahayam, Signals and Systems in Biomedical Engineering: Signal Processing and Physiological Systems Modeling, Kluwer Academic, New York, NY, 2000.

S. Haykin, Neural Networks—A Comprehensive Foundation, Macmillan College Publishing Co., Inc., New York, NY, 1994.
R.B. Northrop, Signals and Systems Analysis in Biomedical Engineering, CRC Press, Boca Raton, FL, 2003.
B. Onaral (Ed.), VI. Biomedical signal analysis, in: J.D. Bronzino (Ed.), The Biomedical Engineering Handbook, CRC Press, Boca Raton, FL, 1995.
A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.
A.V. Oppenheim, A.S. Willsky, I.T. Young, Signals and Systems, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1983.
R.A. Roberts, C.T. Mullis, Digital Signal Processing, Addison-Wesley Publishing Co., Reading, MA, 1987.
M. Smith, Neural Networks for Statistical Modeling, International Thomson Computer Press, Boston, MA, 1996.
S.D. Stearns, R.A. David, Signal Processing Algorithms in Fortran and C, P T R Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.
W.J. Thompkins, Biomedical Digital Signal Processing, P T R Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.
C.S. Williams, Designing Digital Filters, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.
L.A. Zadeh, Fuzzy Sets and Applications, John Wiley & Sons, New York, NY, 1987.
R.E. Ziemer, W.H. Tranter, D.R. Fannin, Signals and Systems: Continuous and Discrete, third ed., Macmillan Publishing Co., New York, NY, 1993.

CHAPTER

12

Bioelectric Phenomena

John D. Enderle, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Describe the history of bioelectric phenomenon.
- Qualitatively explain how signaling occurs among neurons.
- Calculate the membrane potential due to one or more ions.
- Compute the change in membrane potential due to a current pulse through a cell membrane.
- Describe the change in membrane potential with distance after stimulation.
- Explain the voltage clamp experiment and an action potential.
- Simulate an action potential using the Hodgkin-Huxley model.
- Describe the process used for communication among neurons.

12.1 INTRODUCTION

Chapter 3 briefly described the nervous system and the concept of a neuron. Here, the description of a neuron is extended by examining its properties at rest and during excitation. The concepts introduced here are basic and allow further investigation of more sophisticated models of the neuron or groups of neurons by using GENESIS (a general neural simulation program; see suggested reading by J.M. Bower and D. Beeman) or extensions of the Hodgkin-Huxley model by using more accurate ion channel descriptions and neural networks. The models introduced here are an important first step in understanding the nervous system and how it functions.

Models of the neuron presented in this chapter have a rich history of development. This history continues today as new discoveries unfold that supplant existing theories and models. Much of the physiological interest in models of a neuron involves the neuron's use in transferring and storing information, while much engineering interest involves the neuron's use as a template in computer architecture and neural networks. New developments in brain-machine interfacing make understanding the neuron even more important today (see [4] and [1] for additional information). To fully appreciate the operation of a neuron, it is important to understand the properties of a membrane at rest by using standard biophysics, biochemistry, and electric circuit tools. In this way, a more qualitative awareness of signaling via the generation of the action potential can be better understood.

The Hodgkin and Huxley theory that was published in 1952 described a series of experiments that allowed the development of a model of the action potential. This work was awarded a Nobel Prize in 1963 (shared with John Eccles) and is covered in Section 12.6. It is reasonable to question the usefulness of covering the Hodgkin-Huxley model in a textbook today, given all of the advances since 1952. One simple answer is that this model is one of the few timeless classics and should be covered. Another is that all current, and perhaps future, models have their roots in this model.

Section 12.2 describes a brief history of bioelectricity and can be easily omitted on first reading of the chapter. Section 12.3 describes the structure and provides a qualitative description of a neuron. Biophysics and biochemical tools that are useful in understanding the properties of a neuron at rest are presented in Section 12.4. An equivalent circuit model of a cell membrane at rest consisting of resistors, capacitors, and voltage sources is described in Section 12.5. Section 12.6 describes the Hodgkin-Huxley model of a neuron and includes a brief description of their experiments and the mathematical model describing an action potential. Finally, Section 12.7 provides a model of the whole neuron.

12.2 HISTORY

12.2.1 The Evolution of a Discipline: The Galvani-Volta Controversy

In 1791, an article appeared in the Proceedings of the Bologna Academy, reporting experimental results that, it was claimed, proved the existence of animal electricity. This now famous publication was the work of Luigi Galvani. At the time of its publication, this article caused a great deal of excitement in the scientific community and sparked

a controversy that ultimately resulted in the creation of two separate and distinct disciplines: electrophysiology and electrical engineering. The controversy arose from the different interpretations of the data presented in this now famous article. Galvani was convinced that the muscular contractions he observed in frog legs were due to some form of electrical energy emanating from the animal. On the other hand, Allesandro Volta, a professor of physics at the University of Padua, was convinced that the "electricity" described in Galvani's experiments originated not from the animal but from the presence of the dissimilar metals used in Galvani's experiments. Both of these interpretations were important. The purpose of this section, therefore, is to discuss them in some detail, highlighting the body of scientific knowledge available at the time these experiments were performed, the rationale behind the interpretations that were formed, and their ultimate effect.

12.2.2 Electricity in the Eighteenth Century

Before 1800, a considerable inventory of facts relating to electricity in general and bioelectricity in particular had accumulated. The Egyptians and Greeks had known that certain fish could deliver substantial shocks to an organism in their aqueous environment. Static electricity had been discovered by the Greeks, who produced it by rubbing resin (amber or, in Greek, *elektron*) with cat's fur or by rubbing glass with silk. For example, Thales of Miletus reported in 600 BC that a piece of amber when vigorously rubbed with a cloth responded with an "attractive power." Light particles such as chaff, bits of papyrus, and thread jumped to the amber from a distance and were held to it. The production of static electricity at that time became associated with an aura.

More than two thousand years elapsed before the English physician William Gilbert picked up where Thales left off. Gilbert showed that not only amber but also glass, agate, diamond, sapphire, and many other materials when rubbed exhibited the same attractive power described by the Greeks. However, Gilbert did not report that particles could also be repelled. It was not until a century later that electrostatic repulsion was noted by Charles DuFay (1698–1739) in France.

The next step in the progress of electrification was an improvement of the friction process. Rotating rubbing machines were developed to give continuous and large-scale production of electrostatic charges. The first of these frictional electric machines was developed by Otto von Guericke (1602–1685) in Germany. In the eighteenth century, electrification became a popular science, and experimenters discovered many new attributes of electrical behavior. In England, Stephen Gray (1666–1736) proved that electrification could flow hundreds of feet through ordinary twine when suspended by silk threads. Thus, he theorized that electrification was a "fluid." Substituting metal wires for the support threads, he found that the charges would quickly dissipate. Thus, the understanding that different materials can either conduct or insulate began to take shape. The "electrics," like silk, glass, and resin, held charge. The "nonelectrics," like metals and water, conducted charges. Gray also found that electrification could be transferred by proximity of one charged body to another without direct contact. This was evidence of electrification by induction, a principle that was used later in machines that produced electrostatic charges.

In France, Charles F. DuFay, a member of the French Academy of Science, was intrigued by Gray's experiments. DuFay showed by extensive tests that practically all materials, with the exception of metals and those too soft or fluid to be rubbed, could be electrified. Later, however, he discovered that if metals were insulated, they could hold the largest electric charge of all. DuFay found that rubbed glass would repel a piece of gold leaf, whereas rubbed amber, gum, or wax attracted it. He concluded that there were two different kinds of electric "fluids," which he labeled "vitreous" and "resinous." He found that while unlike charges attracted each other, like charges repelled. This indicated that there were two kinds of electricity.

In the American colonies, Benjamin Franklin (1706–1790) became interested in electricity and performed experiments that led to his hypothesis regarding the "one-fluid theory." Franklin stated that there was only one type of electricity and that the electrical effects produced by friction reflected the separation of electric fluid so one body contained an excess and the other a deficit. He argued that "electrical fire" is a common element in all bodies and is normally in a balanced or neutral state. Excess or deficiency of charge, such as that produced by the friction between materials, created an imbalance. Electrification by friction was, thus, a process of separation rather than a creation of charge. By balancing a charge gain with an equal charge loss, Franklin had implied a law—namely, that the quantity of the electric charge is conserved. Franklin guessed that when glass was rubbed, the excess charge appeared on the glass, and he called that "positive" electricity. He thus established the direction of conventional current from positive to negative. It is now known that the electrons producing a current move in the opposite direction.

Out of this experimental activity came an underlying philosophy or law. Up to the end of the eighteenth century, the knowledge of electrostatics was mainly qualitative. There were means for detecting but not for measuring, and the relationships between the charges had not been formulated. The next step was to quantify the phenomena of electrostatic charge forces.

For this determination, the scientific scene shifted back to France and the engineer-turned-physicist Charles A. Coulomb (1726–1806). Coulomb demonstrated that a force is exerted when two charged particles are placed in the vicinity of each other. However, he went a step beyond experimental observation by deriving a general relationship that completely expressed the magnitude of this force. His inverse-square law for the force of attraction or repulsion between charged bodies became one of the major building blocks in understanding the effect of a fundamental property of matter-charge. However, despite this wide array of discoveries, it is important to note that before the time of Galvani and Volta, there was no source that could deliver a continuous flow of electric fluid, a term that we now know implies both charge and current.

In addition to a career as statesman, diplomat, publisher, and signer of both the Declaration of Independence and the Constitution, Franklin was an avid experimenter and inventor. In 1743 at the age of 37, Franklin witnessed with excited interest a demonstration of static electricity in Boston and resolved to pursue the strange effects with investigations of his own. Purchasing and devising various apparati, Franklin became an avid electrical enthusiast. He launched into many years of experiments with electrostatic effects.

Franklin, the scientist, is most popularly known for his kite experiment during a thunderstorm in June 1752 in Philadelphia. Although various European investigators had surmised the identity of electricity and lightning, Franklin was the first to prove by an experimental procedure and demonstration that lightning was a giant electrical spark. Having previously noted the advantages of sharp metal points for drawing "electrical fire," Franklin put them to use as "lightning rods." Mounted vertically on rooftops, they would dissipate the thundercloud charge gradually and harmlessly to the ground. This was the first practical application in electrostatics.

Franklin's work was well received by the Royal Society in London. The origin of such noteworthy output from remote and colonial America made Franklin especially marked. In his many trips to Europe as statesman and experimenter, Franklin was lionized in social circles and eminently regarded by scientists.

12.2.3 Galvani's Experiments

Against such a background of knowledge of the "electric fluid" and the many powerful demonstrations of its ability to activate muscles and nerves, it is readily understandable that biologists began to suspect that the "nervous fluid" or the "animal spirit" postulated by Galen to course in the hollow cavities of the nerves and mediate muscular contraction, and indeed all the nervous functions, was of an electrical nature. Galvani, an obstetrician and anatomist, was by no means the first to hold such a view, but his experimental search for evidence of the identity of the electric and nervous fluids provided the critical breakthrough.

Speculations that the muscular contractions in the body might be explained by some form of animal electricity were common. By the eighteenth century, experimenters were familiar with the muscular spasms of humans and animals that were subjected to the discharge of electrostatic machines. As a result, electric shock was viewed as a muscular stimulant. In searching for an explanation of the resulting muscular contractions, various anatomical experiments were conducted to study the possible relationship of "metallic contact" to the functioning of animal tissue. In 1750, Johann Sulzer (1720–1779), a professor of physiology at Zurich, described a chance discovery that an unpleasant acid taste occurred when the tongue was put between two strips of different metals, such as zinc and copper, whose ends were in contact. With the metallic ends separated, there was no such sensation. Sulzer ascribed the taste phenomenon to a vibratory motion set up in the metals that stimulated the tongue and used other metals with the same results. However, Sulzer's reports went unheeded for a half-century until new developments called attention to his findings.

The next fortuitous and remarkable discovery was made by Luigi Galvani (1737–1798), a descendant of a very large Bologna family, who at age 25 was made Professor of Anatomy at the University of Bologna. Galvani had developed an ardent interest in electricity and its possible relation to the activity of the muscles and nerves. Dissected frog legs were convenient specimens for investigation, and in his laboratory Galvani used them for studies of muscular and nerve activity. In these experiments, he and his associates were studying the responses of the animal tissue to various stimulations. In this setting, Galvani observed that while a freshly prepared frog leg was being probed by a scalpel, the leg jerked

convulsively whenever a nearby frictional electrical machine gave off sparks. Galvani said the following of his experiments:

> I had dissected and prepared a frog, and laid it on a table, on which there was an electrical machine. It so happened by chance that one of my assistants touched the point of his scalpel to the inner crural nerve of the frog; the muscles of the limb were suddenly and violently convulsed. Another of those who were helping to make the experiments in electricity thought that he noticed this happening only at the instant a spark came from the electrical machine. He was struck with the novelty of the action. I was occupied with other things at the time, but when he drew my attention to it, I immediately repeated the experiment. I touched the other end of the crural nerve with the point of my scalpel, while my assistant drew sparks from the electrical machine. At each moment when sparks occurred, the muscle was seized with convulsions.

With an alert and trained mind, Galvani designed an extended series of experiments to resolve the cause of the mystifying muscle behavior. On repeating the experiments, he found that touching the muscle with a metallic object while the specimen lay on a metal plate provided the condition that resulted in the contractions.

Having heard of Franklin's experimental proof that a flash of lightning was of the same nature as the electricity generated by electric machines, Galvani set out to determine whether atmospheric electricity might produce the same results observed with his electrical machine. By attaching the nerves of frog legs to aerial wires and the feet to another electrical reference point known as an electrical ground, he noted the same muscular response during a thunderstorm that he observed with the electrical machine. It was another chance observation during this experiment that lead to further inquiry, discovery, and controversy.

Galvani also noticed that the prepared frogs, which were suspended by brass hooks through the marrow and rested against an iron trellis, showed occasional convulsions regardless of the weather. In adjusting the specimens, he pressed the brass hook against the trellis and saw the familiar muscle jerk occurring each time he completed the metallic contact. To check whether this jerking might still be from some atmospheric effect, he repeated the experiment inside the laboratory. He found that the specimen, laid on an iron plate, convulsed each time the brass hook in the spinal marrow touched the iron plate. Recognizing that some new principle was involved, he varied his experiments to find the true cause. In the process, he found that by substituting glass for the iron plate, the muscle response was not observed, but using a silver plate restored the muscle reaction. He then joined equal lengths of two different metals and bent them into an arc. When the tips of this bimetallic arc touched the frog specimens, the familiar muscular convulsions were obtained. As a result, he concluded not only that metal contact was a contributing factor but also that the intensity of the convulsion varied according to the kinds of metals joined in the arc pair.

Galvani was now faced with trying to explain the phenomena he was observing. He had encountered two electrical effects for which his specimens served as indicator: one from the sparks of the electrical machine and the other from the contact of dissimilar metals. Either the electricity responsible for the action resided in the anatomy of the specimens with the metals serving to release it or the effect was produced by the bimetallic contact, with the specimen serving only as an indicator.

Galvani was primarily an anatomist and seized on the first explanation. He ascribed the results to "animal electricity" that resided in the muscles and nerves of the organism itself.

Using a physiological model, he compared the body to a Leyden jar, in which the various tissues developed opposite electrical charges. These charges flowed from the brain through nerves to the muscles. Release of electrical charge by metallic contact caused the convulsions of the muscles. "The idea grew," he wrote, "that in the animal itself there was indwelling electricity. We were strengthened in such a supposition by the assumption of a very fine nervous fluid that during the phenomena flowed into the muscle from the nerve, similar to the electric current of a Leyden jar." Galvani's hypothesis reflected the prevailing view of his day that ascribed the body activation to a flow of "spirits" residing in the various body parts.

In 1791, Galvani published his paper *De Viribus Electricitatis In Motu Musculari* in the Proceedings of the Academy of Science in Bologna. This paper set forth his experiments and conclusions. Galvani's report created a sensation and implied to many a possible revelation of the mystery of the life force. Men of science and laymen alike, both in Italy and elsewhere in Europe, were fascinated and challenged by these findings. However, no one pursued Galvani's findings more assiduously and used them as a stepping stone to greater discovery than Allesandro Volta.

12.2.4 Volta's Interpretation

Galvani's investigations aroused a virtual furor of interest. Wherever frogs were found, scientists repeated his experiments with routine success. Initially, Galvani's explanation for the muscular contractions was accepted without question—even by the prominent physician Allesandro Volta, who had received a copy of Galvani's paper and verified the phenomenon.

Volta was a respected scientist in his own right. At age 24, Volta published his first scientific paper, *On the Attractive Force of the Electric Fire*, in which he speculated about the similarities between electric force and gravity. Engaged in studies of physics and mathematics and busy with experimentation, Volta's talents were so evident that before the age of 30 he was named the Professor of Physics at the Royal School of Como. Here he made his first important contribution to science with the invention of the electrophorus or "bearer of electricity." This was the first device to provide a replenishable supply of electric charge by induction rather than by friction.

In 1782, Volta was called to the professorship of physics at the University of Padua. There he made his next invention, the condensing electrophorus, a sensitive instrument for detecting electric charge. Earlier methods of charge detection employed the "electroscope," which consisted of an insulated metal rod that had pairs of silk threads, pith balls, or gold foil suspended at one end. These pairs diverged by repulsion when the rod was touched by a charge. The amount of divergence indicated the strength of the charge and thus provided quantitative evidence for Coulomb's Law.

By combining the electroscope with his electrophorus, Volta provided the scientific community with a detector for minute quantities of electricity. Volta continued to innovate and made his condensing electroscope a part of a mechanical balance that made it possible to measure the force of an electric charge against the force of gravity. This instrument was called an electrometer and was of great value in Volta's later investigations of the electricity created by contact of dissimilar metals.

Volta expressed immediate interest on learning of Galvani's 1791 report to the Bologna Academy on the "Forces of Electricity in Their Relation to Muscular Motion." Volta set out quickly to repeat Galvani's experiments and initially confirmed Galvani's conclusions on "animal electricity" as the cause of the muscular reactions. Along with Galvani, he ascribed the activity to an imbalance between electricity of the muscle and that of the nerve, which was restored to equilibrium when a metallic connection was made. On continuing his investigations, however, Volta began to have doubts about the correctness of that view. He found inconsistencies in the balance theory. In his experiments, muscles would convulse only when the nerve was in the electrical circuit made by metallic contact.

In an effort to find the true cause of the observed muscle activity, Volta went back to an experiment previously performed by Sulzer. When Volta placed a piece of tinfoil on the tip and a silver coin at the rear of his tongue and connected the two with a copper wire, he got a sour taste. When he substituted a silver spoon for the coin and omitted the copper wire, he got the same result as when he let the handle of the spoon touch the foil. When using dissimilar metals to make contact between the tongue and the forehead, he got a sensation of light. From these results, Volta came to the conclusion that the sensations he experienced could not originate from the metals as conductors but must come from the ability of the dissimilar metals themselves to generate electricity.

After two years of experimenting, Volta published his conclusions in 1792. While crediting Galvani with a surprising original discovery, he disagreed with him on what produced the effects. By 1794, Volta had made a complete break with Galvani. He became an outspoken opponent of the theory of animal electricity and proposed the theory of "metallic electricity." Galvani, by nature a modest individual, avoided any direct confrontation with Volta on the issue and simply retired to his experiments on animals.

Volta's conclusive demonstration that Galvani had not discovered animal electricity was a blow from which the latter never recovered. Nevertheless, he persisted in his belief in animal electricity and conducted his third experiment, which definitely proved the existence of bioelectricity. In this experiment, he held one foot of the frog nerve-muscle preparation and swung it so the vertebral column and the sciatic nerve touched the muscles of the other leg. When this occurred or when the vertebral column was made to fall on the thigh, the muscles contracted vigorously. According to most historians, it was his nephew Giovanni Aldini (1762–1834) who championed Galvani's cause by describing this important experiment in which he probably collaborated. The experiment conclusively showed that muscular contractions could be evoked without metallic conductors. According to Fulton and Cushing, Aldini wrote:

> Some philosophers, indeed, had conceived the idea of producing contractions in a frog without metals; and ingenious methods, proposed by my uncle Galvani, induced me to pay attention to the subject, in order that I might attain to greater simplicity. He made me sensible of the importance of the experiment and therefore I was long ago inspired with a desire of discovering that interesting process. It will be seen in the Opuscoli of Milan (No. 21) that I showed publicly, to the Institute of Bologna, contractions in a frog without the aid of metals so far back as the year 1794. The experiment, as described in a memoir addressed to M. Amorotti [*sic*] is as follows: I immersed a prepared frog in a strong solution of muriate of soda. I then took it from the solution, and, holding one extremity of it in my hand, I suffered the other to hang freely down. While in this position, I raised up the nerves with a small glass rod, in such a manner that they did not touch the muscles. I then suddenly removed the glass rod, and every time that the spinal marrow

> and nerves touched the muscular parts, contractions were excited. Any idea of a stimulus arising earlier from the action of the salt, or from the impulse produced by the fall of the nerves, may be easily removed. Nothing will be necessary but to apply the same nerves to the muscles of another prepared frog, not in a Galvanic circle; for, in this case, neither the salt, nor the impulse even if more violent, will produce muscular motion.

The claims and counterclaims of Volta and Galvani developed rival camps of supporters and detractors. Scientists swayed from one side to the other in their opinions and loyalties. Although the subject was complex and not well understood, it was on the verge of an era of revelation. The next great contribution to the field was made by Carlo Matteucci, who both confirmed Galvani's third experiment and made a new discovery. Matteucci showed that the action potential precedes the contraction of skeletal muscle. In confirming Galvani's third experiment, which demonstrated the injury potential, Matteucci noted:

> I injure the muscles of any living animal whatever, and into the interior of the wound I insert the nerve of the leg, which I hold, insulated with glass tube. As I move this nervous filament in the interior of the wound, I see immediately strong contractions in the leg. To always obtain them, it is necessary that one point of the nervous filament touches the depths of the wound, and that another point of the same nerve touches the edge of the wound.

By using a galvanometer, Matteucci found that the difference in potential between an injured and uninjured area was diminished during a tetanic contraction. The study of this phenomenon occupied the attention of all succeeding electrophysiologists. More than this, however, Matteucci made another remarkable discovery: that a transient bioelectric event, now designated the action potential, accompanies the contraction of intact skeletal muscle. He demonstrated this by showing that a contracting muscle is able to stimulate a nerve that, in turn, causes contraction of the muscle it innervates. The existence of a bioelectric potential was established through the experiments of Galvani and Matteucci. Soon thereafter, the presence of an action potential was discovered in cardiac muscle and nerves.

Volta, on the other hand, advocated that the source of the electricity was due to the contact of the dissimilar metals only, with the animal tissue acting merely as the indicator. His results differed substantially depending on the pairs of metals used. For example, Volta found that the muscular reaction from dissimilar metals increased in vigor depending on the metals that were used.

In an effort to obtain better quantitative measurements, Volta dispensed with the use of muscles and nerves as indicators. He substituted instead his "condensing electroscope." He was fortunate in the availability of this superior instrument because the contact charge potential of the dissimilar metals was minute, far too small to be detected by the ordinary gold-leaf electroscope. Volta's condensing electroscope used a stationary disk and a removable disk separated by a thin insulating layer of shellac varnish. The thinness of this layer provided a large capacity for accumulation of charge. When the upper disk was raised after being charged, the condenser capacity was released to give a large deflection of the gold leaves.

Volta proceeded systematically to test the dissimilar metal contacts. He made disks of various metals and measured the quantity of the charge on each disk combination by the divergence of his gold foil condensing electroscope. He then determined whether the charge was positive or negative by bringing a rubbed rod of glass or resin near the electroscope. The effect of the rod on the divergence of the gold foil indicated the polarity of the charge.

Volta's experiments led him toward the idea of an electric force or electrical "potential." This, he assumed, resided in contact between the dissimilar metals. As Volta experimented with additional combinations, he found that an electrical potential also existed when there was contact between the metals and some fluids. As a result, Volta added liquids, such as brine and dilute acids, to his conducting system and classified the metal contacts as "electrifiers of the first class" and the liquids as electrifiers of the "second class."

Volta found that there was only momentary movement of electricity in a circuit composed entirely of dissimilar metals. However, when he put two dissimilar metals in contact with a separator soaked with a saline or acidified solution, there was a steady indication of potential. In essence, Volta was assembling the basic elements of an electric battery: two dissimilar metals and a liquid separator. Furthermore, he found that the overall electric effect could be enlarged by multiplying the elements. Thus, by stacking metal disks and the moistened separators vertically, he constructed an "electric pile," the first electric battery. This was the most practical discovery of his career.

12.2.5 The Final Result

Considerable time passed before true explanations became available for what Galvani and Volta had done. Clearly, both demonstrated the existence of a difference in electric potential, but what had produced it eluded them. The potential difference present in the experiments carried out by both investigators is now clearly understood. Although Galvani thought that he had initiated muscular contractions by discharging animal electricity resident in a physiological capacitor consisting of the nerve (inner conductor) and muscle surface (outer conductor), it is now known that the stimulus consists of an action potential that in turn causes muscular contractions.

It is interesting to note that the fundamental unit of the nervous system—the neuron—has an electric potential between the inside and outside of the cell, even at rest. This membrane-resting potential is continually affected by various inputs to the cell. When a certain potential is reached, an action potential is generated along its axon to all of its distant connections. This process underlies the communication mechanisms of the nervous system. Volta's discovery of the electrical battery provided the scientific community with the first steady source of electrical potential, which when connected in an electric circuit consisting of conducting materials or liquids, results in the flow of electrical charge—that is, electrical current. This device launched the field of electrical engineering.

12.3 NEURONS

A reasonable estimate of the human brain is that it contains about 10^{12} neurons partitioned into fewer than 1,000 different types in an organized structure of rather uniform appearance. While not important in this chapter, it is important to note that there are two classes of neuron: the nerve cell and the neuroglial cell. Even though there are 10 to 50 times as many neuroglial cells as nerve cells in the brain, attention is focused here on the nerve cell, since the neuroglial cells are not involved in signaling and primarily provide a support function for the nerve cell. Therefore, the terms *neuron* and *nerve cell* are used interchangeably,

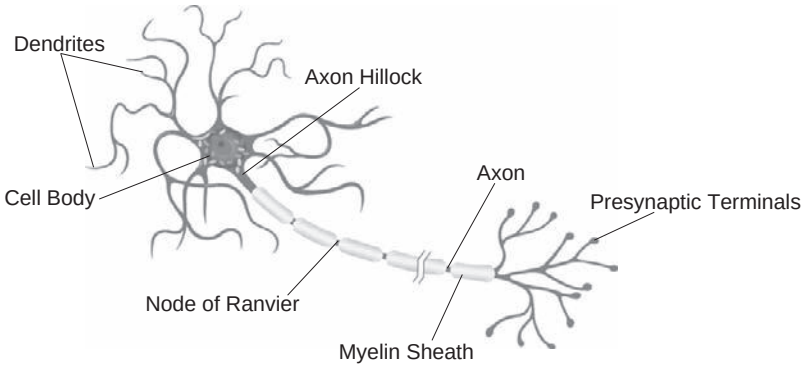


FIGURE 12.1 A typical neuron.

since the primary focus here is to better understand the signaling properties of a neuron. Overall, the complex abilities of the brain are best described by virtue of a neuron's interconnections with other neurons or the periphery and not a function of the individual differences among neurons.

A typical neuron, as shown in Figure 12.1, is defined with four major regions: cell body (also referred to as the soma), dendrites, axon, and presynaptic terminals. The cell body of a neuron contains the nucleus and other organelles needed to nourish the cell and is similar to other cells. Unlike other cells, however, the neuron's cell body is connected to a number of branches called dendrites and a long tube called the axon that connects the cell body to the presynaptic terminals. Some neurons have multiple axons.

Dendrites are the receptive surfaces of the neuron that receive signals from thousands of other neurons passively and without amplification. Located on the dendrite and cell body are receptor sites that receive input from presynaptic terminals from adjacent neurons. Neurons typically have 10^4 to 10^5 synapses. Communication between neurons is through chemical synapses. Chemical synapses, as described in Chapter 3, involve the use of a neurotransmitter that changes the membrane potential of an adjacent neuron. Other cells, such as muscle and cardiac cells, use electrical synapses. An electrical synapse involves the use of a gap junction that directly connects the two cells together through a pore.

Also connected to the neuron cell body is a single axon that ranges in length from 1 meter in the human spinal cord to a few millimeters in the brain. The diameter of the axon also varies from less than 1 to 500 μm. In general, the larger the diameter of the axon, the faster the signal travels. Signals traveling in the axon range from 0.5 m/s to 120 m/s. The purpose of an axon is to serve as a transmission line to move information from one neuron to another at great speeds. Some axons are surrounded by a fatty insulating material called the myelin sheath and have regular gaps, called the nodes of Ranvier, that allow the action potential to jump from one node to the next. The action potential is most easily envisioned as a pulse that travels the length of the axon without decreasing in amplitude.

Most of the remainder of this chapter is devoted to understanding this process. At the end of the axon is a network of up to 10,000 branches with endings called the presynaptic terminals. A diagram of the presynaptic terminal is shown in Figure 3.28. All action potentials that

move through the axon propagate through each branch to the presynaptic terminal. The presynaptic terminals are the transmitting units of the neuron, which, when stimulated, release a neurotransmitter that flows across a gap of approximately 20 nanometers to an adjacent cell, where it interacts with the postsynaptic membrane and changes its potential.

12.3.1 Membrane Potentials

The neuron, like other cells in the body, has a separation of charge across its external membrane. The cell membrane is positively charged on the outside and negatively charged on the inside, as illustrated in Figure 12.2. This separation of charge, due to the selective permeability of the membrane to ions, is responsible for the membrane potential. In the neuron, the potential difference across the cell membrane is approximately 60 mV to 90 mV, depending on the specific cell. By convention, the outside is defined as 0 mV (ground), and the resting potential is $V_m = v_i - v_o = -60$ mV. This charge differential is of particular interest, since most signaling involves changes in this potential across the membrane. Signals such as action potentials are a result of electrical perturbations of the membrane. By definition, if the membrane is more negative than resting potential (i.e., -60 to -70 mV), it is called *hyperpolarization*, and an increase in membrane potential from resting potential (i.e., -60 to -50 mV) is called *depolarization*. As described later, ions travel across the cell membrane through ion selective channels.

Creating a membrane potential of -60 mV does not require the separation of many positive and negative charges across the membrane. The actual number, however, can be found from the relationship $Cdv = dq$, or $C\Delta v = \Delta q$ (Δq = the number of charges times the electron charge of 1.6022×10^{-19}C). Therefore, with $C = 1\ \mu F/cm^2$ and $\Delta v = 60 \times 10^{-3}$, the number of charges equals approximately 1×10^8 per cm^2. These charges are located within a distance of 1 μm from the membrane.

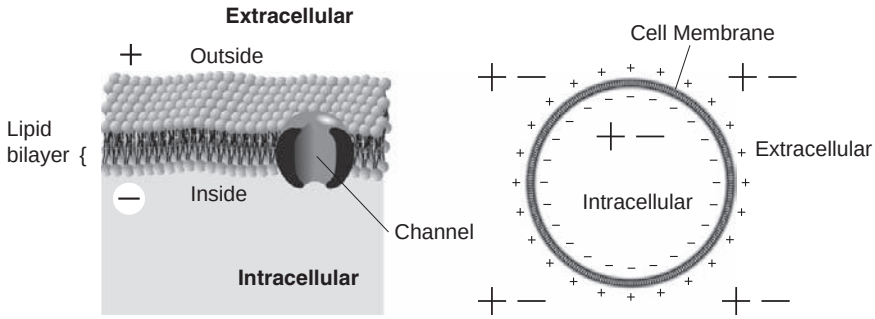


FIGURE 12.2 The separation of charges across a cell membrane. The figure on the left shows a cell membrane with positive ions along the outer surface of the cell membrane and negative ions along the inner surface of the cell membrane. The figure on the right further illustrates separation of charge by showing that only the ions along the inside and outside of the cell membrane are responsible for membrane potential (negative ions along the inside and positive ions along the outside of the cell membrane). Elsewhere the distribution of negative and positive ions are approximately evenly distributed as indicated with the large + − symbols for the illustration on the right. Overall, there is a net excess of negative ions inside the cell and a net excess of positive ions in the immediate vicinity outside the cell. For simplicity, the membrane shown on the right is drawn as the solid circle and ignores the axon and dendrites.

Graded Response and Action Potentials

A neuron can change the membrane potential of another neuron to which it is connected by releasing its neurotransmitter. The neurotransmitter crosses the synaptic cleft, interacts with receptor molecules in the postsynaptic membrane of the dendrite or cell body of the adjacent neuron, and changes the membrane potential of the receptor neuron (Figure 12.3).

The change in membrane potential at the postsynaptic membrane is due to a transformation from neurotransmitter chemical energy to electrical energy. The change in membrane potential depends on how much neurotransmitter is received and can be depolarizing or hyperpolarizing. This type of change in potential is typically called a *graded response*, since it varies with the amount of neurotransmitter received. Another way of envisaging the activity at the synapse is that the neurotransmitter received is integrated or summed, which results in a graded response in the membrane potential. Note that while a signal from a neuron is either inhibitory or excitatory, specific synapses may be excitatory and others inhibitory, providing the nervous system with the ability to perform complex tasks.

The net result of activation of the nerve cell is the action potential. The action potential is a large depolarizing signal of up to 100 mV that travels along the axon and lasts approximately 1–5 ms. Figure 12.4 illustrates a typical action potential. The action potential is an all or none signal that propagates actively along the axon without decreasing in amplitude. When the signal reaches the end of the axon at the presynaptic terminal, the change in potential causes the release of a packet of neurotransmitter. This is a very effective method of signaling over large distances. Additional details about the action potential are described throughout the remainder of this chapter after some tools for better understanding this phenomenon are introduced.

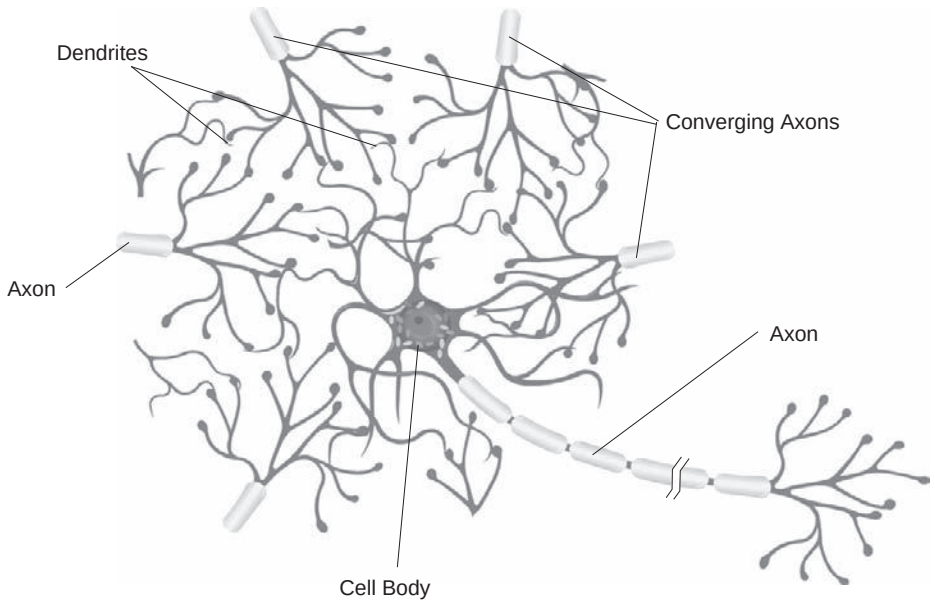


FIGURE 12.3 A typical neuron with presynaptic terminals of adjacent neurons in the vicinity of its dendrites.

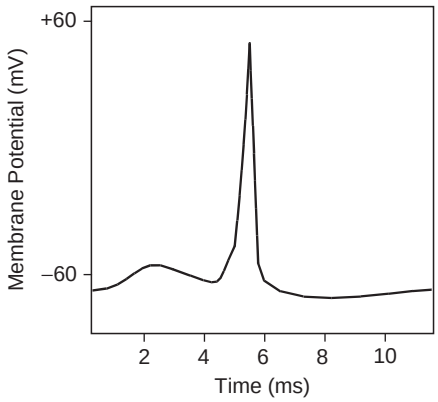


FIGURE 12.4 An action potential.

12.3.2 Resting Potential, Ionic Concentrations, and Channels

A resting membrane potential exists across the cell membrane because of the *differential distribution* of ions in and around the membrane of the nerve cell. The cell maintains these ion concentrations by using a selectively permeable membrane and, as described later, an active ion pump. A selectively permeable cell membrane with ion channels is illustrated in Figure 12.2. The neuron cell membrane is approximately 10 nm thick, and because it consists of a lipid bilayer (i.e., two plates separated by an insulator), it has capacitive properties. The extracellular fluid is composed of primarily Na^+ and Cl^-, and the intracellular fluid (cytosol) is composed of primarily K^+ and A^-. The large organic anions (A^-) are primarily amino acids and proteins that do not cross the membrane. Almost without exception, ions cannot pass through the cell membrane except through a channel.

Channels allow ions to pass through the membrane, are selective, and are either passive or active. Passive channels are always open and are ion specific. Figure 12.5 illustrates a cross section of a cell membrane with passive channels only. As shown, a particular channel allows only one ion type to pass through the membrane and prevents all other ions from crossing the membrane through that channel. Passive channels exist for Cl^-, K^+, and Na^+. In addition, a passive channel exists for Ca^{+2}, which is important in the excitation of the membrane at the synapse, as described in Section 12.8. Active channels, or gates, are either opened or closed in response to an external electrical or chemical stimulation. The active channels are also selective and allow only specific ions to pass through the membrane. Typically, active gates open in response to neurotransmitters and an appropriate change in membrane potential. Figure 12.6 illustrates the concept of an active channel. Here, K^+ passes through an active channel and Cl^- passes through a passive channel. As will be shown, passive channels are responsible for the resting membrane potential, and active channels are responsible for the graded response and action potentials.

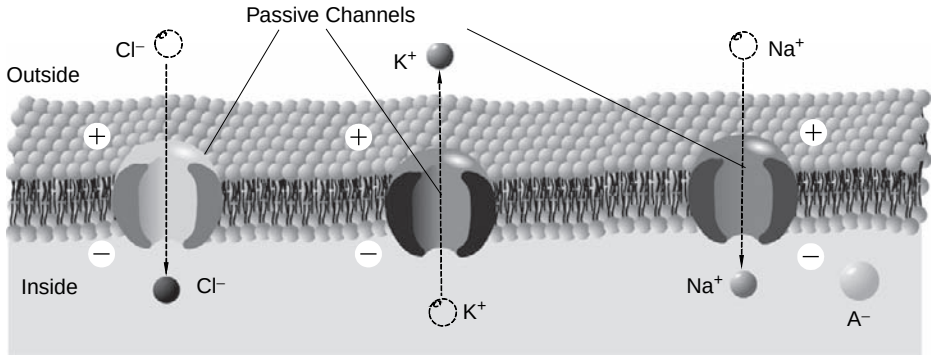


FIGURE 12.5 Idealized cross section of a selectively permeable membrane with channels for ions to cross the membrane. The thickness of the membrane and the size of the channels are not drawn to scale. When the diagram is drawn to scale, the cell membrane thickness is 20 times the size of the ions and 10 times the size of the channels, and the spacing between the channels is approximately 10 times the cell membrane thickness. Note that a potential difference exists between the inside and outside of the membrane, as illustrated with the + and – signs. The membrane is selectively permeable to ions through ion-specific channels—that is, each channel shown here only allows one particular ion to pass through it.

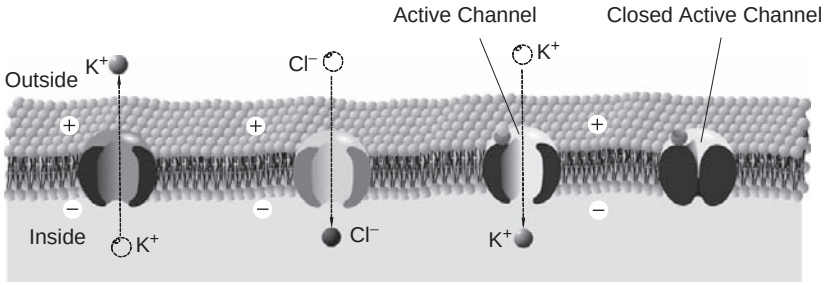


FIGURE 12.6 Passive and active channels provide a means for ions to pass through the membrane. Each channel is ion-specific. As shown, the active channel on the left allows K^+ to pass through the membrane, but the active channel on the right is not open, preventing any ion from passing through the membrane. Also shown is a passive Cl^- channel.

12.4 BASIC BIOPHYSICS TOOLS AND RELATIONSHIPS

12.4.1 Basic Laws

Two basic biophysics tools and a relationship are used to characterize the resting potential across a cell membrane by quantitatively describing the impact of the ionic gradient and electric field.

Fick's Law of Diffusion

The flow of particles due to diffusion is along the concentration gradient, with particles moving from high concentration areas to low ones. Specifically, for a cell membrane, the flow of ions across a membrane is given by

$$J(diffusion) = -D\frac{d[I]}{dx} \tag{12.1}$$

where J is the flow of ions due to diffusion, $[I]$ is the ion concentration, dx is the membrane thickness, and D is the diffusivity constant in m^2/s. The negative sign indicates that the flow of ions is from higher to lower concentration, and $\frac{d[I]}{dx}$ represents the concentration gradient. Fick's Law of diffusion was described in Section 7.3 involving a first-order differential equation that describes change in concentration as a function of time. Here, we are only interested in steady state.

Ohm's Law

Charged particles in a solution experience a force resulting from other charged particles and electric fields present. The flow of ions across a membrane is given by

$$J(drift) = -\mu Z[I]\frac{dv}{dx} \tag{12.2}$$

where J is the flow of ions due to drift in an electric field $\vec{E}$, μ = mobility in m^2/sV, Z = ionic valence, $[I]$ is the ion concentration, v is the voltage across the membrane, and $\frac{dv}{dx}$ is $(-\vec{E})$. Note that Z is positive for positively charged ions (e.g., Z = 1 for Na^+ and Z = 2 for Ca^{+2}) and negative for negatively charged ions (e.g., Z = −1 for Cl^-). Positive ions drift down the electric field and negative ions drift up the electric field.

Figure 12.7 illustrates a cell membrane that is permeable to only K^+ and shows the forces acting on K^+. Assume that the concentration of K^+ is that of a neuron with a higher concentration inside than outside and that the membrane resting potential is negative from inside to outside. Clearly, only K^+ can pass through the membrane, and Na^+, Cl^-, and A^- cannot move through it, since there are no channels for them to pass through. Depending on the actual concentration and membrane potential, K^+ will pass through the membrane until the forces due to drift and diffusion are balanced. The chemical force due to diffusion from inside to outside decreases as K^+ moves through the membrane, and the electric force increases as K^+ accumulates outside the cell until the two forces are balanced.

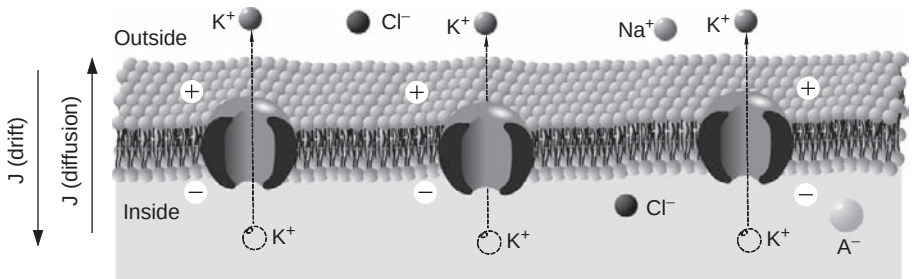


FIGURE 12.7 The direction of the flow of K^+ due to drift and diffusion across a cell membrane that is only permeable to K^+.

Einstein Relationship

The relationship between the drift of particles in an electric field under osmotic pressure described by Einstein in 1905 is given by

$$D = \frac{KT\mu}{q} \tag{12.3}$$

where D is the diffusivity constant, μ is mobility, K is Boltzmann's constant, T is the absolute temperature in degrees Kelvin, and q is the magnitude of the electric charge (i.e., 1.60186×10^{-19} coulombs).

12.4.2 Resting Potential of a Membrane Permeable to One Ion

The flow of ions in response to concentration gradients is limited by the selectively permeable nerve cell membrane and the resultant electric field. As described, ions pass through channels that are selective for that ion only. For clarity, the case of a membrane permeable to one ion only is considered first and then the case of a membrane permeable to more than one ion follows. It is interesting to note that neuroglial cells are permeable to only K^+ and that nerve cells are permeable to K^+, Na^+, and Cl^-. As will be shown, the normal ionic gradient is maintained if the membrane is permeable only to K^+ as in the neuroglial cell.

Consider the cell membrane shown in Figure 12.7 that is permeable only to K^+, and assume that the concentration of K^+ is higher in the intracellular fluid than the extracellular fluid. For this situation, the flow due to diffusion (concentration gradient) tends to push K^+ outside of the cell and is given by

$$J_K\,(diffusion) = -D\frac{d[K^+]}{dx} \tag{12.4}$$

The flow due to drift (electric field) tends to push K^+ inside the cell and is given by

$$J_K\,(drift) = -\mu Z[K^+]\frac{dv}{dx} \tag{12.5}$$

which results in a total flow

$$J_K = J_K\,(diffusion) + J_K\,(drift) = -D\frac{d[K^+]}{dx} - \mu Z[K^+]\frac{dv}{dx} \tag{12.6}$$

Using the Einstein relationship $D = \dfrac{KT\mu}{q}$, the total flow is now given by

$$J_K = -\frac{KT}{q}\mu\frac{d[K^+]}{dx} - \mu Z[K^+]\frac{dv}{dx} \tag{12.7}$$

From Eq. (12.7), the flow of K^+ is found at any time for any given set of initial conditions. In the special case of steady state—that is, at steady state when the flow of K^+ into the cell is exactly balanced by the flow out of the cell or $J_K = 0$—Eq. (12.7) reduces to

$$0 = -\frac{KT}{q}\mu\frac{d[K^+]}{dx} - \mu Z[K^+]\frac{dv}{dx} \tag{12.8}$$

With Z = +1, Eq. (12.8) simplifies to

$$dv = -\frac{KT}{q[K^+]}d[K^+] \tag{12.9}$$

Integrating Eq. (12.9) from outside the cell to inside yields

$$\int_{v_o}^{v_i} dv = -\frac{KT}{q}\int_{[K^+]_o}^{[K^+]_i}\frac{d[K^+]}{[K^+]} \tag{12.10}$$

where v_o and v_i are the voltages outside and inside the membrane, and $[K^+]_o$ and $[K^+]_i$ are the concentrations of potassium outside and inside the membrane. Thus,

$$v_i - v_o = -\frac{KT}{q}\ln\left(\frac{[K^+]_i}{[K^+]_o}\right) = \frac{KT}{q}\ln\left(\frac{[K^+]_o}{[K^+]_i}\right) \tag{12.11}$$

Equation (12.11) is known as the *Nernst equation*, named after German physical chemist Walter Nernst, and $E_K = v_i - v_o$ is known as the *Nernst potential* for K^+. At room temperature, $\frac{KT}{q} = 26\,mV$, and thus the Nernst equation for K^+ becomes

$$E_K = v_i - v_o = 26\ln\frac{[K^+]_o}{[K^+]_i}mV \tag{12.12}$$

While Eq. (12.12) is specifically written for K^+, it can be easily derived for any permeable ion. At room temperature, the Nernst potential for Na^+ is

$$E_{Na} = v_i - v_o = 26\ln\frac{[Na^+]_o}{[Na^+]_i}mV \tag{12.13}$$

and the Nernst potential for Cl^- is

$$E_{Cl} = v_i - v_o = -26\ln\frac{[Cl^-]_o}{[Cl^-]_i} = 26\ln\frac{[Cl^-]_i}{[Cl^-]_o}mV \tag{12.14}$$

The negative sign in Eq. (12.14) is due to Z = −1 for Cl^-.

12.4.3 Donnan Equilibrium

Consider a neuron at steady state that is permeable to more than one ion—for example, K^+, Na^+, and Cl^-. The Nernst potential for each ion can be calculated using Eqs. (12.12) to (12.14). The membrane potential, $V_m = v_i - v_o$, however, is due to the presence of all ions and is influenced by the concentration and permeability of each ion. In this section, the case in which two permeable ions are present is considered. In the next section, the case in which any number of permeable ions are present is considered.

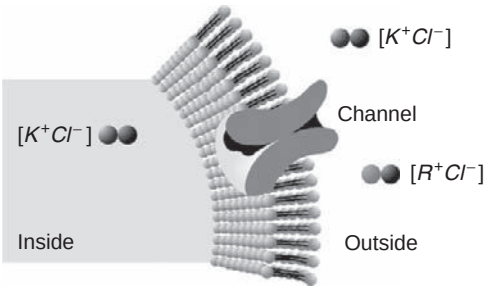


FIGURE 12.8 Membrane is permeable to both K^+ and Cl^-, but not to a large cation R^+.

Suppose a membrane is permeable to both K^+ and Cl^- but not to a large cation, R^+, as shown in Figure 12.8. Under steady-state conditions, the Nernst potentials for both K^+ and Cl^- must be equal—that is $E_K = E_{Cl}$, or

$$E_K = \frac{KT}{q} \ln \frac{[K^+]_o}{[K^+]_i} = E_{Cl} = \frac{KT}{q} \ln \frac{[Cl^-]_i}{[Cl^-]_o} \tag{12.15}$$

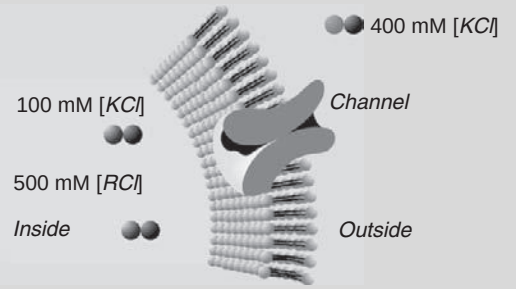
After simplifying,

$$\frac{[K^+]_o}{[K^+]_i} = \frac{[Cl^-]_i}{[Cl^-]_o} \tag{12.16}$$

Equation (12.16) is known as the *Donnan Equilibrium*. An accompanying principle is *space charge neutrality*, which states that the number of cations in a given volume is equal to the number of anions. Thus, at steady-state, ions still diffuse across the membrane, but each K^+ that crosses the membrane must be accompanied by a Cl^- for space charge neutrality to be satisfied. If in Figure 12.8 R^+ were not present, then at steady-state, the concentration of K^+ and Cl^- on both sides of the membrane would be equal. With R^+, the concentrations of [KCl] on both sides of the membrane are different, as shown in Example Problem 12.1, where R^+ is now in the intracellular fluid.

EXAMPLE PROBLEM 12.1

A membrane is permeable to K^+ and Cl^-, but not to a large cation R^+ as shown in the following figure. Find the steady-state concentration for the following initial conditions.



Continued

Solution

By conservation of mass,

$$[K^+]_i + [K^+]_o = 500$$

$$[Cl^-]_i + [Cl^-]_o = 1000$$

and space charge neutrality,

$$[K^+]_i + 500 = [Cl^-]_i$$

$$[K^+]_o = [Cl^-]_o$$

From the Donnan equilibrium,

$$\frac{[K^+]_o}{[K^+]_i} = \frac{[Cl^-]_i}{[Cl^-]_o}$$

Substituting for $[K^+]_o$ and $[Cl^-]_o$ from the conservation of mass equations into the Donnan equilibrium equation gives

$$\frac{500 - [K^+]_i}{[K^+]_i} = \frac{[Cl^-]_i}{1000 - [Cl^-]_i}$$

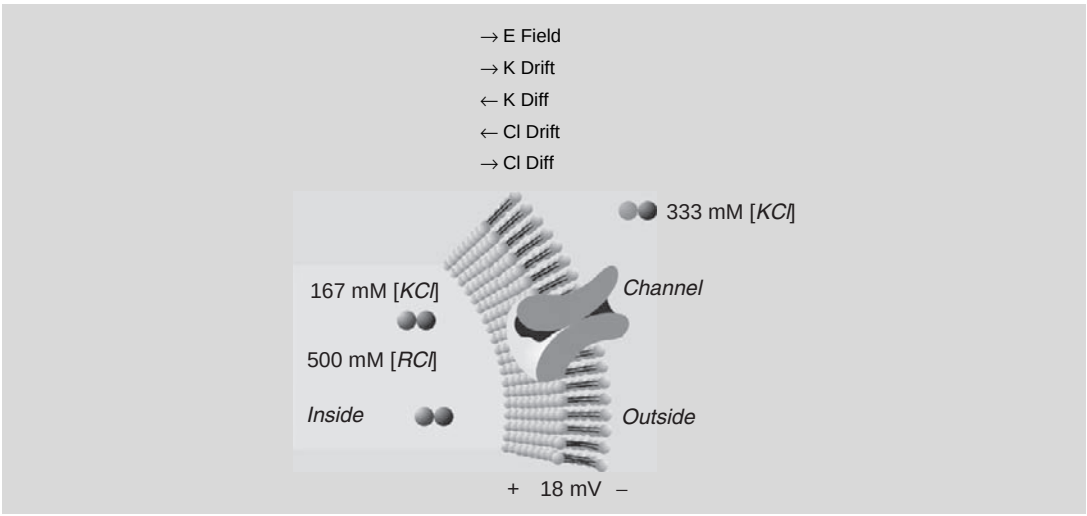
and eliminating $[Cl^-]_i$ by using the space charge neutrality equations gives

$$\frac{500 - [K^+]_i}{[K^+]_i} = \frac{[K^+]_i + 500}{1000 - [K^+]_i - 500} = \frac{[K^+]_i + 500}{500 - [K^+]_i}$$

Solving the preceding equation yields $[K^+]_i = 167$ mM at steady-state. Using the conservation of mass equations and space charge neutrality equation gives $[K^+]_o = 333$ mM, $[Cl^-]_i = 667$ mM, and $[Cl^-]_o = 333$ mM at steady-state. At steady-state and at room temperature, the Nernst potential for either ion is 18 mV, as shown for $[K^+]$:

$$E_K = v_i - v_o = 26 \ln \frac{333}{167} = 18\ mV$$

Summarizing, at steady-state



12.4.4 Goldman Equation

The squid giant axon resting potential is $-60\,mV$, which does not equal the Nernst potential for Na^+ or K^+. As a general rule, when V_m is affected by two or more ions, each ion influences V_m, as determined by its concentration and membrane permeability. The Goldman equation quantitatively describes the relationship between V_m and permeable ions, but only applies when the membrane potential or electric field is constant. This situation is a reasonable approximation for a resting membrane potential.

In this section, the Goldman equation is first derived for K^+ and Cl^-, and then extended to include K^+, Cl^-, and Na^+. The Goldman equation is used by physiologists to calculate the membrane potential for a variety of cells and, in fact, was used by Hodgkin, Huxley, and Katz in studying the squid giant axon.

Consider the cell membrane shown in Figure 12.9. To determine V_m for both K^+ and Cl^-, flow equations for each ion are derived separately under the condition of a constant electric

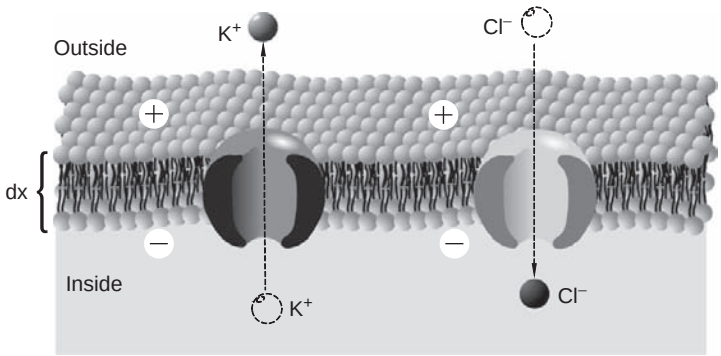


FIGURE 12.9 A cell membrane that is permeable to both K^+ and Cl^-. The width of the membrane is $dx = \delta$.

field and then combined using space charge neutrality to complete the derivation of the Goldman equation.

Potassium Ions

The flow equation for K^+ with mobility μ_K is

$$J_K = -\frac{KT}{q}\mu_K \frac{d[K^+]}{dx} - \mu_K Z_K [K^+]\frac{dv}{dx} \tag{12.17}$$

Under a constant electric field,

$$\frac{dv}{dx} = \frac{\Delta v}{\Delta x} = \frac{V}{\delta} \tag{12.18}$$

Substituting Eq. (12.18) into (12.17) with $Z_K = 1$ gives

$$J_K = -\frac{KT}{q}\mu_K \frac{d[K^+]}{dx} - \mu_K [K^+]\frac{V}{\delta} \tag{12.19}$$

Let the permeability for K^+, P_K, equal

$$P_K = \frac{\mu_K KT}{\delta q} = \frac{D_K}{\delta} \tag{12.20}$$

Substituting Eq. (12.20) into (12.19) gives

$$J_K = \frac{-P_K q}{KT} V[K^+] - P_K \delta \frac{d[K^+]}{dx} \tag{12.21}$$

Rearranging the terms in Eq. (12.21) yields

$$dx = \frac{d[K^+]}{\dfrac{-J_K}{P_K \delta} - \dfrac{qV[K^+]}{KT\delta}} \tag{12.22}$$

Taking the integral of both sides, while assuming that J_K is independent of x, gives

$$\int_0^{\delta} dx = \int_{[K^+]_i}^{[K^+]_o} \frac{d[K^+]}{\dfrac{-J_K}{P_K \delta} - \dfrac{qV[K^+]}{KT\delta}} \tag{12.23}$$

resulting in

$$x\Big|_0^{\delta} = -\frac{KT\delta}{qV} \ln\left(\frac{J_K}{P_K \delta} + \frac{qV[K^+]}{KT\delta}\right)\Bigg|_{[K^+]_i}^{[K^+]_o} \tag{12.24}$$

and

$$\delta = -\frac{KT\delta}{qV} \ln\left(\frac{\dfrac{J_K}{P_K \delta} + \dfrac{qV[K^+]_o}{KT\delta}}{\dfrac{J_K}{P_K \delta} + \dfrac{qV[K^+]_i}{KT\delta}}\right) \tag{12.25}$$

Removing δ from both sides of Eq. (12.25), bringing the term $-\frac{KT}{qV}$ to the other side of the equation, and then taking the exponential of both sides yields

$$e^{-\frac{qV}{KT}} = \frac{\frac{J_K}{P_K\delta} + \frac{qV[K^+]_o}{KT\delta}}{\frac{J_K}{P_K\delta} + \frac{qV[K^+]_i}{KT\delta}} \tag{12.26}$$

Solving for J_K in Eq. (12.26) gives

$$J_K = \frac{qVP_K}{KT}\left(\frac{[K^+]_o - [K^+]_i e^{\frac{-qV}{KT}}}{e^{\frac{-qV}{KT}} - 1}\right) \tag{12.27}$$

Chlorine Ions

The same derivation carried out for K^+ can be repeated for Cl^-, which yields

$$J_{Cl} = \frac{qVP_{Cl}}{KT}\left(\frac{[Cl^-]_o e^{\frac{-qV}{KT}} - [Cl^-]_i}{e^{\frac{-qV}{KT}} - 1}\right) \tag{12.28}$$

where P_{Cl} is the permeability for Cl^-.

Summarizing for Potassium and Chlorine Ions

Using space charge neutrality, $J_K = J_{Cl}$, and Eqs. (12.27) and (12.28) gives

$$P_K\left([K^+]_o - [K^+]_i e^{\frac{-qV}{KT}}\right) = P_{Cl}\left([Cl^-]_o e^{\frac{-qV}{KT}} - [Cl^-]_i\right) \tag{12.29}$$

Solving for the exponential terms yields

$$e^{\frac{-qV}{KT}} = \frac{P_K[K^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Cl}[Cl^-]_o} \tag{12.30}$$

Solving for V gives

$$V = v_o - v_i = -\frac{KT}{q}\ln\left(\frac{P_K[K^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Cl}[Cl^-]_o}\right) \tag{12.31}$$

or in terms of V_m

$$V_m = \frac{KT}{q}\ln\left(\frac{P_K[K^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Cl}[Cl^-]_o}\right) \tag{12.32}$$

This equation is called the *Goldman equation*. Since sodium is also important in membrane potential, the Goldman equation for K^+, Cl^-, and Na^+ can be derived as

$$V_m = \frac{KT}{q}\ln\left(\frac{P_K[K^+]_o + P_{Na}[Na^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Na}[Na^+]_i + P_{Cl}[Cl^-]_o}\right) \tag{12.33}$$

where P_{Na} is the permeability for Na^+. To derive Eq. (12.33), first find J_{Na} and then use space charge neutrality $J_K + J_{Na} = J_{Cl}$. Equation (12.33) then follows. In general, when the

TABLE 12.1 Approximate Intracellular and Extracellular Concentrations of the Important Ions across a Squid Giant Axon, Ratio of Permeabilities, and Nernst Potentials

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities	Nernst Potential (mV)
K^+	400	20	1	−74
Na^+	50	440	0.04	55
Cl^-	52	560	0.45	−60

Note: Permeabilities are relative—that is, P_K:P_{Na}:P_{Cl}—and not absolute. Data were recorded at 6.3°C, resulting in KT/q approximately equal to 25.3 mV.

TABLE 12.2 Approximate Intracellular and Extracellular Concentrations of the Important Ions across a Frog Skeletal Muscle, Ratio of Permeabilities, and Nernst Potentials

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities	Nernst Potential (mV)
K^+	140	2.5	1.0	–105
Na^+	13	110	0.019	56
Cl^-	3	90	0.381	–89

Note: Data were recorded at room temperature, resulting in KT/q approximately equal to 26 mV.

permeability to one ion is exceptionally high, as compared with the other ions, then V_m predicted by the Goldman equation is very close to the Nernst equation for that ion.

Tables 12.1 and 12.2 contain the important ions across the cell membrane, the ratio of permeabilities, and Nernst potentials for the squid giant axon and frog skeletal muscle. The squid giant axon is extensively reported and used in experiments due to its large size, lack of myelination, and ease of use. In general, the intracellular and extracellular concentration of ions in vertebrate neurons is approximately three to four times less than the squid giant axon.

EXAMPLE PROBLEM 12.2

Calculate V_m for the squid giant axon at 6.3°C.

Solution

Using Equation (12.33) and the data in Table 12.1 gives

$$V_m = 25.3 \times \ln\left(\frac{1 \times 20 + 0.04 \times 440 + 0.45 \times 52}{1 \times 400 + 0.04 \times 50 + 0.45 \times 560}\right) mV = -60\, mV$$

12.4.5 Ion Pumps

At rest, separation of charge and ionic concentrations across the cell membrane must be maintained, or V_m changes. That is, the flow of charge into the cell must be balanced by the flow of charge out of the cell. For Na^+, the concentration and electric gradient creates a force

that drives Na^+ into the cell at rest. At V_m, the K^+ force due to diffusion is greater than that due to drift and results in an efflux of K^+ out of the cell. Space charge neutrality requires that the influx of Na^+ be equal to the flow of K^+ out of the cell. Although these flows cancel out each other and space charge neutrality is maintained, this process cannot continue unopposed. Otherwise, $[K^+]_i$ goes to zero as $[Na^+]_i$ increases, with subsequent change in V_m as predicted by the Goldman equation.

Any change in the concentration gradient of K^+ and Na^+ is prevented by the *Na-K* pump. The pump transports a steady stream of Na^+ out of the cell and K^+ into the cell. Removal of Na^+ from the cell is against its concentration and electric gradient and is accomplished with an active pump that consumes metabolic energy. The biochemical reactions that govern the *Na-K* pump are given in Chapter 8. The *Na-K* pump is also used to maintain cell volume. Figure 12.10 illustrates an *Na-K* pump along with passive channels.

The *Na-K* pump has been found to be electrogenic—that is, there is a net transfer of charge across the membrane. Nonelectrogenic pumps operate without any net transfer of charge. For many neurons, the *Na-K* ion pump removes three Na^+ ions for every two K^+ ions moved into the cell, which makes V_m slightly more negative than predicted with only passive channels.

In general, when the cell membrane is at rest, the active and passive ion flows are balanced and a permanent potential exists across a membrane only under the following conditions:

1. The membrane is impermeable to some ion(s).
2. An active pump is present.

The presence of the *Na-K* pump forces V_m to a given potential based on the $K^+$and Na^+ concentrations that are determined by the active pump. Other ion concentrations are determined by V_m. For instance, since Cl^- moves across the membrane only through passive channels, the Cl^- concentration ratio at rest is determined from the Nernst equation with $E_{Cl} = V_m$, or

$$\frac{[Cl^-]_i}{[Cl^-]_o} = e^{\frac{qV_m}{KT}} \tag{12.34}$$

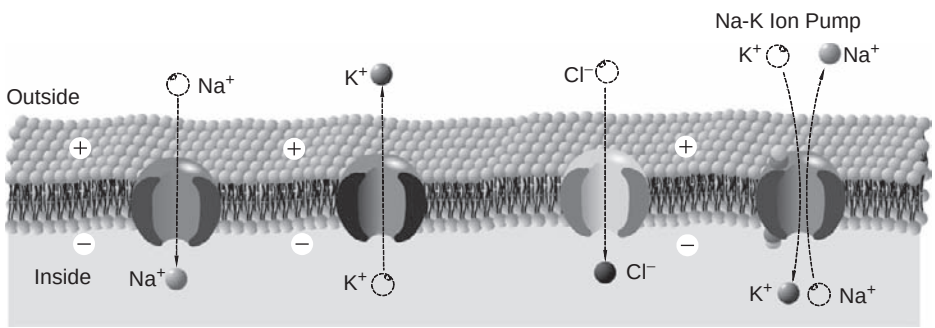


FIGURE 12.10 An active pump is illustrated along with passive channels.

EXAMPLE PROBLEM 12.3

Consider a membrane in which there is an active K^+ pump, passive channels for K^+ and Cl^-, and a non-steady-state initial concentration of $[KCl]$ on both sides of the membrane. Find an expression for the active K^+ pump.

Solution

From space charge neutrality, $J_{Cl} = J_K$, or

$$J_K = J_p - \mu_K Z_K [K^+] \frac{dv}{dx} - \frac{KT}{q} \mu_K \frac{d[K^+]}{dx}$$

$$J_{Cl} = -\mu_{Cl} Z_{Cl} [Cl^-] \frac{dv}{dx} - \frac{KT}{q} \mu_{Cl} \frac{d[Cl^-]}{dx}$$

where J_p is the flow due to the active K^+ pump. Solving for $\frac{dv}{dx}$ using the J_{Cl} equation with $Z_{Cl} = -1$ gives

$$\frac{dv}{dx} = \frac{KT}{q[Cl^-]} \frac{d[Cl^-]}{dx}$$

By space charge neutrality, $[Cl^-] = [K^+]$, which allows rewriting the previous equation as

$$\frac{dv}{dx} = \frac{KT}{q[K^+]} \frac{d[K^+]}{dx}$$

At steady state, both flows are zero, and with $Z_K = 1$, the J_K equation with $\frac{dv}{dx}$ substitution is given as

$$\begin{aligned} J_K = 0 &= J_p - \mu_K [K^+] \frac{dv}{dx} - \frac{KT\mu_K}{q} \frac{d[K^+]}{dx} \\ &= J_p - \mu_K [K^+] \frac{KT}{q[K^+]} \frac{d[K^+]}{dx} - \frac{KT\mu_K}{q} \frac{d[K^+]}{dx} \\ &= J_p - \frac{2KT\mu_K}{q} \frac{d[K^+]}{dx} \end{aligned}$$

Moving J_p to the left side of the equation, multiplying both sides by dx, and then integrating yields

$$-\int_0^{\delta} J_p dx = -\frac{2KT\mu_K}{q} \int_{[K^+]_i}^{[K^+]_o} d[K^+]$$

or

$$J_p = \frac{2KT\mu_K}{q\delta} \left([K^+]_o - [K^+]_i \right)$$

Note: If *no* pump was present in this example, then at steady-state, the concentration on both sides of the membrane would be the same.

12.5 EQUIVALENT CIRCUIT MODEL FOR THE CELL MEMBRANE

In this section, an equivalent circuit model is developed using the tools previously developed. Creating a circuit model is helpful when discussing the Hodgkin-Huxley model of an action potential in the next section, a model that introduces voltage- and time-dependent ion channels. As previously described, the nerve has three types of passive electrical characteristics: electromotive force, resistance, and capacitance. The nerve membrane is a lipid bilayer that is pierced by a variety of different types of ion channels, where each channel is characterized as being passive (always open) or active (gates that can be opened). Each ion channel is also characterized by its selectivity. In addition, there is the active *Na-K* pump that maintains V_m across the cell membrane at steady state.

12.5.1 Electromotive, Resistive, and Capacitive Properties

Electromotive Force Properties

The three major ions, K^+, Na^+, and Cl^-, are differentially distributed across the cell membrane at rest using passive ion channels, as illustrated in Figure 12.5. This separation of charge exists across the membrane and results in a voltage potential V_m, as described by Eq. (12.33) (the Goldman equation).

Across each ion-specific channel, a concentration gradient exists for each ion that creates an electromotive force, a force that drives that ion through the channel at a constant rate. The Nernst potential for that ion is the electrical potential difference across the channel and is easily modeled as a battery, as shown in Figure 12.11 for K^+. The same model is applied for Na^+ and Cl^-, with values equal to the Nernst potentials for each.

Resistive Properties

In addition to the electromotive force, each channel also has resistance—that is, it resists the movement of ions through the channel. This is mainly due to collisions with the channel wall, where energy is given up as heat. The term conductance, *G*, measured in siemens (S), which is the ease with which the ions move through the membrane, is typically used to

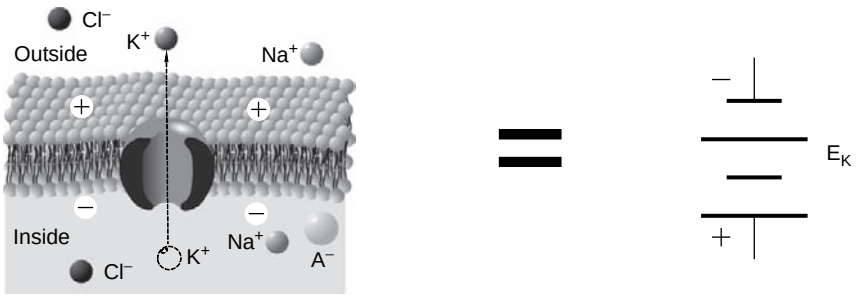


FIGURE 12.11 A battery is used to model the electromotive force for a K^+ channel with a value equal to the K^+ Nernst potential. The polarity of the battery is given with the ground on the outside of the membrane in agreement with convention. From Table 12.1, note that the Nernst potential for K^+ is negative, which reverses the polarity of the battery, driving K^+ out of the cell.

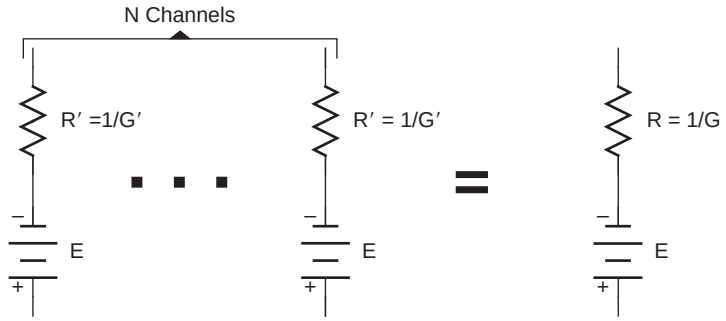


FIGURE 12.12 The equivalent circuit for N ion channels is a single resistor and battery.

represent resistance. Since the conductances (channels) are in parallel, the total conductance is the total number of channels, N, times the conductance for each channel, G':

$$G = N \times G'$$

It is usually more convenient to write the conductance as resistance $R = \frac{1}{G}$, measured in ohms (Ω). An equivalent circuit for the channels for a single ion is now given as a resistor in series with a battery, as shown in Figure 12.12.

Conductance is related to membrane permeability, but they are not interchangeable in a physiological sense. Conductance depends on the state of the membrane, varies with ion concentration, and is proportional to the flow of ions through a membrane. Permeability describes the state of the membrane for a particular ion. Consider the case in which there are no ions on either side of the membrane. No matter how many channels are open, G = 0 because there are no ions available to flow across the cell membrane (due to a potential difference). At the same time, ion permeability is constant and determined by the state of the membrane.

Equivalent Circuit for Three Ions

Each of the three ions $K^+, Na^+,$ and Cl^- are represented by the same equivalent circuit, as shown in Figure 12.12, with Nernst potentials and appropriate resistances. Combining the three equivalent circuits into one circuit with the extracellular fluid and cytoplasm connected by short circuits completely describes a membrane at rest (Figure 12.13).

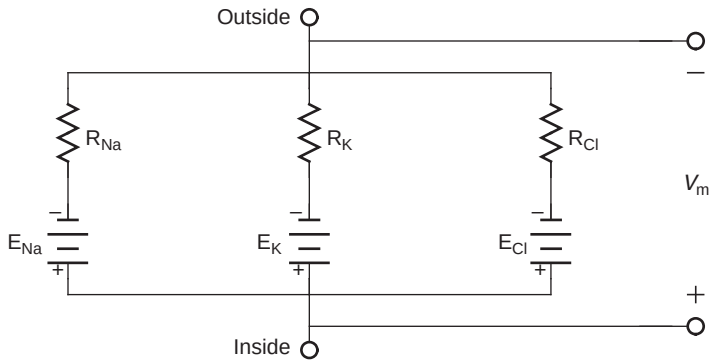


FIGURE 12.13 Model of the passive channels for a small area of nerve at rest, with each ion channel represented by a resistor in series with a battery.

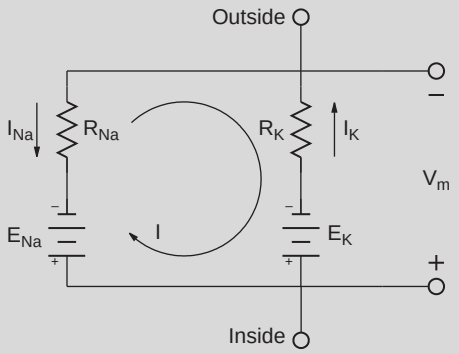
EXAMPLE PROBLEM 12.4

Find V_m for the frog skeletal muscle (Table 12.2) if the Cl^- channels are ignored. Use $R_K = 1.7\ \text{k}\Omega$ and $R_{Na} = 15.67\ \text{k}\Omega$.

Solution

The following diagram depicts the membrane circuit with mesh current I, current I_{Na} through the sodium channel, and current I_K through the potassium channel. Current I is found using mesh analysis:

$$E_{Na} + IR_{Na} + IR_K - E_K = 0$$



and then solving for I,

$$I = \frac{E_K - E_{Na}}{R_{Na} + R_K} = \frac{(-105 - 56) \times 10^{-3}}{(15.67 + 1.7) \times 10^3} = -9.27\ \mu A$$

yields

$$V_m = E_{Na} + IR_{Na} = -89\ mV$$

Notice that $I = -I_{Na}$ and $I = -I_K$, or $I_{Na} = I_K$ as expected. Physiologically, this implies that the inward Na^+ current is exactly balanced by the outward bound K^+ current.

EXAMPLE PROBLEM 12.5

Find Vm for the frog skeletal muscle if $R_{Cl} = 3.125\ \text{k}\Omega$.

Solution

To solve, first find a Thevenin's equivalent circuit for the circuit in Example Problem 12.4:

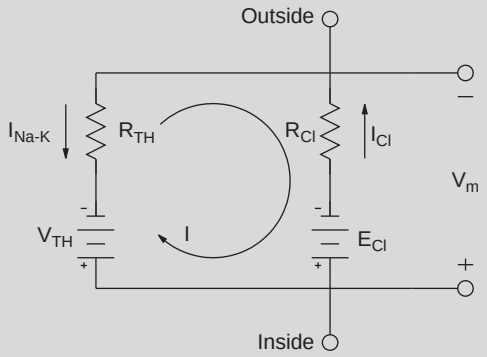
$$V_{TH} = V_m = -89\ \text{mV}$$

and

$$R_{TH} = \frac{R_{Na} \times R_K}{R_{Na} + R_K} = 1.534\ \text{k}\Omega$$

Continued

The Thevenin's equivalent circuit is shown in the following figure.



Since $E_{Cl} = V_{TH}$ according to Table 12.2, no current flows. This is the actual situation in most nerve cells. The membrane potential is determined by the relative conductances and Nernst potentials for K^+ and Na^+. The Nernst potentials are maintained by the *Na-K* active pump that maintains the concentration gradient. Cl^- is usually passively distributed across the membrane.

The Na-K *Pump*

As shown in Example Problem 12.4 and Section 12.4, there is a steady flow of K^+ ions out of the cell and Na^+ ions into the cell even when the membrane is at the resting potential. Left unchecked, this would drive E_K and E_{Na} toward 0. To prevent this, current generators depicting the *Na-K* pump are used that are equal to and the opposite of the passive currents and incorporated into the model, as shown in Figure 12.14.

12.5.2 Capacitive Properties

Capacitance occurs whenever electrical conductors are separated by an insulating material. In the neuron, the cytoplasm and extracellular fluid are the electrical conductors,

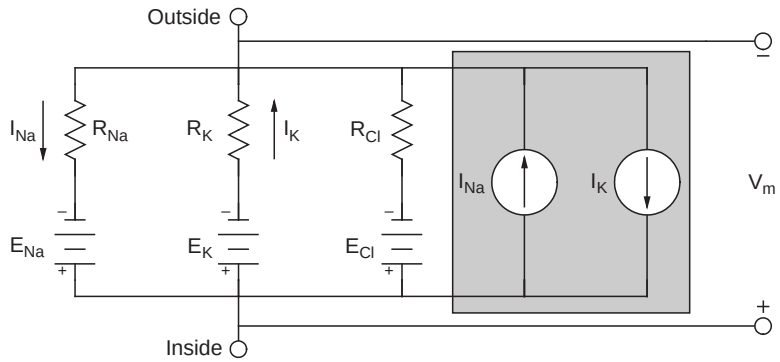


FIGURE 12.14 Circuit model of the three passive channels for a small area of the nerve at rest with each ion channel represented by a resistor in series with a battery. The *Na-K* active pump is modeled as two current sources within the shaded box.

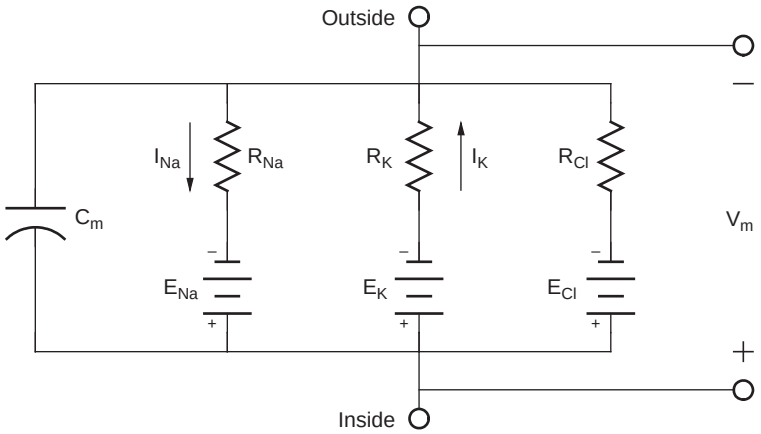


FIGURE 12.15 Circuit model of a small area of the nerve at rest with all of its passive electrical properties. The *Na-K* active pump shown in Figure 12.14 is removed, since it does not contribute electrically to the circuit.

and the lipid bilayer of the membrane is the insulating material (see Figure 12.3). Capacitance for a neuron membrane is approximately $1\mu F/cm^2$. Membrane capacitance implies that ions do not move through the membrane except through ion channels.

The membrane can be modeled using the circuit in Figure 12.15 by incorporating membrane capacitance with the electromotive and resistive properties. A consequence of membrane capacitance is that changes in membrane voltage are not immediate but follow an exponential time course due to first-order time constant effects. To appreciate the effect of capacitance, the circuit in Figure 12.15 is reduced to Figure 12.16 by using a Thevenin's equivalent for the batteries and the resistors with R_{TH} and V_{TH} given in Eqs. (12.35) and (12.36).

$$R_{TH} = \frac{1}{\frac{1}{R_K} + \frac{1}{R_{Na}} + \frac{1}{R_{Cl}}} \tag{12.35}$$

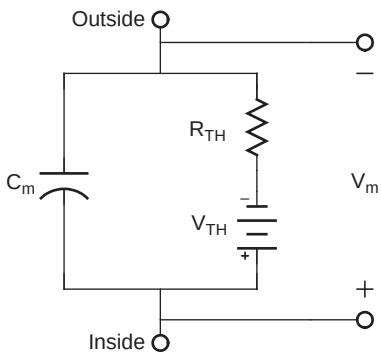


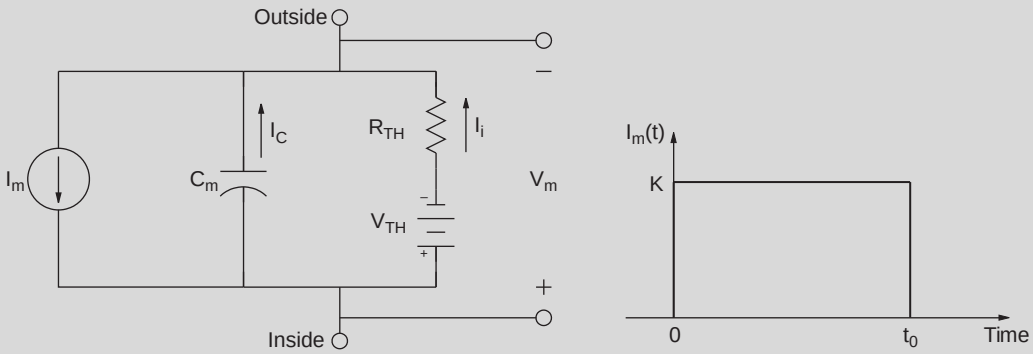
FIGURE 12.16 Thévenin's equivalent circuit of the model in Figure 12.15.

$$V_{TH} = \frac{R_{Na}R_{Cl}E_K + R_K R_{Cl}E_{Na} + R_K R_{Na}E_{Cl}}{R_{Na}R_{Cl} + R_K R_{Cl} + R_K R_{Na}} \tag{12.36}$$

The time constant for the membrane circuit model is $\tau = R_{TH}C_m$, and at 5τ the response is within 1 percent of steady-state. The range for τ is from 1 to 20 ms in a typical neuron. In addition, at steady-state the capacitor acts as an open circuit and $V_{TH} = V_m$, as it should.

EXAMPLE PROBLEM 12.6

Compute the change in V_m due to a current pulse through the cell membrane.



Solution

Experimentally, the stimulus current is a pulse passed through the membrane from an intracellular electrode to an extracellular electrode, as depicted in the circuit diagram above. The membrane potential, V_m, due to a current pulse, I_m, with amplitude K and duration t_0 applied at $t = 0$, is found by applying Kirchhoff's current law at the cytoplasm, yielding

$$-I_m + \frac{V_m - V_{TH}}{R_{TH}} + C_m \frac{dV_m}{dt} = 0$$

The Laplace transform of the node equation is

$$-I_m(s) + \frac{V_m(s)}{R_{TH}} - \frac{V_{TH}}{sR_{TH}} + sC_m V_m(s) - C_m V_m(0^+) = 0$$

Combining common terms gives

$$\left(s + \frac{1}{C_m R_{TH}}\right) V_m(s) = V_m(0^+) + \frac{I_m(s)}{C_m} + \frac{V_{TH}}{sC_m R_{TH}}$$

The Laplace transform of the current pulse is $I_m(s) = \frac{K}{s}\left(1 - e^{-st_0}\right)$. Substituting $I_m(s)$ into the node equation and rearranging terms yields

$$V_m(s) = \frac{V_m(0^+)}{\left(s + \frac{1}{C_m R_{TH}}\right)} + \frac{K(1 - e^{st_0})}{sC_m\left(s + \frac{1}{C_m R_{TH}}\right)} + \frac{V_m}{sC_m R_{TH}\left(s + \frac{1}{C_m R_{TH}}\right)}$$

Performing a partial fraction expansion, and noting that $V_m(0^+) = V_{TH}$, gives

$$V_m(s) = KR_{TH}\left(\frac{1}{s} - \frac{1}{\left(s + \frac{1}{C_m R_{TH}}\right)}\right)\left(1 - e^{-st_0}\right) + \frac{V_{TH}}{s}$$

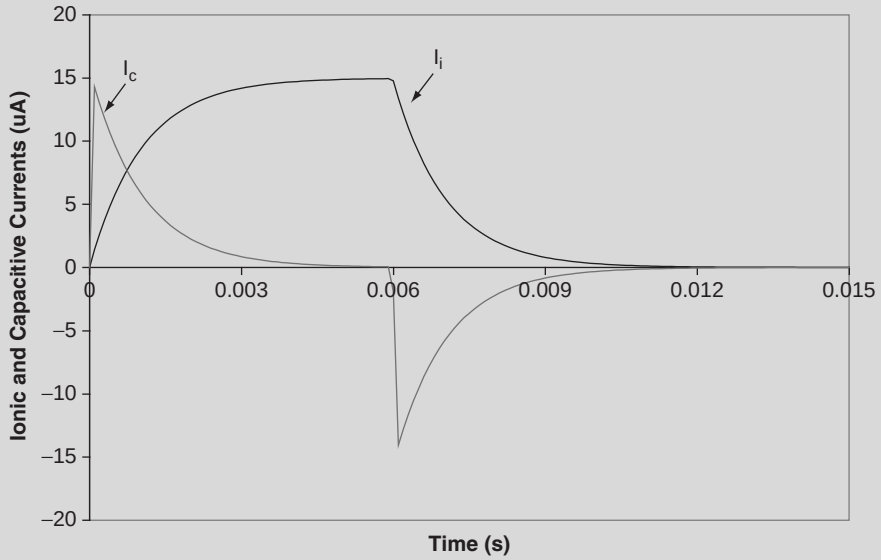
Transforming back into the time domain yields the solution

$$V_m(t) = V_{TH} + R_{TH}K\left(1 - e^{-\frac{t}{R_{TH}C_m}}\right)u(t) - R_{TH}K\left(1 - e^{-\frac{t-t_0}{R_{TH}C_m}}\right)u(t - t_0)$$

The ionic current (I_i) and capacitive current (I_c) are shown in the following figure, where

$$I_i = \frac{V_m - V_{TH}}{R_{Th}}$$

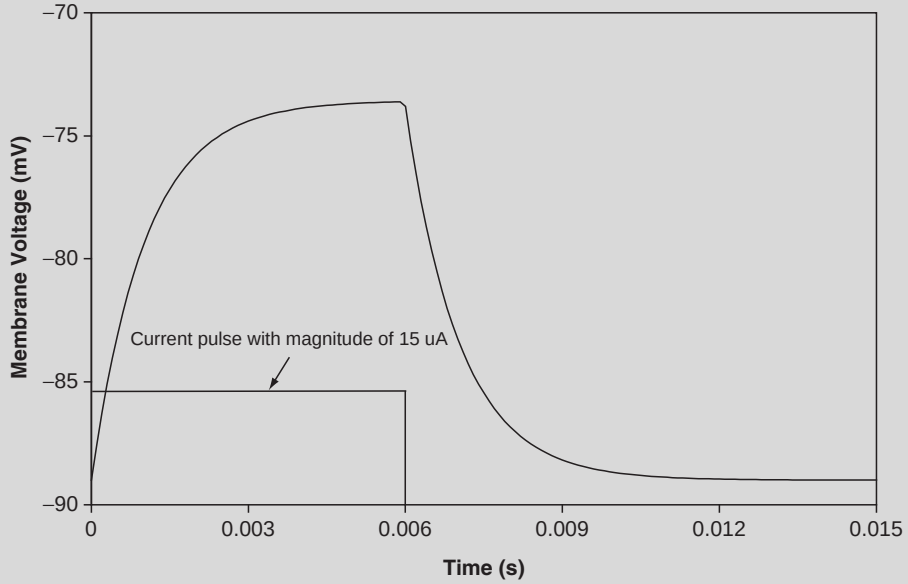
$$I_c = C_m \frac{dV_m}{dt}$$

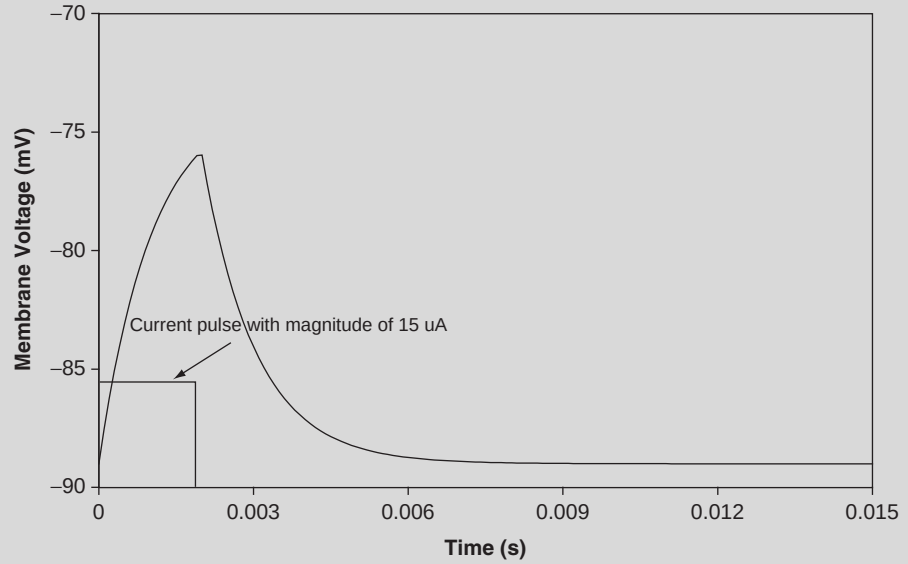


Shown in the following figures are graphs of V_m in response to a 15 μA current pulse of 6 ms (top) and 2 ms (bottom) using parameters for the frog skeletal muscle. The time constant is

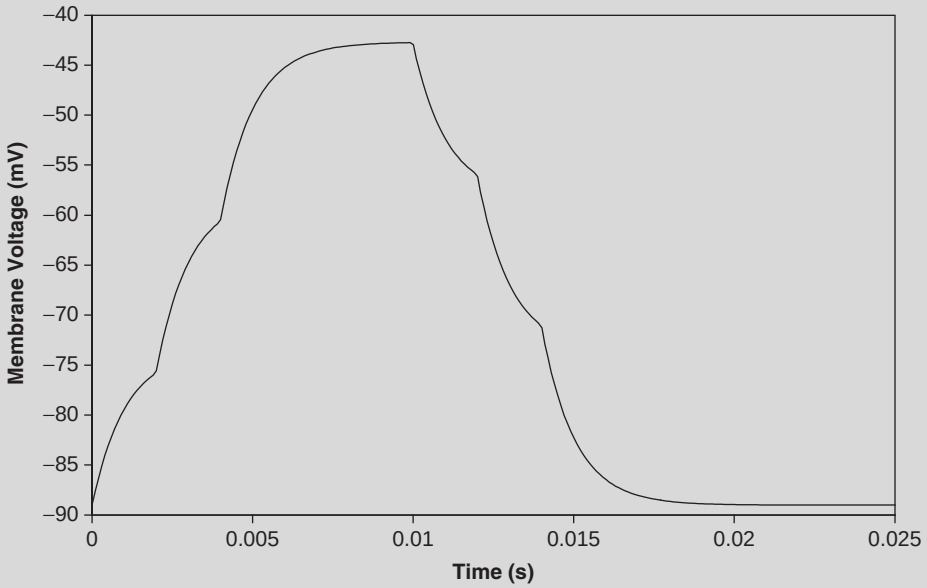
Continued

approximately 1 ms. Note that in the figure on the top, V_m reaches steady state before the current pulse returns to zero, and in the figure on the bottom, V_m falls short of the steady state value reached on the top. The value of the time constant is important in the integration of currents (packets of neurotransmitter) at the synapse. Each packet of neurotransmitter acts as a current pulse, as described in Section 12.8. Note that the longer the time constant, the more time the membrane is excited. Most excitations are not synchronous, but because of τ, a significant portion of the stimulus is added together that might cause an action potential.





The following figure is due to a series of 15 μA current pulses of 6 ms duration, with the onsets occurring at 0, 2, 4, 6, and 8 ms. Since the pulses occur within 5τ of the previous pulses, the effect of each on V_m is additive, allowing the membrane to depolarize to approximately −45 mV. If the pulses are spaced at intervals greater than 5τ, then V_m would be a series of pulse responses, as previously illustrated.



12.5.3 Changes in Membrane Potential with Distance

The circuit model in Figure 12.15 or 12.16 describes a small area or section of the membrane. In Example Problem 12.6, a current pulse was injected into the membrane and resulted in a change in V_m. The change in V_m in this section of the membrane causes current to flow into the adjacent membrane sections, which causes a change in V_m in each section and so on, continuing throughout the surface of the membrane. Since the volume inside the dendrite is much smaller than the extracellular space, there is significant resistance to the flow of current in the cytoplasm from one membrane section to the next as compared with the flow of current in the extracellular space. The larger the diameter of the dendrite, the smaller the resistance to the spread of current from one section to the next. To model this effect, a resistor, R_a, is placed in the cytoplasm connecting each section together, as shown in Figure 12.17. This model is actually a three-dimensional surface and continues in the x, y, and z directions. The outside resistance is negligible, since it has a greater volume and is modeled as a short circuit.

Suppose a current is injected into a section of the dendrite as shown in Figure 12.18, similar to the situation in Example Problem 12.6, where t_0 is large. At steady state, the transient response due to C_m has expired, and only current through the resistance is important. Most of the current flows out through the section into which the current was injected, since it has

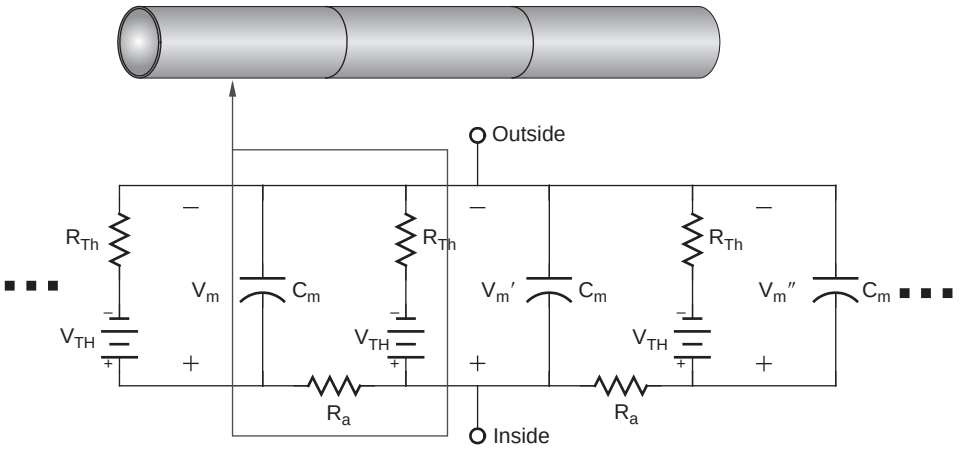


FIGURE 12.17 Equivalent circuit of series of membrane sections connected with axial resistance, R_a.

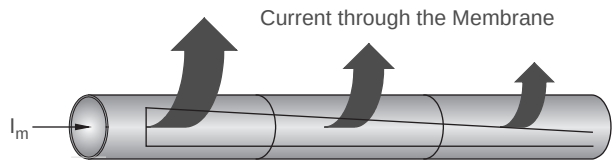


FIGURE 12.18 The flow of current through a dendrite at steady-state. Since current seeks the path of least resistance, most of the current leaves the dendrite at the injection site and becomes smaller with distance from the injection site.

the smallest resistance (R_{TH}) in relation to the other sections. The next largest current flowing out of the membrane occurs in the next section, since it has the next smallest resistance, $R_{TH} + R_a$. The change in V_m, ΔV_m, from the injection site is independent of C_m and depends solely on the relative values of R_{TH} and R_a. The resistance seen n sections from the injection site is $R_{TH} + nR_a$. Since current decreases with distance from the injection site, then ΔV_m also decreases with distance from the injection site because it equals the current through that section times R_{TH}. The change in membrane potential, ΔV_m, decreases exponentially with distance and is given by

$$\Delta V_m = V_o e^{\frac{-x}{\lambda}} \tag{12.37}$$

where $\lambda = \sqrt{\frac{R_{TH}}{R_a}}$ is the membrane length constant, x is the distance away from injection site, and V_o is the change in membrane potential at the injection site.[1] The range of values for λ is 0.1 to 1 mm. The larger the value of λ, the greater the effect of the stimulation along the length of the membrane.

[1]Equation 12.37 is the solution of the one dimensional cable equation for the dendrite using partial differential equations (see Keener and Sneyd for details).

12.6 THE HODGKIN-HUXLEY MODEL OF THE ACTION POTENTIAL

Alan Lloyd Hodgkin and Andrew Fielding Huxley published five papers in 1952 that described a series of experiments and an empirical model of an action potential in a squid giant axon. Their first four papers described the experiments that characterized the changes in the cell membrane that occurred during the action potential. The last paper presented the empirical model. The empirical model they developed is not a physiological model based on the laws and theory developed in this chapter but a model based on curve fitting using an exponential function. In this section, highlights of the Hodgkin-Huxley experiments are presented along with the empirical model. All of the figures presented in this section were simulated using SIMULINK and the Hodgkin-Huxley empirical model parameterized with their squid giant axon data.

12.6.1 Action Potentials and the Voltage Clamp Experiment

The ability of nerve cells to conduct action potentials makes it possible for signals to be transmitted over long distances within the nervous system. An important feature of the action potential is that it does not decrease in amplitude as it is conducted away from its site of initiation. An action potential occurs when V_m reaches a value called the *threshold potential* at the axon hillock (see Figure 12.1). Once V_m reaches threshold, time- and voltage-dependent conductance changes occur in the active Na^+ and K^+ gates that drive V_m toward E_{Na}, then back to E_K, and finally to the resting potential. These changes in conductance were first described by Hodgkin and Huxley (and Katz as a coauthor on one paper and a collaborator on several others). Figure 12.19 illustrates a stylized action potential with the threshold potential at approximately −40 mV.

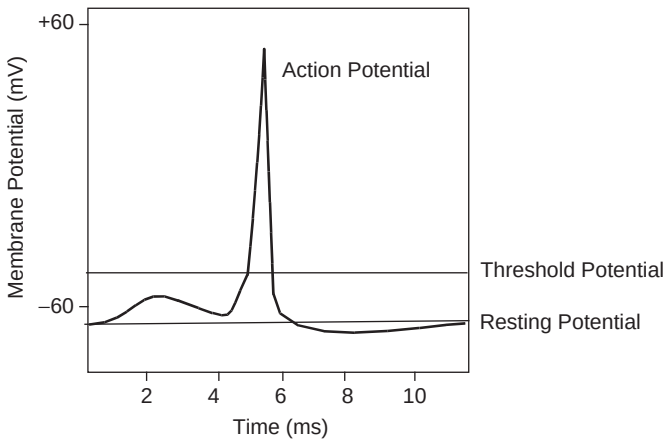


FIGURE 12.19 Stylized diagram of an action potential once threshold potential is reached at approximately 5 ms. The action potential is due to voltage and time-dependent changes in conductance. The action potential rise is due to Na^+, and the fall is due to K^+ conductance changes.

Stimulation of the postsynaptic membrane along the dendrite and cell body must occur for V_m to rise to the threshold potential at the axon hillock. As previously described, the greater the distance from the axon hillock, the smaller the contribution of postsynaptic membrane stimulation to the change in V_m at the axon hillock. Also, because of the membrane time constant, there is a time delay in stimulation at the postsynaptic membrane and the resultant change in V_m at the axon hillock. Thus, time and distance are important functions in describing the graded response of V_m at the axon hillock.

Once V_m reaches threshold, active Na^+ conductance gates are opened and an inward flow of Na^+ ions results, causing further depolarization. This depolarization increases Na^+ conductance, consequently inducing more Na^+ current. This iterative cycle, shown in Figure 12.20, continues driving V_m to E_{Na} and concludes with the closure of the Na^+ gates. A similar but slower change in K^+ conductance occurs that drives V_m back to the resting potential. Once an action potential is started, it continues until completion. This is called the "all or none" phenomenon. The active gates for Na^+ and K^+ are both functions of V_m and time.

The action potential moves through the axon at high speeds and appears to jump from one node of Ranvier to the next in myelinated neurons. This occurs because the membrane capacitance of the myelin sheath is very small, making the membrane appear only resistive with almost instantaneous changes in V_m possible.

To investigate the action potential, Hodgkin and Huxley used an unmyelinated squid giant axon in their studies because of its large diameter (up to 1 mm) and long survival time of several hours in seawater at 6.3°C. Their investigations examined the then existing theory that described an action potential as due to enormous changes in membrane permeability that allowed all ions to freely flow across the membrane, driving V_m to zero. As they discovered, this was not the case. The success of the Hodgkin-Huxley studies was based on two new experimental techniques, the space clamp and voltage clamp, and collaboration with Cole and Curtis from Columbia University.

The space clamp allowed Hodgkin and Huxley to produce a constant V_m over a large region of the membrane by inserting a silver wire inside the axon and thus eliminating R_a. The voltage clamp allowed the control of V_m by eliminating the effect of further depolarization due to the influx of I_{Na} and efflux of I_K as membrane permeability changed. Selection of the squid giant axon was fortunate for two reasons: it was large and survived a very long time in seawater, and it had only two types of voltage-time-dependent permeable channels. Other types of neurons have more than two voltage-time-dependent permeable channels that would have made the analysis extremely difficult or even impossible.

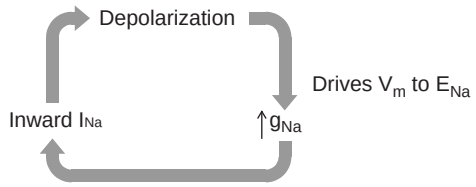


FIGURE 12.20 The conductance gate for sodium.

Voltage Clamps

To study the variable voltage-time-resistance channels for K^+ and Na^+, Hodgkin and Huxley used a voltage clamp to separate these two dynamic mechanisms so only the time-dependent features of the channel were examined. Figure 12.21 shows the voltage clamp experiment by using the equivalent circuit model previously described. The channels for K^+ and Na^+ are represented using variable voltage-time resistances, and the passive gates for Na^+, K^+, and Cl^- are given by a leakage channel with resistance R_l (that is, the Thevenin's equivalent circuit of the passive channels). The function of the voltage clamp is to suspend the interaction between Na^+ and K^+ channel resistance and the membrane potential, as shown in Figure 12.22. If the membrane voltage is not clamped, then changes in Na^+ and K^+ channel resistance modify membrane voltage, which then changes Na^+ and K^+ channel resistance, and so on, as previously described.

A voltage clamp is created by using two sets of electrodes, as shown in Figure 12.23. In an experiment, one pair injects current, I_m, to keep V_m constant, and another pair is used to observe V_m. To estimate the conductance in the Na^+ and K^+ channels, I_m is also measured during the experiment. Meters for recording V_m and I_m are shown in Figure 12.21. They

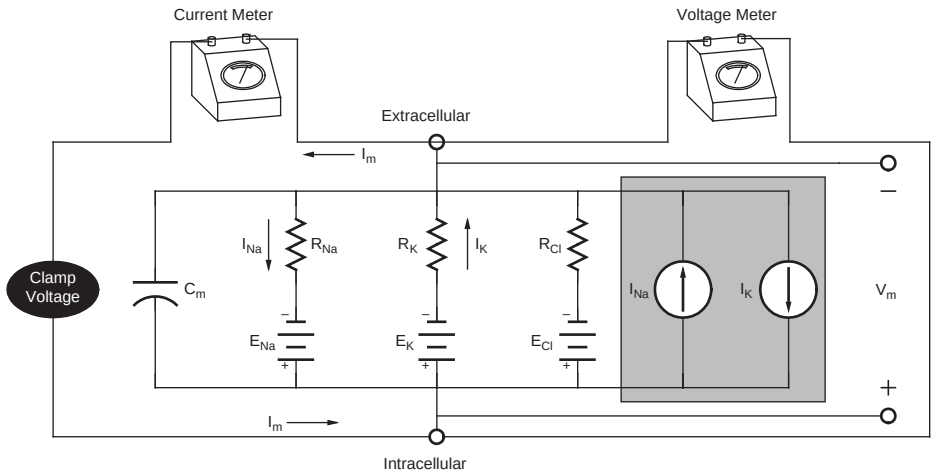


FIGURE 12.21 Equivalent circuit model of an unmyelinated section of squid giant axon under voltage clamp conditions. The channels for K^+ and Na^+ are now represented using variable voltage-time resistances, and the passive gates for Na^+, K^+, and Cl^- are given by a leakage channel with resistance R_l. The *Na-K* pump is illustrated within the shaded area of the circuit. In the experiment, the membrane is immersed in the seawater bath.

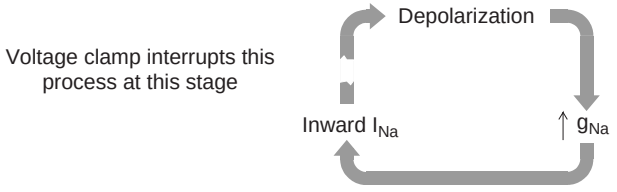


FIGURE 12.22 Voltage clamp experiment interrupts the cycle shown in Figure 12.20.

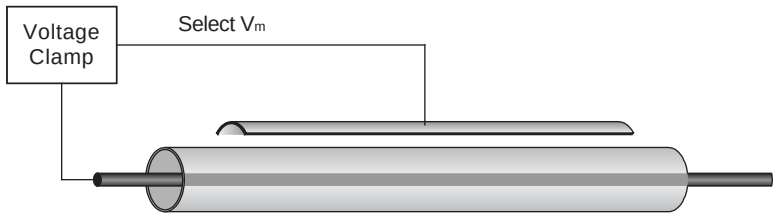


FIGURE 12.23 Physical set up for the voltage clamp experiment.

are placed outside the seawater bath. Today, these would be connected to an analog-to-digital converter (ADC) with data stored in a hard disk of a computer. Back in 1952, these meters were strip chart recorders. The application of a clamp voltage, V_c, causes a change in Na^+ conductance that results in an inward flow of Na^+ ions. This causes the membrane potential to be more positive than V_c. The clamp removes positive ions from inside the cell, which results in no net change in V_m. The current, I_m, is the dependent variable in the voltage clamp experiment, and V_c is the independent variable.

To carry out the voltage clamp experiment, the investigator first selects a clamp voltage and then records the resultant membrane current, I_m, that is necessary to keep V_m at the clamp voltage. Figure 12.24 shows the resulting I_m due to a clamp voltage of -20 mV. Initially, the step change in V_m causes a large current to pass through the membrane that is primarily due to the capacitive current. The clamp voltage also creates a constant leakage current through the membrane that is equal to

$$I_l = \frac{V_c - E_l}{R_l} \tag{12.38}$$

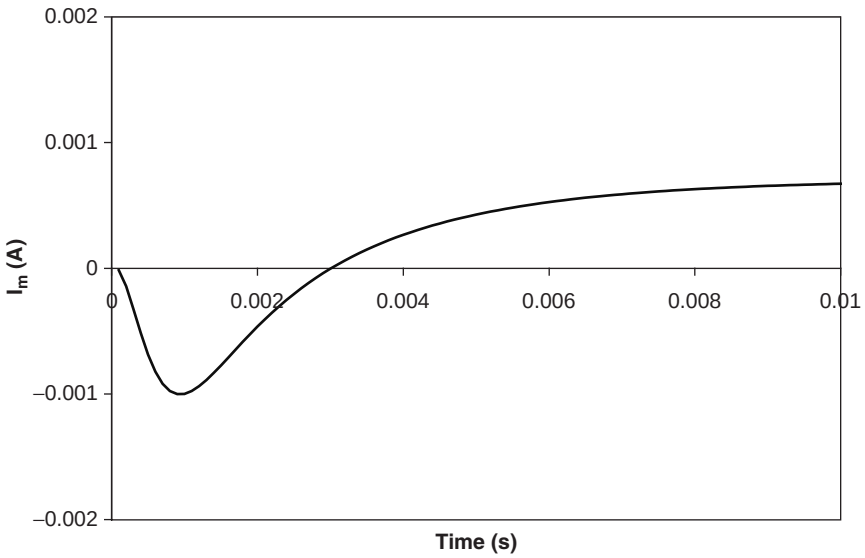


FIGURE 12.24 Membrane current I_m due to a -20 mV voltage clamp.

Subtracting both the capacitive and leakage current from I_m leaves only the Na^+ and K^+ currents. To separate the Na^+ and K^+ currents, Hodgkin and Huxley substituted a large impermeable cation for Na^+ in the external solution. This eliminated the Na^+ current and left only the K^+ current. Returning the Na^+ current to the external solution allowed the Na^+ current to be estimated by subtracting the capacitive, leakage, and K^+ currents from I_m. The Na^+ and K^+ currents due to a clamp voltage of -20 mV are shown in Figure 12.25. Since the clamp voltage in Figure 12.25 is above threshold, the Na^+ and K^+ channel resistances are engaged and follow a typical profile. The Na^+ current rises to a peak first and then returns to zero as the clamp voltage is maintained. The K^+ current falls to a steady-state current well after the Na^+ current peaks and is maintained at this level until the clamp voltage is removed. This general pattern holds for both currents for all clamp voltages above threshold.

The Na^+ and K^+ channel resistance or conductance is easily determined by applying Ohm's law to the circuit in Figure 12.20 and the current waveforms in Figure 12.21:

$$I_K = \frac{V_m - E_K}{R_K} = G_K(V_m - E_K) \tag{12.39}$$

$$I_{Na} = \frac{E_{Na} - V_m}{R_{Na}} = G_{Na}(E_{Na} - V_m) \tag{12.40}$$

These conductances are plotted as a function of clamp voltages ranging from -50 mV to $+20$ mV in Figure 12.26.

For all clamp voltages above threshold, the rate of onset for opening Na^+ channels is more rapid than for K^+ channels, and the Na^+ channels close after a period of time, while

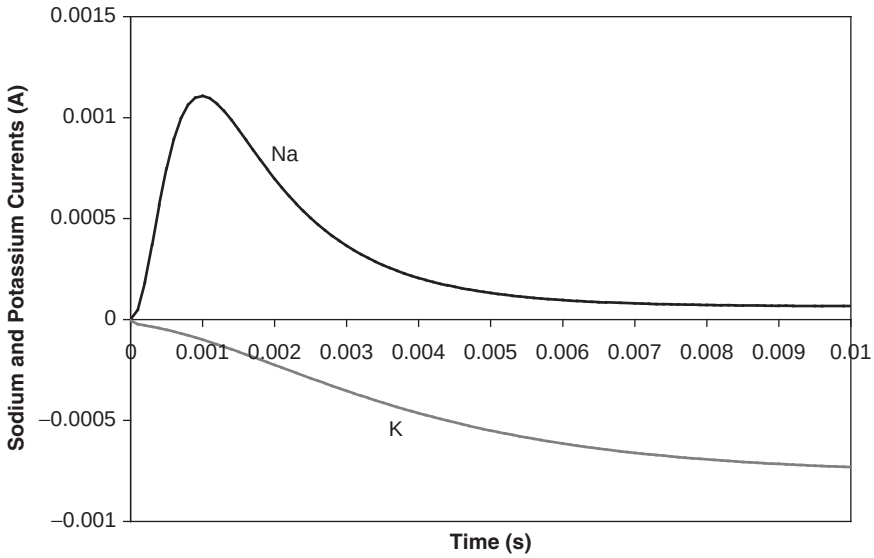


FIGURE 12.25 Diagram illustrating sodium and potassium currents due to a -20 mV voltage clamp.

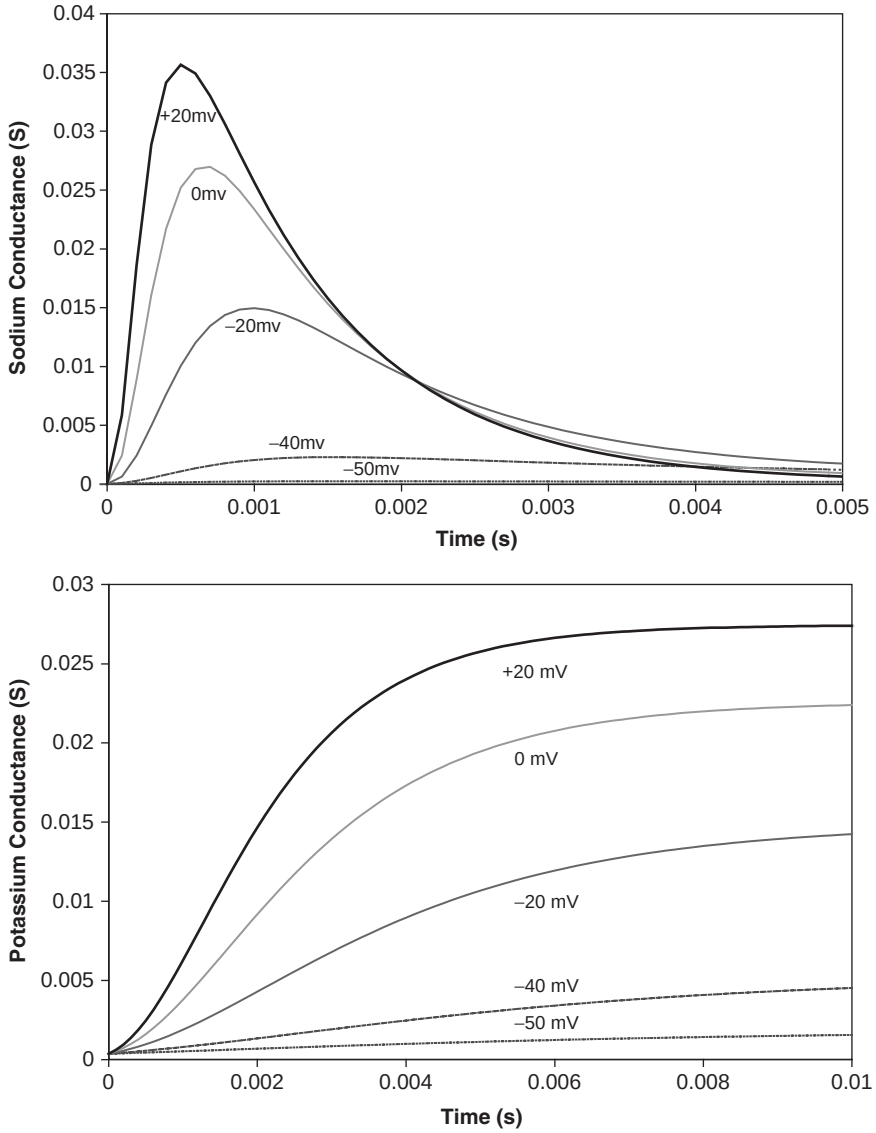


FIGURE 12.26 Diagram illustrating the change in Na^+ and K^+ conductance with clamp voltage ranging from −50 mV (below threshold) to +20 mV. Note that the time scales are different in the two conductance plots.

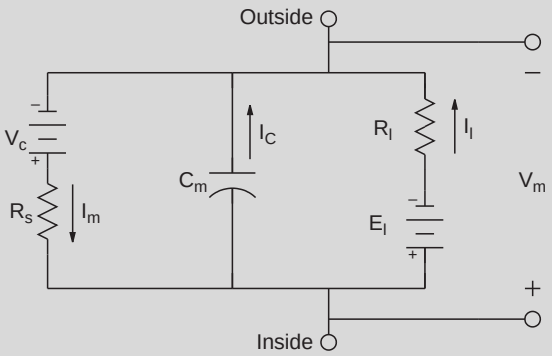
K^+ channels remain open while the voltage clamp is maintained. Once the Na^+ channels close, they cannot be opened until the membrane has been hyperpolarized to its resting potential. The time spent in the closed state is called the refractory period. If the voltage clamp is turned off before the time course for Na^+ is complete (returns to zero), G_{Na} almost immediately returns to zero, and G_K returns to zero slowly regardless of whether the time course for Na^+ is complete.

EXAMPLE PROBLEM 12.7

Compute I_c and I_l through a cell membrane for a subthreshold clamp voltage.

Solution

Assume that the Na^+ and K^+ voltage-time-dependent channels are not activated because the stimulus is below threshold. This eliminates these gates from the analysis although this is not actually true as shown in Example Problem 12.9. The cell membrane circuit is given by



where R_s is the resistance of the wire. Applying Kirchhoff's current law at the cytoplasm gives

$$C_m \frac{dV_m}{dt} + \frac{V_m - E_l}{R_l} + \frac{V_m - V_c}{R_s} = 0$$

Rearranging the terms in the previous equation yields

$$C_m \frac{dV_m}{dt} + \frac{R_l + R_s}{R_l R_s} V_m = \frac{R_l V_c + R_s E_l}{R_l R_s}$$

With the initial condition $V_m(0) = E_l$, the solution is given by

$$V_m = \frac{R_l V_c + R_s E_l}{R_l + R_s} + \frac{R_l (E_l - V_c)}{R_l + R_s} e^{-\frac{(R_l + R_s)t}{R_l R_s C_m}}$$

Now

$$I_c = C_m \frac{dV_m}{dt} = \frac{E_l - V_c}{R_s} e^{-\frac{(R_l + R_s)t}{R_l R_s C_m}}$$

and $I_l = \dfrac{V_m - E_l}{R_l}$. At steady-state, $I_l = \dfrac{V_c - E_l}{R_l + R_s}$.

Reconstructing the Action Potential

By analyzing the estimated G_{Na} and G_K from voltage clamp pulses of various amplitudes and durations, Hodgkin and Huxley were able to obtain a complete set of nonlinear empirical equations that described the action potential. Simulations using these

equations accurately describe an action potential in response to a wide variety of stimulations. Before presenting these equations, it is important to qualitatively understand the sequence of events that occur during an action potential by using previously described data and analyses. The start of an action potential begins with a depolarization above threshold that causes an increase in G_{Na} and results in an inward Na^+ current. The Na^+ current causes a further depolarization of the membrane, which then increases the Na^+ current. This continues to drive V_m to the Nernst potential for Na^+. As shown in Figure 12.26, G_{Na} is a function of both time and voltage, which peaks and then falls to zero.

During the time it takes for G_{Na} to return to zero, G_K continues to increase, which hyperpolarizes the cell membrane and drives V_m from E_{Na} toward E_K. The increase in G_K results in an outward K^+ current. The K^+ current causes further hyperpolarization of the membrane, which then increases K^+ current. This continues to drive V_m to the Nernst potential for K^+, which is below resting potential. Figure 12.27 illustrates the changes in V_m, G_{Na}, and G_K during an action potential.

The circuit shown in Figure 12.16 is a useful tool for modeling the cell membrane during small subthreshold depolarizations. This model assumes that the K^+ and Na^+ currents are small enough to neglect. As illustrated in Example Problem 12.6, a current pulse sent through the cell membrane briefly creates a capacitive current, which decays exponentially and creates an exponentially increasing I_l. Once the current pulse is turned off, capacitive current flows again and exponentially decreases to zero. The leakage current also exponentially decays to zero.

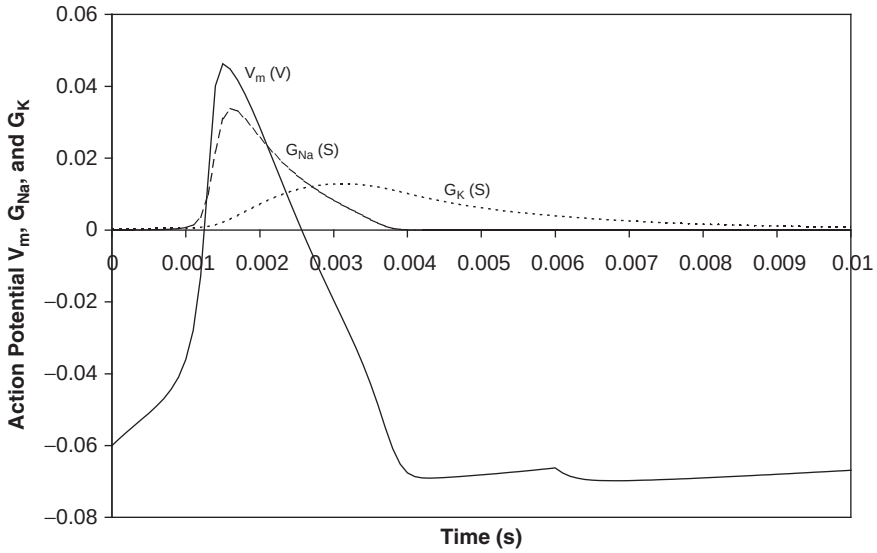


FIGURE 12.27 V_m, G_{Na}, and G_K during an action potential.

As the current pulse magnitude is increased, depolarization of the membrane increases, causing activation of the Na^+ and K^+ voltage-time-dependent channels. For sufficiently large depolarizations, the inward Na^+ current exceeds the sum of the outward K^+ and leakage currents ($I_{Na} > I_K + I_l$). The value of V_m at this current is called *threshold*. Once the membrane reaches threshold, the Na^+ and K^+ voltage-time channels are engaged and run to completion, as shown in Figure 12.27.

If a slow rising stimulus current is used to depolarize the cell membrane, then threshold will be higher. During the slow approach to threshold, inactivation of G_{Na} channels occurs and activation of G_K channels develops before threshold is reached. The value of V_m, where $I_{Na} > I_K + I_l$ is satisfied, is much larger than if the approach to threshold occurs quickly.

12.6.2 Equations Describing G_{Na} and G_K

The empirical equation used by Hodgkin and Huxley to model G_{Na} and G_K is of the form

$$G(t) = \left(A + Be^{-Ct}\right)^D \tag{12.41}$$

Values for the parameters $A, B, C,$ and D were estimated from the voltage clamp data that were collected on the squid giant axon. Not evident in Eq. (12.41) is the voltage dependence of the conductance channels. The voltage dependence is captured in the parameters as described in this section. In each of the conductance models, D is selected as 4 to give a best fit to the data. Figure 12.27 was actually calculated using SIMULINK, a simulation package that is part of MATLAB, and the parameter estimates found by Hodgkin and Huxley. Details concerning the simulation are covered later in this section.

Potassium

The potassium conductance waveform shown in Figure 12.26 is described by a rise to a peak while the stimulus is applied. This aspect is easily included in a model of G_K by using the general Hodgkin-Huxley expression as follows.

$$G_K = \bar{G}_K n^4 \tag{12.42}$$

where $\bar{G}_K$ is maximum K^+ conductance and n is thought of as a rate constant and given as the solution to the following differential equation:

$$\frac{dn}{dt} = \alpha_n(1 - n) - \beta_n n \tag{12.43}$$

where

$$\alpha_n = 0.01\frac{V+10}{e^{\frac{V+10}{10}}-1}$$

$$\beta_n = 0.125e^{\frac{V}{80}}$$

$$V = V_{rp} - V_m$$

V_{rp} is the membrane potential at rest without any membrane stimulation. Note that V is the displacement from resting potential and should be negative. Clearly, G_K is a time-dependent variable, since it depends on Eq. (12.43), and a voltage-dependent variable, since n depends on voltage because of α_n and β_n.

Sodium

The sodium conductance waveform in Figure 12.26 is described by a rise to a peak and a subsequent decline. These aspects are included in a model of G_{Na} as the product of two functions, one describing the rising phase and the other describing the falling phase, and modeled as

$$G_{NA} = \bar{G}_{Na}m^3h \tag{12.44}$$

where $\bar{G}_{Na}$ is maximum Na^+ conductance, and m and h are thought of as rate constants and given as the solutions to the following differential equations:

$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m \tag{12.45}$$

where

$$\alpha_m = 0.1\frac{V+25}{e^{\frac{V+25}{10}}-1}$$

$$\beta_m = 4e^{\frac{V}{18}}$$

and

$$\frac{dh}{dt} = \alpha_h(1-h) - \beta_h h \tag{12.46}$$

where

$$\alpha_h = 0.07e^{\frac{V}{20}}$$

$$\beta_h = \frac{1}{e^{\frac{V+30}{10}}+1}$$

Note that m describes the rising phase and h describes the falling phase of G_{Na}. The units for the $\alpha_i's$ and $\beta_i's$ in Eqs. (12.43), (12.45), and (12.46) are ms^{-1}, while n, m, and h are dimensionless and range in value from 0 to 1.

EXAMPLE PROBLEM 12.8

Calculate G_K and G_{Na} at resting potential for the squid giant axon using the Hodgkin-Huxley model. Parameter values are $\bar{G}_K = 36 \times 10^{-3}$ S and $\bar{G}_{Na} = 120 \times 10^{-3}$ S.

Solution

At resting potential, G_K and G_{Na} are constant with values dependent on n, m, and h. Since the membrane is at steady state, $\frac{dn}{dt} = 0$, $\frac{dm}{dt} = 0$, and $\frac{dh}{dt} = 0$. Using Eqs. (12.43), (12.45), and (12.46), at resting potential and steady state

$$n = \frac{\alpha_n^0}{\alpha_n^0 + \beta_n^0}$$

$$m = \frac{\alpha_m^0}{\alpha_m^0 + \beta_m^0}$$

$$h = \frac{\alpha_h^0}{\alpha_h^0 + \beta_h^0}$$

where α_i^0 is α at $V = 0$ for $i = n, m,$ and $h,$ and β_i^0 is β at $V = 0$ for $i = n, m,$ and h. Calculations yield $\alpha_n^0 = 0.0582$, $\beta_n^0 = 0.125$, $n = 0.31769$, $\alpha_m^0 = 0.2236$, $\beta_m^0 = 4$, $m = 0.05294$, $\alpha_h^0 = 0.07$, $\beta_h^0 = 0.04742$, and $h = 0.59615$. Therefore, at resting potential and steady state

$$G_K = \bar{G}_K n^4 = 36.0 \times 10^{-3} (0.31769)^4 = 0.3667 \times 10^{-3} \text{ S}$$

and

$$G_{Na} = \bar{G}_{Na} m^3 h = 120.0 \times 10^{-3} (0.05294)^3 \times 0.59615 = 0.010614 \times 10^{-3} \text{ S}$$

12.6.3 Equation for the Time Dependence of the Membrane Potential

Figure 12.28 shows a model of the cell membrane that is stimulated via an external stimulus, I_m, which is appropriate for simulating action potentials. Applying Kirchhoff's current law at the cytoplasm yields

$$I_m = G_K(V_m - E_K) + G_{Na}(V_m - E_{Na}) + \frac{(V_m - E_l)}{R_l} + C_m \frac{dV_m}{dt} \tag{12.47}$$

where G_K and G_{Na} are the voltage-time dependent conductances given by Eqs. (12.42) and (12.44).

EXAMPLE PROBLEM 12.9

For the squid giant axon, compute the size of the current pulse (magnitude and pulse width) necessary to raise the membrane potential from its resting value of −60 mV to −40 mV and then back to its resting potential. Neglect any changes in K^+ and Na^+ conductances from resting potential, but include G_K and G_{Na} at resting potential in the analysis. Hodgkin-Huxley parameter values

Continued

for the squid giant axon are $G_l = \dfrac{1}{R_l} = 0.3 \times 10^{-3}$ S, $\bar{G}_K = 36 \times 10^{-3}$ S, $\bar{G}_{Na} = 120 \times 10^{-3}$ S, $E_K = -72 \times 10^{-3}$V, $E_l = -49.4 \times 10^{-3}$V, $E_{Na} = 55 \times 10^{-3}$V, and $C_m = 1 \times 10^{-6}$F.

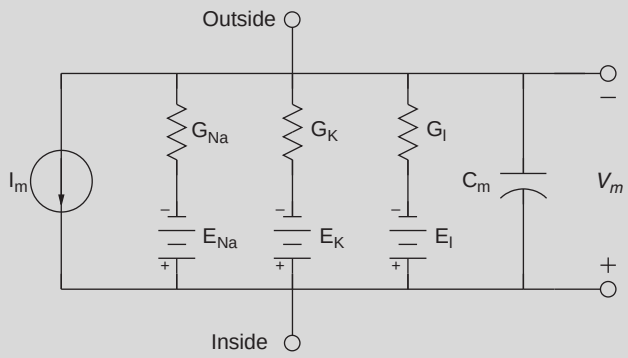
Solution

Let current I_m be given by $I_m = K(u(t) - u(t - t_0))$. In Example Problem 12.8, the conductances at resting potential were calculated as $G_K = 0.3667 \times 10^{-3}$ S and $G_{Na} = 0.010614 \times 10^{-3}$ S. Since G_K and G_{Na} remain constant for a subthreshold current stimulus in this problem, the circuit in Figure 12.28 reduces to the following circuit. For ease in analysis, this circuit is replaced by the following Thevenin's equivalent circuit with

$$R_{Th} = \frac{1}{G_{Na} + G_K + G_l} = 1.4764\, k\Omega$$

and

$$V_{TH} = -60 \text{ mV}.$$



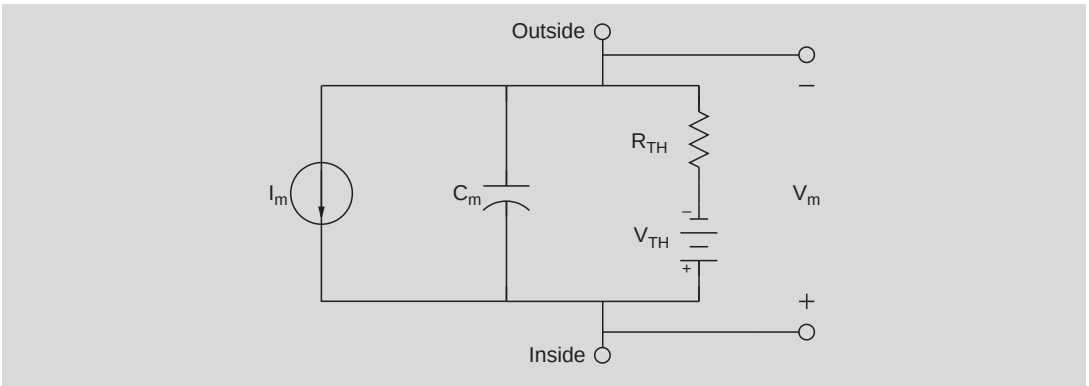
Since the solution in Example Problem 12.6 is the same as the solution in this problem, we have

$$V_m(t) = V_{TH} + R_{TH}K\left(1 - e^{-\frac{t}{R_{TH}C_m}}\right)u(t) - R_{TH}K\left(1 - e^{-\frac{t-t_o}{R_{TH}C_m}}\right)u(t - t_0)$$

For convenience, assume the current pulse $t_0 > 5\tau$. Therefore, for $5\tau < t \leq t_o$, $V_m = -40$ mV according to the problem statement at steady state, and from the preceding equal, V_m reduces to

$$V_m = -0.040 = V_{TH} + K \times R_{TH} = -0.06 + K \times 1476.4$$

which yields $K = 13.6\ \mu$A. Since $\tau = R_{TH}C_m = 1.47$ ms, any value for t_0 greater than $5\tau = 7.35$ ms brings V_m to -40 mV with $K = 13.6\ \mu$A. Naturally, a larger current pulse magnitude is needed for an action potential because as V_m exponentially approaches threshold (reaching it with a duration of infinity), the Na^+ conductance channels become inactive and shut down.



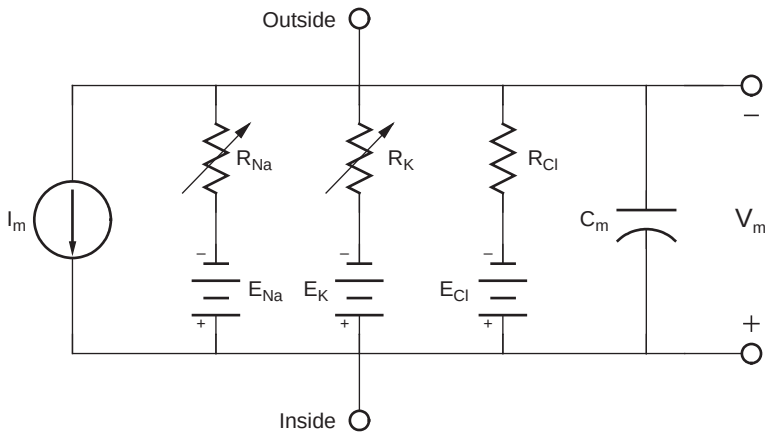


FIGURE 12.28 Circuit model of an unmyelinated section of squid giant axon. The channels for K^+ and Na^+ are represented using the variable voltage-time conductances given in Eqs. (12.42) and (12.44). The passive gates for Na^+, K^+, and Cl^- are given by a leakage channel with resistance, R_l, and Nernst potential, E_l. The *Na-K* pump is not drawn for ease in analysis, since it does not contribute any current to the rest of the circuit.

To find V_m during an action potential, four differential equations (Eqs. (12.43), (12.45), (12.46), and (12.47)) and six algebraic equations ($\alpha_i's$ and $\beta_i's$ in Eqs. (12.43), (12.45), and (12.46)) need to be solved. Since the system of equations is nonlinear due to the n^4 and m^3 conductance terms, an analytic solution is not possible. To solve for V_m, it is therefore necessary to simulate the solution. There are many computer tools that allow a simulation solution of nonlinear systems. SIMULINK, a general purpose toolbox in MATLAB that simulates solutions for linear and nonlinear, continuous, and discrete dynamic systems, is used in this textbook. SIMULINK is a popular and widely used simulation program with a user-friendly interface that is fully integrated within MATLAB. SIMULINK is interactive and works on most computer platforms. Analogous to an analog computer, programs for SIMULINK are developed based on a block diagram of the system.

The SIMULINK program for an action potential is shown in Figures 12.29 to 12.32. The block diagram is created by solving for the highest derivative term in Eq. (12.47), which yields Eq. (12.48). The SIMULINK program is then created by using integrators, summers, and so on.

$$\frac{dV_m}{dt} = \frac{1}{C_m}(I_m + G_K(E_K - V_m) + G_{Na}(E_{Na} - V_m) + G_l(E_l - V_m)) \tag{12.48}$$

Figure 12.25 shows the main block diagram. Figures 12.29 to 12.31 are subsystems that were created for ease in analysis. The Workspace output blocks were used to pass simulation results to MATLAB for plotting. Parameter values used in the simulation were based on the empirical results from Hodgkin and Huxley, with $G_l = \dfrac{1}{R_l} = 0.3 \times 10^{-3}$ S, $\bar{G}_K = 36 \times 10^{-3}$ S, $\bar{G}_{Na} = 120 \times 10^{-3}$ S, $E_K = -72 \times 10^{-3}$ V, $E_l = -49.4 \times 10^{-3}$ V, $E_{Na} = 55 \times 10^{-3}$ V,

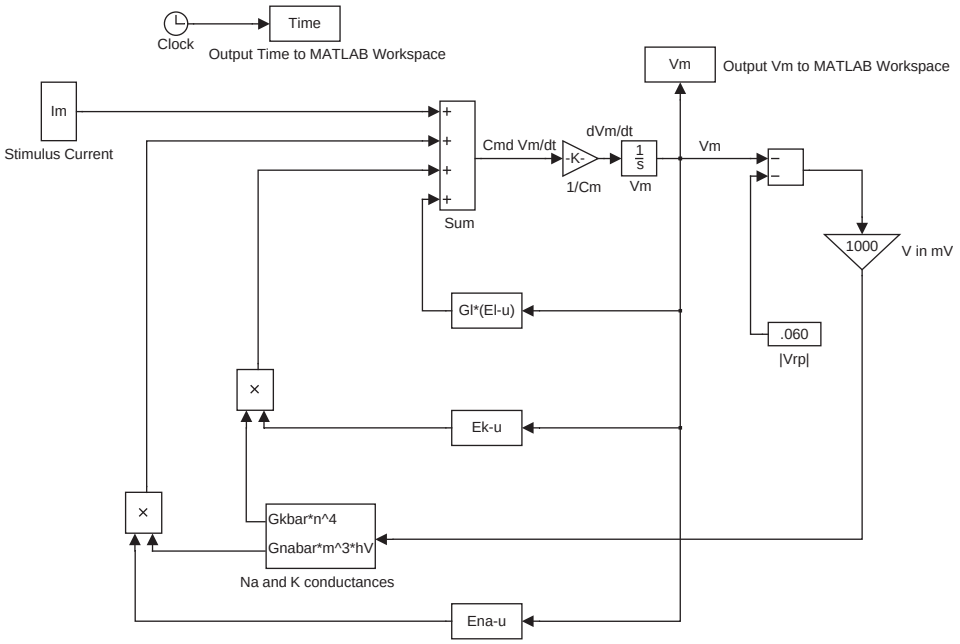


FIGURE 12.29 Main block diagram for simulating an action potential using SIMULINK. The stimulus current is a pulse created by subtracting two step functions as described in Figure 12.30. The Na^+ and K^+ conductance function blocks are described in Figures 12.31 and 12.32.

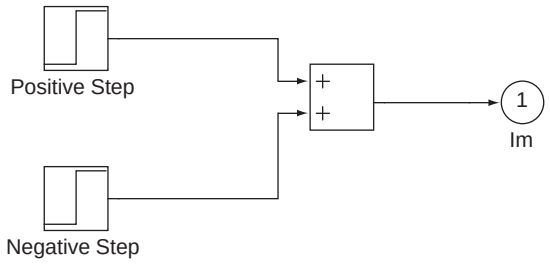


FIGURE 12.30 The stimulus current.

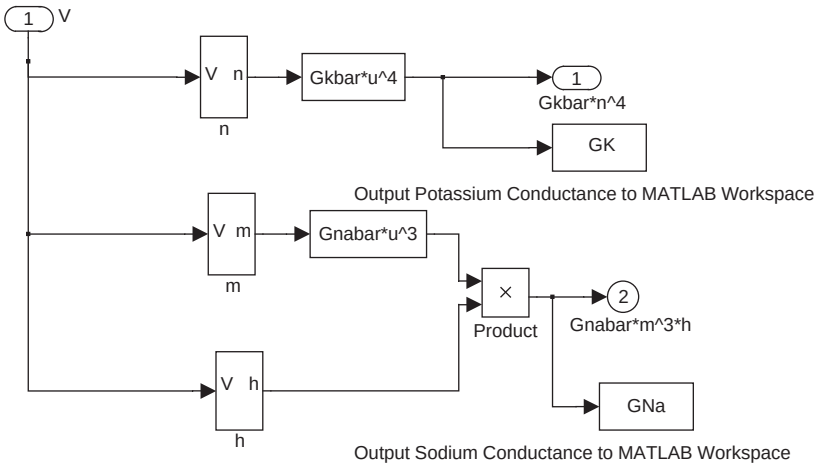


FIGURE 12.31 SIMULINK program for the K^+ and Na^+ conductance channels.

and $C_m = 1 \times 10^{-6}$ F. Figure 12.23 is a SIMULINK simulation of an action potential. The blocks Gl*(El-u), Ek-u, and Ena-u are function blocks that were used to represent the terms $G_l\,(E_l - V_m)$, $(E_{\mathrm{K}} - V_m)$, and $(E_{\mathrm{Na}} - V_m)$ in Eq. (12.48), respectively.

The stimulus pulse current was created by using the SIMULINK step function as shown in Figure 12.30. The first step function starts at $t = 0$ with magnitude K, and the other one starts at $t = t_0$ with magnitude $-K$. The current pulse should be sufficient to quickly bring V_{m} above threshold.

Figure 12.31 illustrates the SIMULINK program for the conductance channels for Na^+ and K^+. Function blocks Gkbar*u^4, Gnabar*u^3, and Gnabar*m^3*h represent $\bar{G}_K\ n^4$, $\bar{G}_{Na}\ m^3$, and $\bar{G}_{Na}\ m^3h$, respectively. The subsystems n, m, and h are described in Figure 12.32 and are based on six algebraic equations for α_i's and β_i's in Eqs. (12.43), (12.45), and (12.46).

12.7 MODEL OF A WHOLE NEURON

This section brings together the entire neuron, combing the dendrite, soma, axon, and presynaptic terminal. Dendrites and axons can be modeled as a series of cylindrical compartments, each connected together with an axial resistance, as described in Section 12.5.3. Both the axon and dendrites are connected to the soma. Of course, real neurons have many different arrangements, such as the dendrite connected to the axon, which then connects to the soma. The basic neuron consists of many dendrites, one axon, and one soma. Note that the dendrite and axon do not have to have constant diameter cylinders but may narrow toward the periphery.

As described previously, Figure 12.17 illustrates a generic electrical dendrite compartment model with passive channels, and Figure 12.28 illustrates the axon compartment with active channels at the axon hillock and the node of Ranvier. To model the myelinated portion of the axon, a set of passive compartments, like the dendrite compartment, can be used with capacitance, passive ion channels, and axial resistance. Shown in Figure 12.33 is a portion of the axon with myelin sheath, with three passive channels, and an active component

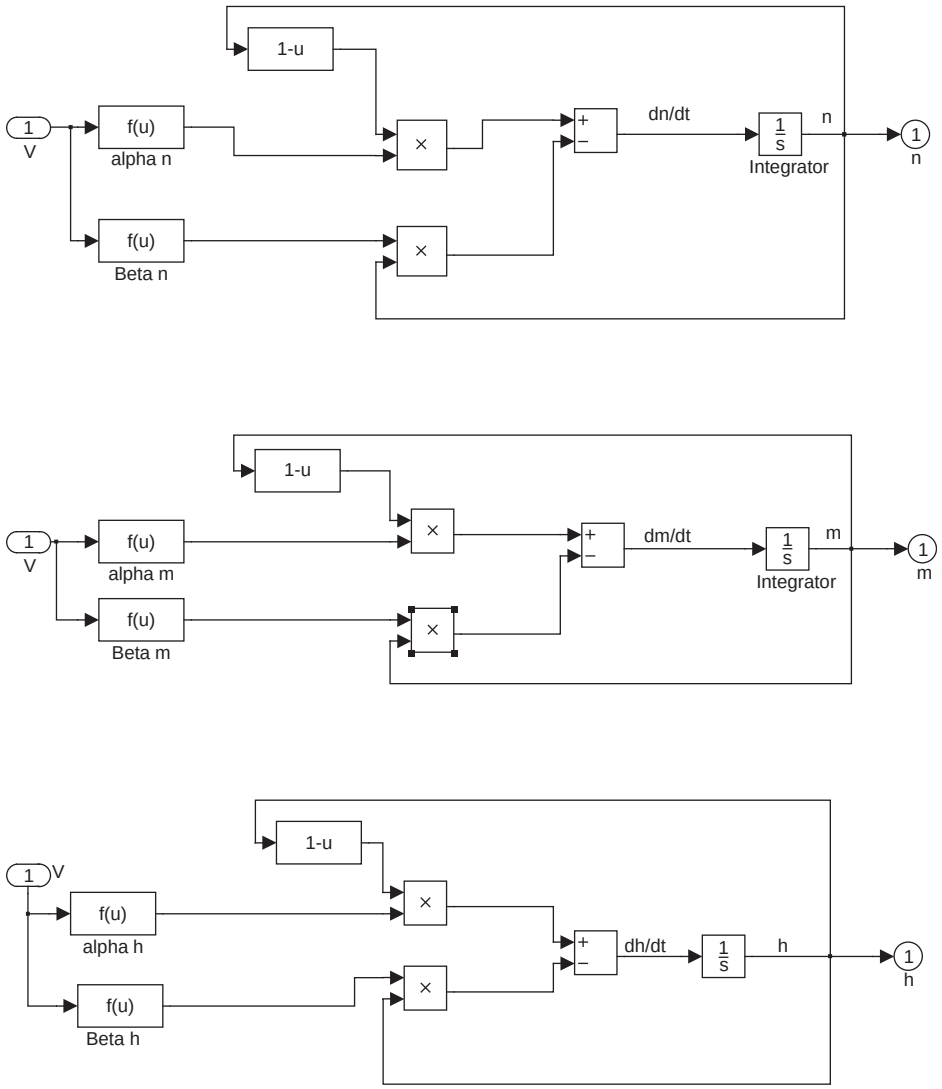


FIGURE 12.32 SIMULINK program for the alpha and beta terms in Equations (12.43), (12.45) and (12.46).

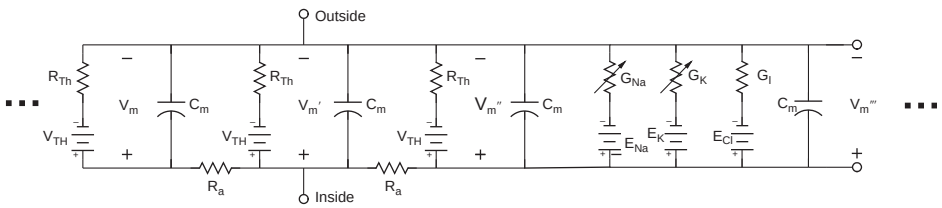


FIGURE 12.33 A segment of the axon with active and passive compartments.

for the node of Ranvier. The structure in Figure 12.33 can be modified for any number of compartments as appropriate. The soma can be modeled as an active or passive compartment depending on the type of neuron.

To model the neuron in Figure 12.33, Kirchhoff's current law is applied for each compartment (i.e., each line in Eq. (12.49) is for a compartment), giving

$$\begin{aligned}
&\ldots + C_m \frac{dV_m}{dt} + \frac{(V_m - V_{TH})}{R_{TH}} + \frac{(V_m - V'_m)}{R_a} \\
&+C_m \frac{dV'_m}{dt} + \frac{(V'_m - V_{TH})}{R_{TH}} + \frac{(V'_m - V''_m)}{R_a} \\
&+C_m \frac{dV''_m}{dt} + \frac{(V''_m - V_{TH})}{R_{TH}} + \frac{(V''_m - V'''_m)}{R_a} \\
&+G_K\left(V'''_m - E_K\right) + G_{Na}\left(V'''_m - E_{Na}\right) + \frac{\left(V'''_m - E_l\right)}{R_l} + C_m \frac{dV'''_m}{dt} + \ldots
\end{aligned} \tag{12.49}$$

Because neurons usually have other channels in addition to the three of the squid giant axon, a model of the neuron should have the capability of including other channels, such as a fast sodium channel, delayed potassium conductance, high threshold calcium conductance, and so forth. Additional ion channels can be added for each compartment in Eq. (12.49) by adding

$$\sum_{i=1}^{n} G_i(V_m - E_i)$$

for each compartment for channels $i = 1, n$. The values of $C_m, R_{TH}, R_a,$ and G_i are dependent on the size of the compartment and the type of neuron modeled.

A complete model of the neuron can be constructed by including as many dendritic branches as needed, each described using Figure 12.17 and each modeled by

$$\ldots + C_m \frac{dV_m}{dt} + \frac{(V_m - V_{TH})}{R_{TH}} + \frac{(V_m - V'_m)}{R_a} + C_m \frac{dV'_m}{dt} + \frac{(V'_m - V_{TH})}{R_{TH}} + \frac{(V'_m - V''_m)}{R_a} + \ldots \tag{12.50}$$

a soma with passive or active properties using either

$$C_m \frac{dV_m}{dt} + \frac{(V_m - V_{TH})}{R_{TH}} + \frac{(V_m - V'_m)}{R_a} \tag{12.51}$$

or

$$G_K\left(V'''_m - E_K\right) + G_{Na}\left(V'''_m - E_{Na}\right) + \frac{\left(V'''_m - E_l\right)}{R_l} + C_m \frac{dV'''_m}{dt} \tag{12.52}$$

and an axon using Eq. (12.49) as described in Rodriguez and Enderle [3]. Except for the terminal compartment, two inputs are needed for the dendrite compartment: the input defined by the previous compartment's membrane potential and the next compartment's

membrane potential. Additional neurons can be added using the same basic neuron, interacting with each other using the current from the adjacent neuron (presynaptic terminal) to stimulate the next neuron.

For illustration purposes, the interaction between two adjacent neurons is modeled using SIMULINK, shown in Figure 12.34, and the results shown in Figure 12.35. Three voltage-dependent channels for Na^{+}, K^{+}, and Ca^{+2}, and also a leakage channel are used for the axon. We use a myelinated axon with four passive compartments between each node of Ranvier. The total axon consists of three active compartments and two myelinated passive segments. The dendrite consists of five passive compartments, and the soma is a passive spherical compartment. The stimulus is applied at the terminal end of the dendrite of the first neuron. It is modeled as an active electrode compartment. The size of each axon compartment is the same but different than the dendrite compartment. The input to the first neuron is shown in Figure 12.36.

12.8 CHEMICAL SYNAPSES

The previous section describes the movement of a signal through a change in membrane potential from the dendrite, soma, and axon to the presynaptic terminal. In this section, we examine the process that occurs at the presynaptic terminal, called the presynaptic neuron, and the interaction with the postsynaptic terminal on an adjacent neuron's dendrite, called the postsynaptic neuron. The major action in the communication between two adjacent neurons is the movement of a neurotransmitter from the presynaptic terminal to the postsynaptic terminal via diffusion, transferring electrical energy to chemical energy, and then back into electrical energy.

Another type of synapse is the gap junction that connects two neurons. Under this arrangement, ions move directly from one cell to the other. Since this form of communication is rather rare among neurons, we will not cover it any further here. Gap junctions are typically observed in signaling among smooth muscle fibers.

More than 40 neurotransmitters have been discovered, some that are excitatory and others that are inhibitionary. An example of an excitatory neurotransmitter is acetylcholine (ACh), which depolarizes the postsynaptic neuron's dendrite by opening sodium channels. An example of an inhibitionary neurotransmitter is gamma-aminobutyric acid (GABA), which opens the chloride channels that then hyperpolarizes the postsynaptic neuron's dendrite.

Figure 12.3 shows a set of converging presynaptic terminals on the postsynaptic neuron's dendritic membrane. The space between the two is called the synaptic gap. The distance between the two neurons is quite small and ranges from 200 to 500 angstroms. Important features in this form of communication is that it is one-way and analog, which allow for the summing of all asynchronous inputs. This one-way communication allows a precise communication and control among neurons with signals going in one direction only. Keep in mind that the gap junction is not a one-way communication channel, as it allows ions to move in either direction. It should also be noted that a small percentage of presynaptic terminals also converge on the soma, which eliminates the change in membrane potential as a function of distance that occurs on the dendrite.

Figure 12.37 depicts a single presynaptic terminal (top) and the postsynaptic terminal (bottom). The neurotransmitter is stored in numerous vesicles in the presynaptic neuron,

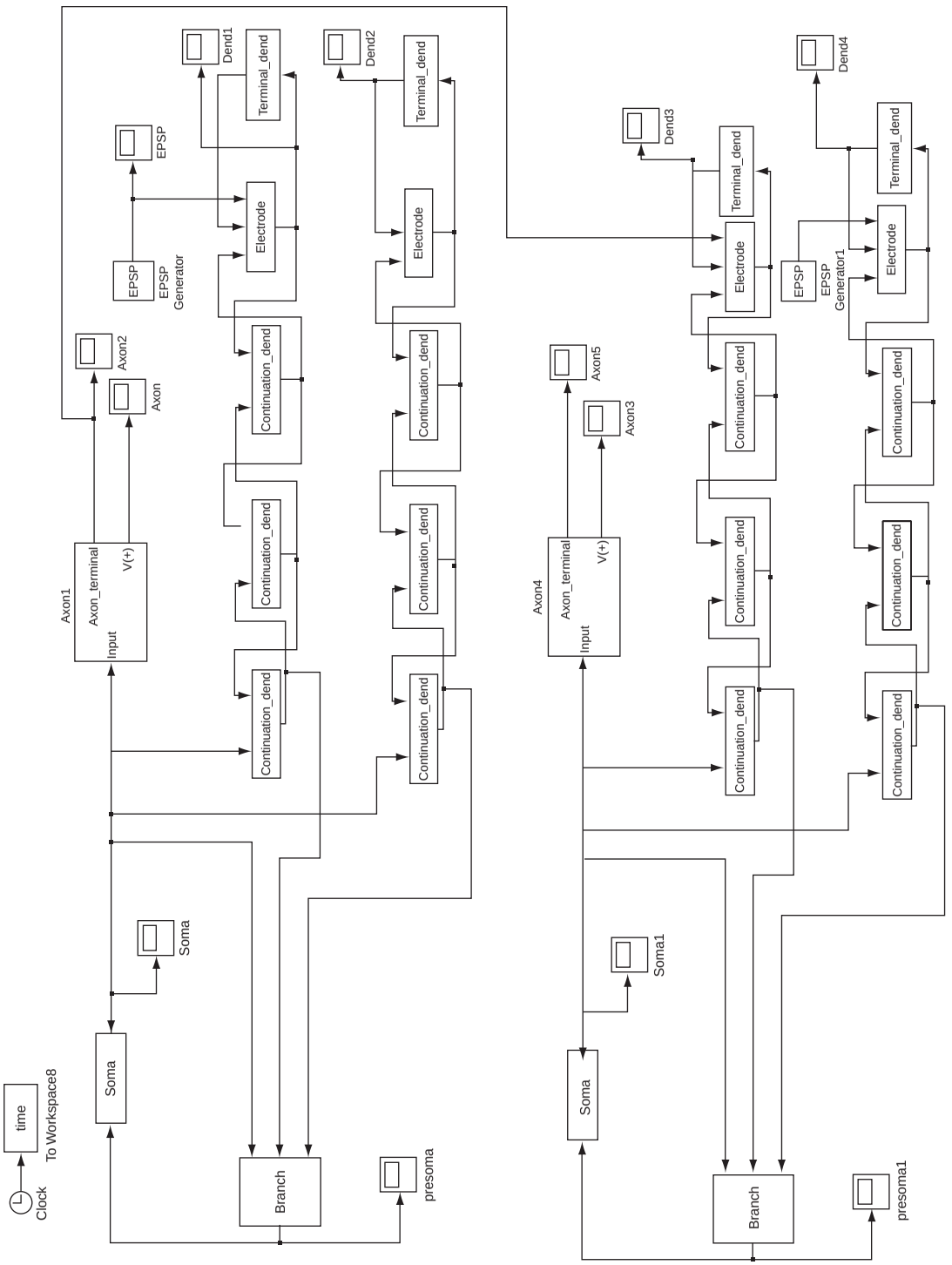


FIGURE 12.34 SIMULINK model for two adjacent neurons.

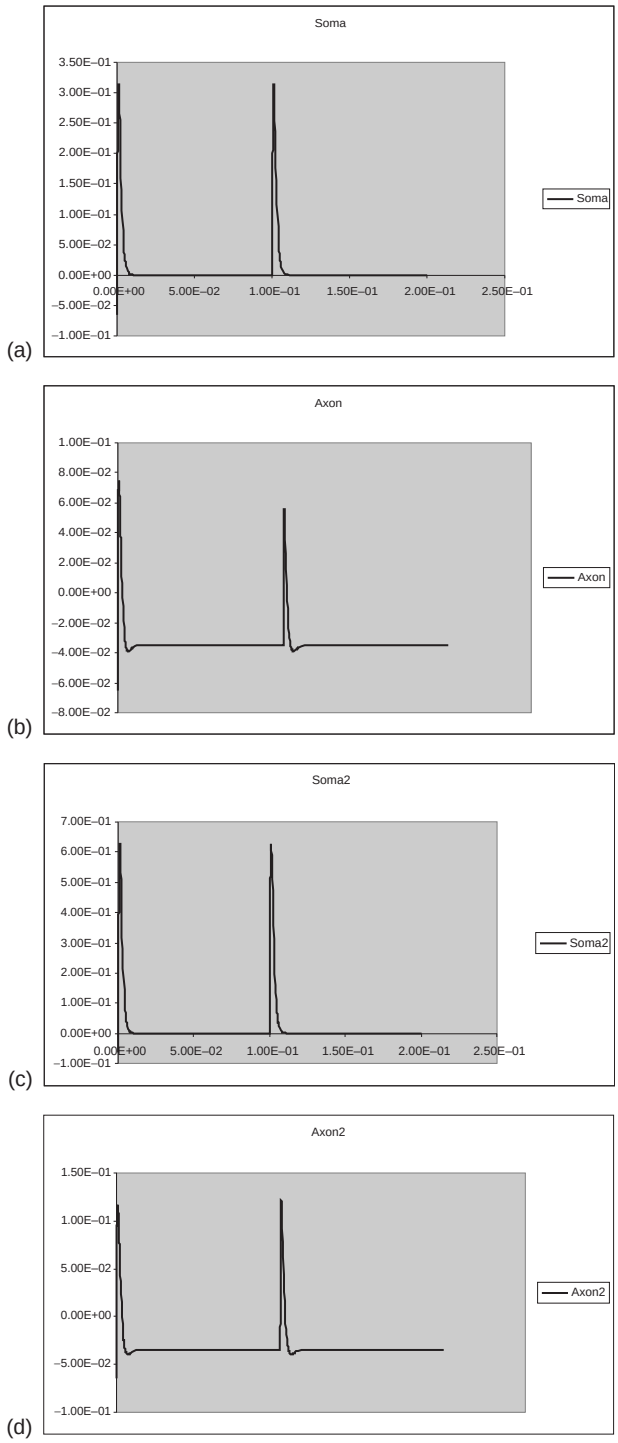


FIGURE 12.35 (a) Soma of first neuron, (b) axon of first neuron, (c) soma of second neuron, and (d) axon of second neuron.

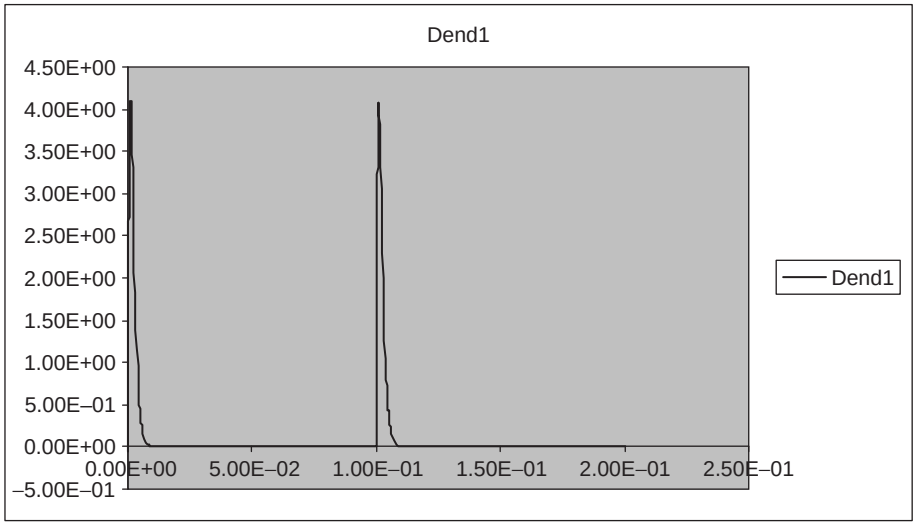


FIGURE 12.36 The stimulus to the first neuron.

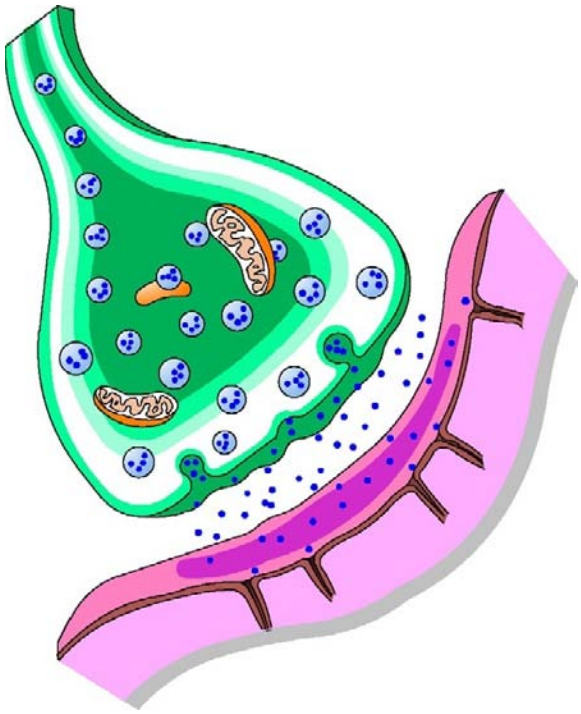


FIGURE 12.37 Presynaptic (top) and postsynaptic (bottom) terminals illustrating a neurotransmitter moving across the synaptic gap.

enough for responding to thousands of action potentials, where each vesicle stores approximately 10,000 molecules of neurotransmitter in vesicles. The mitochondria supplies ATP to resupply the vesicles with the neurotransmitter. The resupplying of the neurotransmitter into the vesicles occurs even while the presynaptic neuron is delivering the neurotransmitter into the synaptic gap during an action potential.

After an action potential reaches the presynaptic terminal, it depolarizes the membrane and causes the movement of a few vesicles toward the membrane. Once in the vicinity of the membrane, the vesicles then fuse with the membrane, as shown in Figure 12.38. This process usually occurs in less than 1 ms. The fused vesicle then opens and the neurotransmitter is released into the synaptic gap. This movement is not highly synchronized but occurs in a random fashion. After moving through the synaptic gap by diffusion, the neurotransmitter binds to a receptor in the postsynaptic neuron. The binding of the neurotransmitter to the postsynaptic receptor causes a change in membrane potential, either depolarization by an excitatory neurotransmitter, or hyperpolarization by an inhibitionary neurotransmitter. This process usually occurs in less than 1 ms.

Note that there are a large number of presynaptic terminals (10,000 to 200,000) that converge on the dendrite of a postsynaptic neuron. Each molecule of neurotransmitter causes a change in membrane potential, which all sum according to superposition as described in Example Problem 12.6. Note that synchronization is not essential because of the capacitive nature of the membrane. The change in membrane potential at the postsynaptic dendrite then travels to the soma as described in Section 12.5. If at the axon hillock the membrane potential crosses threshold, an action potential is created that then moves without attenuation down the axon, where the process repeats itself. If the membrane potential does not reach threshold at the axon hillock, it then returns to the steady-state resting potential.

12.8.1 Calcium Ions

While the exact mechanism for the movement of the neurotransmitter vesicles to the presynaptic terminal membrane and its release into the synaptic gap is unknown, it is believed to involve voltage-gated calcium ion channels. The concentration of Ca^{+2} is kept quite low in the cytosol during periods of rest. When an action potential arrives at the presynaptic

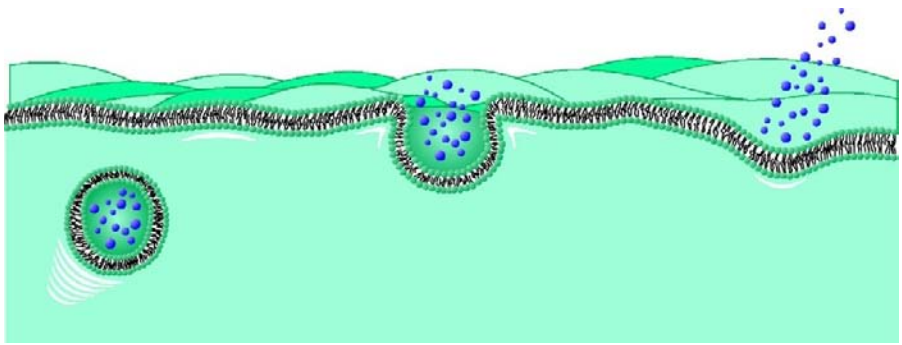


FIGURE 12.38 The movement of the vesicle through the cytosol to the cell membrane, where it fuses and then releases the neurotransmitter in the synaptic gap.

terminal, the concentration of Ca^{+2} in the cytosol quickly rises. The amount of Ca^{+2} in the cytosol appears to be proportional to the number of action potentials, and in turn, the amount of neurotransmitter released.

Keener and Sneyd [2] provide an excellent summary in Chapter 7 of the various theories involving the role of Ca^{+2} in the release of neurotransmitter into the synaptic gap. Two current models involve different roles for the change in membrane potential, one indicating that it has a direct effect on the amount of neurotransmitter released, and the other acting directly on the Ca^{+2} entering the cytosol.

For illustrative purposes, consider a model for the movement of the neurotransmitter vesicles that involves the transport of Ca^{+2} across the cell membrane using a carrier-mediated transport process as described in Section 8.4.3.[2] When an action potential depolarizes the presynaptic terminal, a voltage-gated channel in the membrane is activated that allows up to four Ca^{+2} to bind to a binding protein, R, on the outside of the cell membrane. When totally bound, the protein joined by four Ca^{+2}, P, moves to the interior of the cell membrane, where it binds with a vesicle, V. A carrier, C, is then joined with PV, forming the bound carrier W. This reaction releases the four Ca^{+2} and the binding protein R. The bound carrier W then moves to the outside of the cell membrane through the opened voltage-gated channel, where vesicle releases n packets of the neurotransmitter N. The overall reaction is given by

$$
\begin{aligned}
&4Ca_o^{+2} + R_o \xrightarrow{B_1} P_o \underset{B_{-2}}{\overset{B_2}{\rightleftharpoons}} P_i \\
&P_i + V_i \xrightarrow{B_3} P_iV_i \\
&C_i + P_iV_i \xrightarrow{B_4} W_i + 4Ca_o^{+2} + R_i \\
&W_i \underset{B_{-5}}{\overset{B_5}{\rightleftharpoons}} W_o \xrightarrow{B_6} V_o + C_o \\
&V_o \xrightarrow{B_7} nN \\
&R_i \xrightarrow{B_8} R_o \\
&C_o \xrightarrow{B_9} C_i
\end{aligned}
\tag{12.53}
$$

where the subscript i and o are the inside and outside. Figure 12.39 shows a model of the four stages of the binding protein with Ca^{+2}, which are described by the following equations.

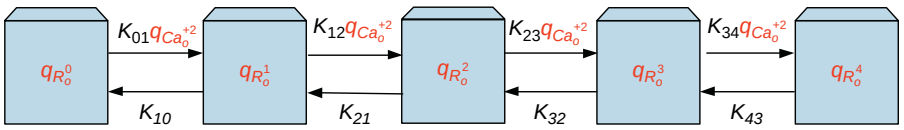


FIGURE 12.39 Carrier-mediated transport of calcium ions into the presynaptic terminal that initiates the movement of the neurotransmitter out of the cell.

[2]A more complex model is given in Section 7.1.3 of Keener and Sneyd that involves probabilistic considerations.

$$
\begin{aligned}
\dot{q}_{Ca_0^{+2}} &= J_{Ca^{+2}} + K_{10}q_{R_o^1} + K_{21}q_{R_o^2} + K_{32}q_{R_o^3} + K_{43}q_{R_o^4} \\
&\quad -K_{01}q_{Ca_0^{+2}}q_{R_o^0} - K_{12}q_{Ca_0^{+2}}q_{R_o^1} - K_{23}q_{Ca_0^{+2}}q_{R_o^2} - K_{34}q_{Ca_0^{+2}}q_{R_o^3} \\
\dot{q}_{R_o^0} &= B_8 q_{R_i} + K_{10}q_{R_o^1} - K_{01}q_{Ca_0^{+2}}q_{R_o^0} \\
\dot{q}_{R_o^1} &= K_{21}q_{R_o^2} + K_{01}q_{Ca_0^{+2}}q_{R_o^0} - K_{10}q_{R_o^1} - K_{12}q_{Ca_0^{+2}}q_{R_o^1} \\
\dot{q}_{R_o^2} &= K_{32}q_{R_o^3} + K_{12}q_{Ca_o^{+2}}q_{R_o^1} - K_{21}q_{R_o^2} - K_{23}q_{Ca_o^{+2}}q_{R_o^2} \\
\dot{q}_{R_o^3} &= K_{43}q_{R_o^4} + K_{23}q_{Ca_o^{+2}}q_{R_o^2} - K_{32}q_{R_o^3} - K_{34}q_{Ca_o^{+2}}q_{R_o^3} \\
\dot{q}_{R_o^4} &= K_{34}q_{Ca_o^{+2}}q_{R_o^3} - K_{43}q_{R_o^4} - B_1 q_{R_o^4}
\end{aligned}
\tag{12.54}
$$

where R_o^i has i Ca_o^{+2} bound to it, the transfer rates, $K_{i,j}$, are functions of membrane voltage, and $J_{Ca^{+2}}$ is the flow of Ca^{+2} out of the cell. Here, we assume that the quantity of Ca^{+2} is constant and given by $\gamma = q_{Ca_i^{+2}} + q_{Ca_o^{+2}} - q_{R_o^0} - q_{R_o^1} - q_{R_o^2} - q_{R_o^3} - q_{R_o^4}$. The flow $J_{Ca^{+2}}$ is due to an active pump that keeps the concentration of Ca^{+2} in the cytosol low, while the concentration outside the cell is very high. One can imagine a similar process for the $J_{Ca^{+2}}$ active pump as given for the *Na-K* pump described in Section 8.4.4. The transfer rates are zero at resting potential and increase as the membrane depolarizes (Keener and Sneyd use exponential functions for the transfer rates). Next, we have

$$
\begin{aligned}
\dot{q}_{P_o} &= B_1 q_{R_o^4} + B_{-2}q_{P_i} - B_2 q_{P_o} \\
\dot{q}_{P_i} &= B_2 q_{P_o} - B_{-2}q_{P_i} - B_3 q_{P_i} q_{V_i} \\
\dot{q}_{V_i} &= J_{V_i} - B_3 q_{P_i} q_{V_i} \\
\dot{q}_{V_o} &= B_6 q_{W_o} - B_7 q_{V_o} \\
\dot{q}_{W_i} &= B_4 q_{C_i} q_{P_i V_i} + B_{-5}q_{W_o} - B_5 q_{W_i} \\
\dot{q}_{W_o} &= B_5 q_{W_i} - B_{-5}q_{W_o} - B_6 q_{W_o} \\
\dot{q}_{C_i} &= B_9 q_{C_o} - B_4 q_{C_i} q_{P_i V_i} \\
\dot{q}_{C_o} &= B_6 q_{W_o} - B_9 q_{C_o} \\
\dot{q}_{N_o} &= nB_7 q_{V_o}
\end{aligned}
\tag{12.55}
$$

where J_{V_i} is the rate of the creation of new vesicles.

It should be noted that the vesicles absorbed into the cell membrane during the movement of the neurotransmitter out of the cell, after a period of time, leave the cell membrane and reform in the cytosol with the same characteristics as before. The vesicle then synthesizes the neurotransmitter using enzyme proteins in its membrane using ATP from the mitochondria. For example, ACh is resupplied inside the vesicle using a carrier-mediated enzyme choline acetyltransferase that binds acetyl coenzyme and choline similar to the reaction described in Section 8.4.3. The entire process from receipt of the action potential at the presynaptic terminal to release of the neurotransmitter takes less than 1 ms.

12.8.2 Postsynaptic Neurons

Once the neurotransmitter is released from the presynaptic neuron, it diffuses across the synaptic gap and binds to the receptors on the postsynaptic neuron's dendrite. Once bound to the dendrite, the neurotransmitter moves into the cytosol by carrier-mediated transport, where once in the cytosol, an ion channel opens. If the neurotransmitter is excitatory, Na^+ flows into the open channel, and the membrane depolarizes. If the neurotransmitter is inhibitory, Cl^- flows into the open channel, and the membrane hyperpolarizes. This process takes less than 1 ms. The change in membrane potential lasts up to 15 ms due to the capacitive nature of the membrane, thus allowing for temporal summation of all the neurotransmitter secreted by the many presynaptic terminals.

Consider the neurotransmitter ACh and its carrier-mediated transport into the cytosol of the postsynaptic neuron. ACh binds to the enzyme choline acetyltransferase, C, that transports it across the membrane, allowing it to pass into the cytosol. Using the model illustrated in Figure 8.21, we have

$$\begin{aligned} &ACh_0 + C_0 \xrightarrow{K_1} P_o \underset{K_2}{\overset{K_2}{\rightleftharpoons}} P_i \xrightarrow{K_3} ACh_i + C_i \\ &ACh_i \xrightarrow{K_4} A_i + Ch_i \\ &C_o \underset{K_2}{\overset{K_2}{\rightleftharpoons}} C_i \end{aligned} \tag{12.56}$$

where P is the bound substrate and carrier complex, A is acetyl coenzyme A, and Ch is choline. As before, i and o subscripts refer to the inside and outside of the cell. We assume that A and Ch are moved from the cytosol into the synaptic gap by diffusion and that the flow of ACh into the synaptic gap from the presynaptic neuron is given by N_o (now called ACh_o) from Eq. (12.55). The equations that describe this system are given by

$$\begin{aligned} \dot{q}_{ACh_o} &= -K_1 q_{ACh_o} q_{C_o} + nB_7 q_{V_o} \\ \dot{q}_{C_o} &= -K_1 q_{ACh_o} q_{C_o} + K_2 q_{C_i} - K_2 q_{C_o} \\ \dot{q}_{P_o} &= K_1 q_{ACh_o} q_{C_o} + K_2 q_{P_i} - K_2 q_{P_o} \\ \dot{q}_{P_i} &= K_2 q_{P_o} - K_1 q_{P_i} \\ \dot{q}_{ACh_i} &= K_1 q_{P_i} - K_4 q_{ACh_i} \\ \dot{q}_{C_i} &= K_1 q_{P_i} + K_2 q_{C_o} - K_2 q_{C_i} \\ \dot{q}_{A_i} &= K_4 q_{ACh_i} - D_A q_{A_i} \\ \dot{q}_{Ch_i} &= K_4 q_{ACh_i} - D_{Ch} q_{Ch_i} \end{aligned} \tag{12.57}$$

To capture the change in conductance as a function of ACh_i, and its impact on the membrane potential, we introduce another channel in the dendritic membrane section, as shown in Figure 12.40, where the stimulus current, I_s, from the neurotransmitter is

$$I_s = \frac{V_m - V_{TH}}{R(ACh_i)} \tag{12.58}$$

where $R(ACh_i)$ is the channel resistance that depends on the quantity of ACh_i.

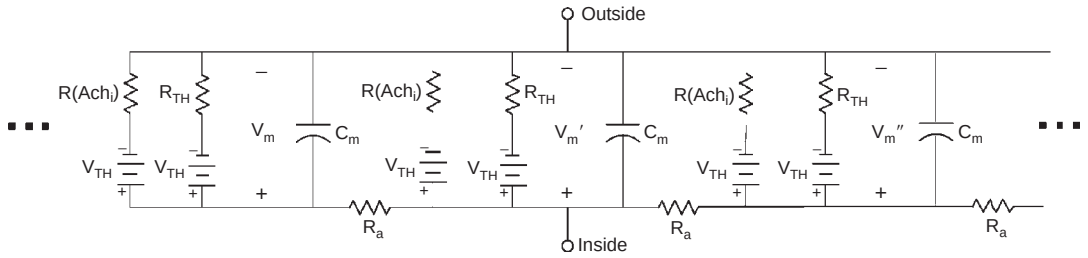


FIGURE 12.40 Equivalent circuit of series of dendritic membrane sections that includes a neurotransmitter channel. Each section is connected together with an axial resistance, R_a.

A complete model of the neuron can be constructed by including as many dendritic branches as needed, each described using Figure 12.17 and modeled by

$$\begin{aligned}&\ldots + C_m\frac{dV_m}{dt} + \frac{(V_m - V_{TH})}{R(ACh_i)} + \frac{(V_m - V_{TH})}{R_{TH}} + \frac{(V_m - V'_m)}{R_a}\\ &+C_m\frac{dV'_m}{dt} + \frac{(V'_m - V_{TH})}{R(ACh_i)} + \frac{(V'_m - V_{TH})}{R_{TH}} + \frac{(V'_m - V''_m)}{R_a} + \ldots\end{aligned} \tag{12.59}$$

Although this chapter has focused on the neuron, it is important to note that numerous other cells have action potentials that involve signaling or triggering. Many of the principles discussed in this chapter apply to these other cells as well, but the action potential defining equations are different. For example, the cardiac action potential can be defined with a DiFranceso-Noble, Luo-Rudy, or other models instead of a Hodgkin-Huxley model of the neuron.

12.9 EXERCISES

1. Assume a membrane is permeable to only Ca^{+2}. (a) Derive the expression for the flow of Ca^{+2}. (b) Find the Nernst potential for Ca^{+2}.
2. Assume that a membrane is permeable to Ca^{+2} and Cl^{-} but not to a large cation R^{+}. The inside concentrations are $[RCl] = 100$ mM and $[CaCl_2] = 200$ mM, and the outside concentration is $[CaCl_2] = 300$ mM. (a) Derive the Donnan equilibrium. (b) Find the steady-state concentration for Ca^{+2}.
3. Assume that a membrane is permeable to Ca^{+2} and Cl^{-}. The initial concentrations on the inside are different from the outside, and these are the only ions in the solution. (a) Write an equation for J_{Ca} and J_{Cl}. (b) Write an expression for the relationship between J_{Ca} and J_{Cl}. (c) Find the steady-state voltage. (d) Find the relationship between the voltage across the membrane and $[CaCl_2]$ before steady-state.
4. Assume that a membrane is permeable to only ion R^{+3}. The inside concentration is $[RCl_3] = 2$ mM, and the outside concentration is $[RCl_3] = 1.4$ mM. (a) Write an expression for the flow of R^{+3}. (b) Derive the Nernst potential for R^{+3} at steady-state.

5. Derive the Goldman equation for a membrane in which Na^+, K^+, and Cl^- are the only permeable ions.
6. Calculate V_m for the frog skeletal muscle at room temperature.
7. The following steady-state concentrations and permeabilities are given for a red blood cell membrane.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	140	4	1.0
Na^+	11	145	0.54
Cl^-	80	116	0.21

(a) Find the Nernst potential for each ion. (b) What is the resting potential predicted by the Goldman equation?

8. The following steady-state concentrations and permeabilities are given for a skeletal muscle membrane.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	150	5	0.1
Na^+	12	145	0.001
Cl^-	4	116	1.0

(a) Find the Nernst potential for each ion. (b) What is the resting potential predicted by the Goldman equation?

9. The following steady-state concentrations and permeabilities are given for a membrane.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	140	2.5	1.0
Na^+	13	110	0.019
Cl^-	3	90	0.381

(a) Find the Nernst potential for K^+. (b) What is the resting potential predicted by the Goldman equation? (c) Explain whether space charge neutrality is satisfied. (d) Explain why the steady-state membrane potential does not equal zero.

10. A membrane has the following concentrations and permeabilities.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	?	4	?
Na^+	41	276	0.017
Cl^-	52	340	0.412

The resting potential of the membrane is -52 mV at room temperature. Find the K^+ cytoplasm concentration.

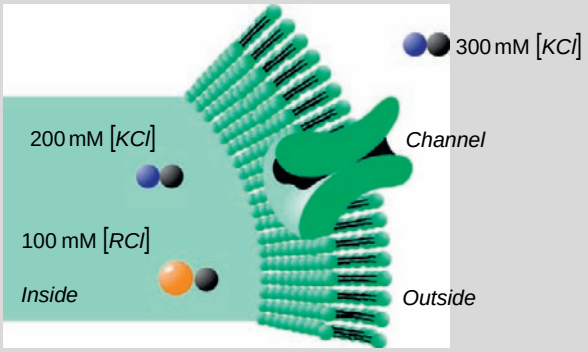
11. The following steady-state concentrations and permeabilities are given for a membrane. Note that A^+ is not permeable.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	136	15	1.0
Na^+	19	155	0.019
Cl^-	78	112	0.381
A^+	64	12	—

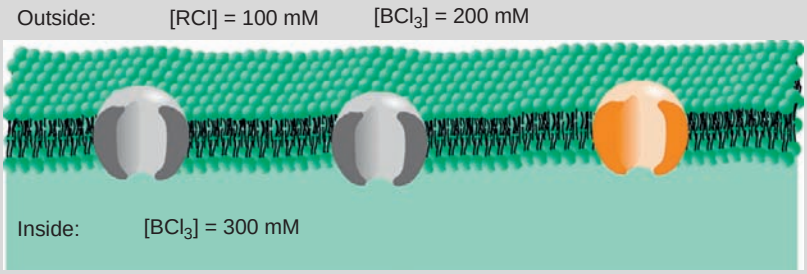
Continued

(a) Find the Nernst potential for Cl^-. (b) What is the resting potential predicted by the Goldman equation? (c) Explain whether space charge neutrality is satisfied. (d) Explain why the steady-state membrane potential does not equal zero. (e) Explain why the resting potential does not equal the Nernst potential of any of the ions.

12. A membrane is permeable to K^+ and Cl^- but not to cation R^+. Find the steady-state equilibrium concentrations for the following system's initial conditions.

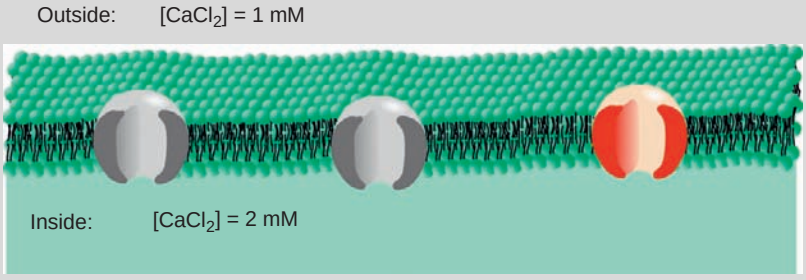


13. A membrane is permeable to B^{+3} and Cl^- but not to a large cation R^+. The following initial concentrations are given.

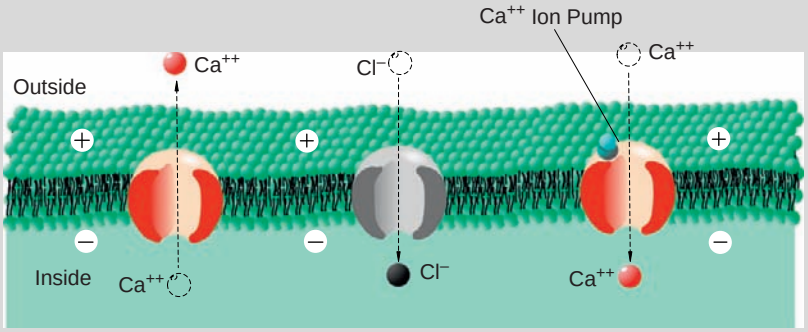


(a) Derive the Donnan equilibrium. (b) Find the steady-state concentration for B^{+3}.

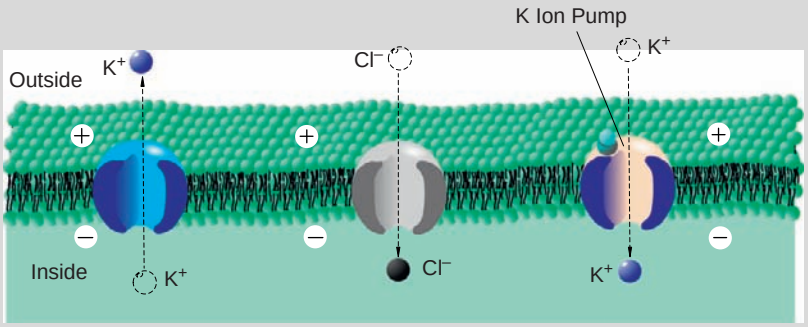
14. The following membrane is permeable to Ca^{+2} and Cl^-. (a) Write expressions for the flow of Ca^{+2} and Cl^- ions. (b) Write an expression for the relationship between J_{Ca} and J_{Cl}. (c) Find the steady-state voltage. (d) Find the relationship between voltage across the membrane and the concentration of $CaCl_2$ before steady-state is reached.



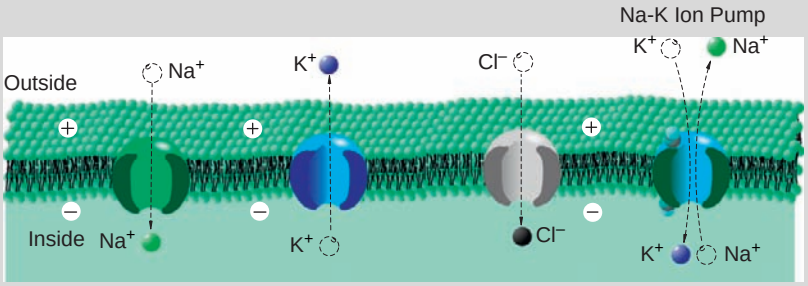
15. The following membrane has an active Ca^{+2} pump. Assume that the membrane is permeable to both Ca^{+2} and Cl^{-}, and the Ca^{+2} pump flow is J_p. The width of the membrane is δ. Find the pump flow as a function of $\left[Ca^{+2}\right]$.



16. The membrane shown is permeable to K^{+} and Cl^{-}. The active pump transports K^{+} from the outside to the inside of the cell. The width of the membrane is δ.
(a) Write an equation for the flow of each ion. (b) Find the flows at steady-state. (c) Find the pump flow as a function of $\left([K^{+}]_i - [K^{+}]_o\right)$. (d) Qualitatively describe the ion concentration on each side of the membrane.



17. The following membrane is given with two active pumps. Assume that the membrane is permeable to $Na^{+}, K^{+},$ and Cl^{-} and $J_p(K) = J_p(Na) = J_p$. The width of the membrane is δ. Solve for the quantity $\left([Cl^{-}]_i - [Cl^{-}]_o\right)$ as a function of J_p.



Continued

18. The following steady-state concentrations and permeabilities are given for a membrane. Note that A^+ is not permeable. The ion channel resistances are $R_K = 1.7\ \text{k}\Omega, R_{Na} = 9.09\ \text{k}\Omega,$ and $R_{Cl} = 3.125\ \text{k}\Omega$.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	168	6	1.0
Na^+	50	337	0.019
Cl^-	41	340	0.381
A^+	64	12	—

(a) Find the Nernst potential for each ion. (b) Draw a circuit model for this membrane. (Hint: See Figure 12.13.) (c) Find the membrane resting potential using the circuit in part (b). (d) Find the Thevenin's equivalent circuit for the circuit in part (b).

19. Suppose the membrane in Figure 12.13 is given with $R_K = 0.1\ \text{k}\Omega, R_{Na} = 2\ \text{k}\Omega, R_{Cl} = 0.25\ \text{k}\Omega,$ $E_K = -74\ \text{mV}, E_{Na} = 55\ \text{mV},$ and $E_{Cl} = -68\ \text{mV}$. (a) Find V_m. (b) Find the Thevenin's equivalent circuit.

20. Suppose a membrane has an active *Na-K* pump with $R_K = 0.1\ \text{k}\Omega, R_{Na} = 2\ \text{k}\Omega, R_{Cl} = 0.25\ \text{k}\Omega,$ $E_K = -74\ \text{mV}, E_{Na} = 55\ \text{mV},$ and $E_{Cl} = -68\ \text{mV},$ as shown in Figure 12.14. Find I_{Na} and I_K for the active pump.

21. The following steady-state concentrations and permeabilities are given for a membrane. Note that A^+ is not permeable.

Ion	Cytoplasm (mM)	Extracellular Fluid (mM)	Ratio of Permeabilities
K^+	140	2.5	1.0
Na^+	13	110	0.019
Cl^-	3	90	0.381
A^+	64	12	—

(a) If $R_K = 1.7\ \text{k}\Omega$ and $\text{R}_{\text{Cl}} = 3.125\ \text{k}\Omega,$ then find R_{Na}. (b) Find the Thevenin's equivalent circuit model.

22. Suppose that a membrane that has an active *Na-K* pump with $R_K = 0.1\ \text{k}\Omega,$ $R_{Na} = 2\ \text{k}\Omega,$ $R_{Cl} = 0.25\ \text{k}\Omega,$ $E_K = -74\ \text{mV},$ $E_{Na} = 55\ \text{mV},$ $E_{Cl} = -68\ \text{mV},$ and $C_m = 1\ \mu\text{F},$ as shown in Figure 12.15, is stimulated by a current pulse of 10 μA for 6 ms. (a) Find V_m. (b) Find the capacitive current. (c) Calculate the size of the current pulse applied at 6 ms for 1 ms, necessary to raise V_m to -40 mV. (d) If the threshold voltage is -40 mV and the stimulus is applied as in part (c), then explain whether an action potential occurs.

23. Suppose that a membrane that has an active *Na-K* pump with $R_K = 2.727\ \text{k}\Omega,$ $R_{Na} = 94.34\ \text{k}\Omega,$ $R_{Cl} = 3.33\ \text{k}\Omega,$ $E_K = -72\ \text{mV},$ $E_{Na} = 55\ \text{mV},$ $E_{Cl} = -49.5\ \text{mV},$ and $C_m = 1\mu\text{F},$ shown in Figure 12.15, is stimulated by a current pulse of 13 μA for 6 ms. Find (a) V_m, (b) I_K, and (c) the capacitive current.

24. Suppose a membrane has an active *Na-K* pump with $R_K = 1.75\ \text{k}\Omega,$ $R_{Na} = 9.09\ \text{k}\Omega,$ $R_{Cl} = 3.125\ \text{k}\Omega,$ $E_K = -85.9\ \text{mV},$ $E_{Na} = 54.6\ \text{mV},$ $E_{Cl} = -9.4\ \text{mV},$ and $C_m = 1\mu\text{F}$ as shown in Figure 12.15. (a) Find the predicted resting membrane potential. (b) Find V_m if a small subthreshold current pulse is used to stimulate the membrane.

25. Suppose a membrane has an active *Na-K* pump with $R_K = 2.727\ \text{k}\Omega, R_{Na} = 94.34\ \text{k}\Omega$, $R_{Cl} = 3.33\ \text{k}\Omega$, $E_K = -72\ \text{mV}, E_{Na} = 55\ \text{mV}, E_{Cl} = -49.5\ \text{mV}$, and $C_m = 1\ \mu\text{F}$, as shown in Figure 12.15. Design a stimulus that will drive V_m to threshold at 3 ms. Assume that the threshold potential is -40 mV. (a) Find the current pulse magnitude and duration. (b) Find and sketch V_m.

26. Suppose a current pulse of 20 μA is passed through the membrane of a squid giant axon. The Hodgkin-Huxley parameter values for the squid giant axon are $G_l = \dfrac{1}{R_l} = 0.3 \times 10^{-3}\ \text{S}$, $\bar{G}_K = 36 \times 10^{-3}\ \text{S}$, $\bar{G}_{Na} = 120 \times 10^{-3}\ \text{S}$, $E_K = -72 \times 10^{-3}\ \text{V}, E_l = -49.4 \times 10^{-3}\ \text{V}$, $E_{Na} = 55 \times 10^{-3}\ \text{V}$ and $C_m = 1 \times 10^{-6}\ \text{F}$. Simulate the action potential. Plot (a) V_m, G_{Na}, and G_K versus time, (b) V_m, n, m, and h versus time, and (c) V_m, I_{Na}, I_K, I_c, and I_l versus time.

27. Suppose a current pulse of 20 μA is passed through an axon membrane. The parameter values for the axon are $G_l = \dfrac{1}{R_l} = 0.3 \times 10^{-3}\ \text{S}$, $\bar{G}_K = 36 \times 10^{-3}\ \text{S}$, $E_K = -12 \times 10^{-3}\ \text{V}$, $E_l = 10.6 \times 10^{-3}\ \text{V}$, $E_{Na} = 115 \times 10^{-3}\ \text{V}$, and $C_m = 1 \times 10^{-6}\ \text{F}$. Assume that Eqs. (12.41) to (12.48) describe the axon. Simulate the action potential. Plot (a) V_m, G_{Na}, and G_K versus time, (b) V_m, n, m, and h versus time, and (c) V_m, I_{Na}, I_K, I_c, and I_l versus time.

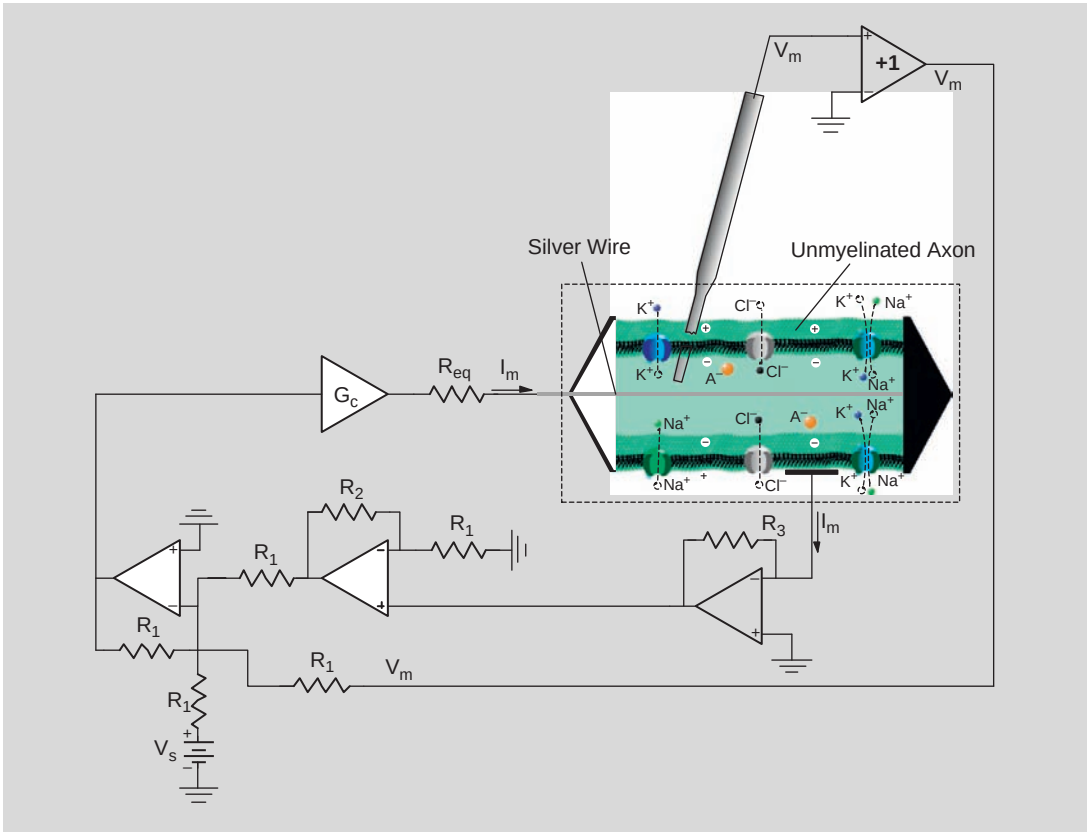
28. This exercise examines the effect of the threshold potential on an action potential for the squid giant axon. The Hodgkin-Huxley parameter values for the squid giant axon are $G_l = \dfrac{1}{R_l} = 0.3 \times 10^{-3}\ \text{S}$, $\bar{G}_K = 36 \times 10^{-3}\ \text{S}$, $\bar{G}_{Na} = 120 \times 10^{-3}\ \text{S}, E_K = -72 \times 10^{-3}\ \text{V}$, $E_l = -49.4 \times 10^{-3}\ \text{V}$, $E_{Na} = 55 \times 10^{-3}\ \text{V}$, and $C_m = 1 \times 10^{-6}\ \text{F}$. (a) Suppose a current pulse of -10 μA, which hyperpolarizes the membrane, is passed through the membrane of a squid giant axon for a very long time. At time $t = 0$, the current pulse is removed. Simulate the resultant action potential. (b) The value of the threshold potential is defined as when $I_{Na} > I_K + I_l$. Changes in threshold potential can be easily implemented by changing the value of 25 in the equation for α_m to a lower value. Suppose the value of 25 is changed to 10 in the equation defining α_m and a current pulse of -10 μA (hyperpolarizes the membrane) is passed through the membrane of a squid giant axon for a very long time. At time $t = 0$, the current pulse is removed. Simulate the resultant action potential.

29. Simulate the plots shown in Figures 12.20 to 12.22 in the voltage clamp mode.

30. Select an input current waveform necessary to investigate the refractory period that follows an action potential. (Hint: Use a two-pulse current input.) What is the minimum refractory period? If the second pulse is applied before the minimum refractory period, how much larger is the stimulus magnitude that is needed to generate an action potential?

31. Explain whether the following circuit allows the investigator to conduct a voltage clamp experiment. The unmyelinated axon is sealed at both ends. A silver wire is inserted inside the axon that eliminates R_a. Clearly state any assumptions.

Continued



References

[1] T.D. Coates, Neural Interfacing: Forging the Human-Machine Connection, Morgan & Claypool Publishers, LaPorte, Colorado, 2008.
[2] J. Keener, J. Sneyd, Mathematical Physiology, Springer, New York, 1998.
[3] F. Rodriguez Campos, J.D. Enderle, Porting Genesis to SIMULINK, in: Proceedings of the 30th IEEE EMBS Annual International Conference, September 2–5, 2004.
[4] J.C. Sanchez, J.C. Principe, Brain-Machine Interface Engineering, Morgan & Claypool Publishers, LaPorte, Colorado, 2007.

Suggested Readings

A.T. Bahill, Bioengineering: Biomedical, Medical and Clinical Engineering, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.
J.D. Bronzino, The Biomedical Engineering Handbook, CRC Press, Inc., Boca Raton, FL, 2000.
J.M. Bower, D. Beeman, The Book of Genesis: Exploring Realistic Neural Models with the General Neural Simulation System, Springer-Verlag, New York, 1998.
S. Deutsch, A. Deutsch, Understanding the Nervous System—An Engineering Perspective, IEEE Press, New York, 1993.
D. DiFrancesco, D. Noble, A model of cardiac electrical activity incorporating ionic pumps and concentration changes, Philos. Trans. R. Soc. Lond. B Biol. Sci. 307 (1985) 307–353.

J.D. Enderle, Neural Control of Saccades, in: J. Hyönä, D. Munoz, W. Heide, R. Radach (Eds.), The Brain's Eyes: Neurobiological and Clinical Aspects to Oculomotor Research, Progress in Brain Research, vol. 140, Elsevier, Amsterdam, 2002, pp. 21–50.

J.F. Fulton, H. Cushing, A bibliographical study of the Galvani and Aldini writings on animal electricity, Ann. Sci. 1 (1936) 239–268.

A.C. Guyton, J.E. Hall, Textbook on Medical Physiology, ninth ed., W.B. Saunders Co., Philadelphia, 1995.

B. Hille, Ionic Channels of Excitable Membranes, second ed., Sunderland, Massachusetts, 1992.

A. Hodgkin, A. Huxley, B. Katz, Measurement of current-voltage relations in the membrane of the giant axon of Loligo, J. Physiol. (London) 116 (1952) 424–448.

A. Hodgkin, A. Huxley, Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo, J. Physiol. (London) 116 (1952) 449–472.

A. Hodgkin, A. Huxley, The components of membrane conductance in the giant axon of Loligo, J. Physiol. (London) 116 (1952) 473–496.

A. Hodgkin, A. Huxley, The dual effect of membrane potential on sodium conductance in the giant axon of Loligo, J. Physiol. (London) 116 (1952) 497–506.

A. Hodgkin, A. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, J. Physiol. (London) 117 (1952) 500–544.

E.R. Kandel, J.H. Schwartz, T.M. Jessell, Principles of Neural Science, fifth ed., McGraw-Hill, New York, 2000.

C. Koch, I. Segev, Methods in Neuronal Modeling, From Ions to Networks, MIT Press, Cambridge, Massachusetts, 1998.

C. Luo, Y. Rudy, A dynamic model of the cardiac ventricular action potential: I. Simulations of ionic currents and concentration changes, Circ. Res. 74 (1994) 1071.

G.G. Matthews, Cellular Physiology of Nerve and Muscle, Blackwell Scientific Publications, Boston, 1991.

W. Nernst, Die elektromotorishe Wirksamkeit der Jonen, Z. Physik. Chem. 4 (1889) 129–188.

R. Plonsey, R.C. Barr, Bioelectricity: A Quantitative Approach, Plenum Press, New York, 1982.

R. Northrop, Introduction to Dynamic Modeling of Neurosensory Systems, CRC Press, Boca Raton, 2001.

J. Rinzel, Electrical excitability of cells, theory and experiment: Review of the Hodgkin-Huxley foundation and an update, in: M. Mangel (Ed.), Bull. Math. Biology: Classics of Theoretical Biology, vol. 52, 1990, , pp. 5–23.

J.V. Tranquillo, Quantitative Neurophysiology, Morgan & Claypool Publishers, LaPorte, Colorado, 2008.

This page intentionally left blank

CHAPTER

13

Physiological Modeling

John D. Enderle, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Describe the process used to build a mathematical physiological model.
- Qualitatively describe a saccadic eye movement.
- Describe the saccadic eye movement system with a second order model.
- Explain the importance of the pulse-step saccadic control signal.
- Explain how a muscle operates using a nonlinear and linear muscle model.
- Simulate a saccade with a fourth order saccadic eye movement model.
- Estimate the parameters of a model using system identification.

13.1 INTRODUCTION

A *quantitative* physiological model is a mathematical representation that approximates the behavior of an actual physiological system. *Qualitative* physiological models, most often used by biologists, describe the actual physiological system without the use of mathematics. Quantitative physiological models, however, are much more useful and are the subject of this chapter. Physiological systems are almost always dynamic and mathematically characterized with differential equations. The modeling techniques developed in this chapter are intimately tied to many other interdisciplinary areas, such as physiology, biophysics, and biochemistry, and involve electrical and mechanical analogs. A model is usually constructed using basic and natural laws. This chapter extends this experience by presenting models that are more complex and involve larger systems.

Creating a model is always accompanied by carrying out an experiment and obtaining data. The best experiment is one that provides data that are related to variables used in the model. Consequently, the design and execution of an experiment is one of the most important and time-consuming tasks in modeling. A model constructed from basic and natural laws then becomes a tool for explaining the underlying processes that cause the experimental data and predicting the behavior of the system to other types of stimuli. Models serve as vehicles for thinking, organizing complex data, and testing hypotheses. Ultimately, modeling's most important goals are the generation of new knowledge, prediction of observations before they occur, and assistance in designing new experiments.

Figure 13.1 illustrates the typical steps in developing a model. The first step involves observations from an experiment or a phenomenon that leads to a conjecture or a verbal description of the physiological system. An initial hypothesis is formed via a mathematical model. The strength of the model is tested by obtaining data and testing the model against the data. If the model performs adequately, the model is satisfactory, and a solution is stated. If the model does not meet performance specifications, then the model is updated, and additional experiments are carried out. Usually some of the variables in the model are observable and some are not. New experiments provide additional data that increase the understanding of the physiological system by providing information about previously unobservable variables, which improves the model. The process of testing the model against the data continues until a satisfactory solution is attained. Usually a statistical test is performed to test the goodness of fit between the model and the data. One of the characteristics of a good model is how well it predicts the future performance of the physiological system.

The introduction of the digital computer, programming languages, and simulation software has caused a rapid change in the use of physiological models. Before digital computers, mathematical models of biomedical systems were either oversimplified or involved a great deal of hand calculation, as described in the Hodgkin-Huxley investigations published in 1952. Today, digital computers have becomes so common that the terms *modeling* and *simulation* have almost become synonymous. This has allowed the development of much more realistic or homeomorphic models that include as much knowledge as possible about the structure and interrelationships of the physiological system without any overriding concern about the number of calculations. Models of neuron networks and muscle cross-bridge models involving thousands of differential equations are becoming quite commonplace.

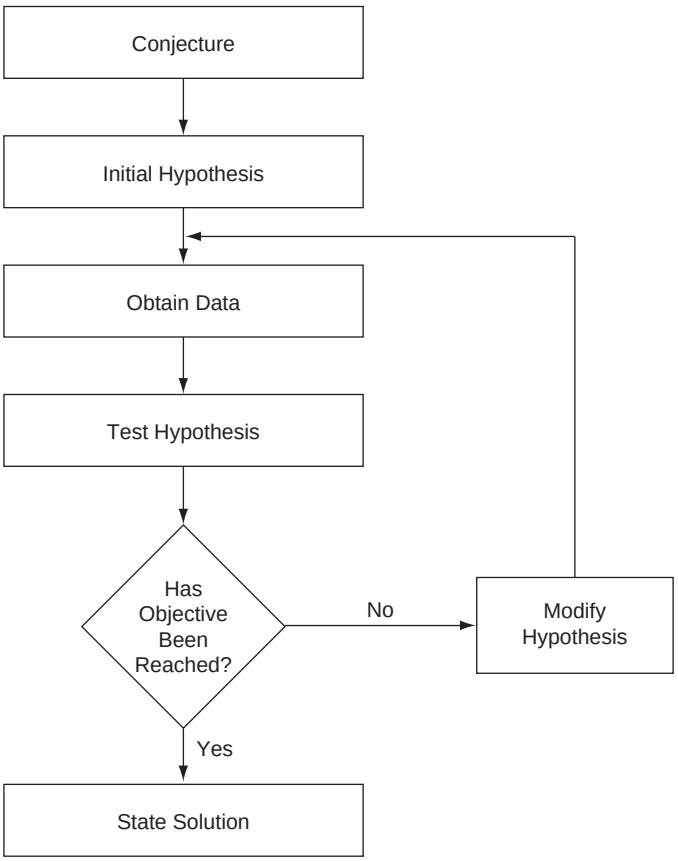


FIGURE 13.1 Flow chart for modeling.

While models can continue to be made more complex, it is important to evaluate the value added with each stage of complexity. The model should be made as simple as possible to explain the data but not so simple that it becomes meaningless. On the other hand, a model that is made too complex is also of little use.

13.1.1 Deterministic and Stochastic Models

A deterministic model is one that has an exact solution that relates the independent variables of the model to one another and to the dependent variable. For a given set of initial conditions, a deterministic model yields the same solution each and every time. A stochastic model involves random variables that are functions of time and include probabilistic considerations. For a given set of initial conditions, a stochastic model yields a different solution each and every time. Suffice it to say that solutions involving stochastic models are much more involved than the solution for a deterministic model.

It is interesting to note that all deterministic models include some measurement error. The measurement error introduces a probabilistic element into the deterministic model, so it might be considered stochastic. However, in this chapter, models are deterministic if their principle features lead to definitive predictions. On the other hand, models are stochastic if their principle features depend on probabilistic elements. This chapter is primarily concerned with deterministic models.

13.1.2 Solutions

There are two types of solutions available to the modeler. A closed form solution exists for models that can be solved by analytic techniques, such as solving a differential equation using the classical technique or by using Laplace transforms. For example, given the following differential equation

$$\ddot{x} + 4\dot{x} + 3x = 9$$

with initial conditions $x(0) = 0$ and $\dot{x}(0) = 1$, the solution is found as

$$x(t) = \left(e^{-3t} - 4e^{-t} + 3\right)u(t)$$

A numerical or simulation solution exists for models that have no closed form solution. Consider the following function:

$$x = \int_{-20}^{20} \frac{e^{\frac{-1}{2}\left(\frac{t-7}{33}\right)^2}}{33\sqrt{2\pi}}\, dt$$

This function (the area under a Gaussian curve) has no closed form solution and must be solved using an approximation technique, such as the trapezoidal rule for integration. Most nonlinear differential equations do not have an exact solution and must be solved via an iterative method or simulation package such as SIMULINK. This was the situation in Chapter 12 when the Hodgkin-Huxley model was solved.

Inverse Solutions

Engineers often design and build systems to a predetermined specification. They often use a model to predict how the system will behave because a model is efficient and economical. The model that is built is called a plant and consists of parameters that completely describe the system: the characteristic equation. The engineer selects the parameters of the plant to achieve a certain set of specifications such as rise time, settling time, or peak overshoot time.

In contrast, biomedical engineers involved with physiological modeling do not build the physiological system but only observe the behavior of the system—the input and output of the system—and then characterize it with a model. Characterizing the model as illustrated in Figure 13.1 involves identifying the form or structure of the model, collecting data, and then using the data to estimate the parameters of the model. The goal of physiological modeling is not to design a system but to identify the components (or parameters) of the system. Most often, data needed for building the model are not the data that can be

collected using existing bioinstrumentation, biosensors, and biosignal processing, as discussed in Chapters 8–10. Typically, the recorded data are transformed from measurement data into estimates of the variables used in the model. Collecting appropriate data is usually the most difficult aspect of the discovery process.

Model building typically involves estimating the parameters of the model that optimize, in a mean square error sense, the output of the model or model prediction, $\hat{x}_i$, and the data, x_i. For example, one metric for estimating the parameters of a model, S, is given by minimizing the sum of squared errors between the model prediction and the data

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

where ε_i is the error between the data x_i and the model prediction $\hat{x}_i$. This technique provides an unbiased estimate with close correspondence between the model prediction and the data.

In order to provide a feeling for the modeling process described in Figure 13.1, this chapter focuses on one particular system—the fast eye movement system—the modeling of which began with early muscle modeling experiences in the 1920s and continues today with neural network models for the control of the fast eye movement system. This physiological system is probably the best understood of all systems in the body. Some of the reasons for this success are the relative ease in obtaining data, the simplicity of the system in relation to others, and the lack of feedback during dynamic changes in the system. In this chapter, a qualitative description of the fast eye movement system is presented, followed by the first model of the system by Westheimer, who used a second-order model published in 1954. The 1964 model of the system by Robinson is next presented because of its fundamental advances in describing the input to the system. With the physical understanding of the system in place, a detailed presentation of muscle models is given with the early work of Levin and Wyman in 1927, and Fenn and Marsh in 1935. These muscle models are important in developing a realistic model that accurately depicts the system. Using the more accurate muscle models, the fast eye movement model is revisited by examining the model presented by Bahill and coworkers and then several models by Enderle and coworkers. Next, the control mechanism for this system is described from the basis of physiology, systems control theory, and neural networks based on anatomical pathways. Finally, the topic of system identification or parameter estimation closes the chapter. The literature on the fast eye movement system is vast, and the material covered in this chapter is not exhaustive but rather a representative sample from the field. Not covered in this chapter at all is how visual information is collected and processed by the body and how the body reacts to the information.

13.2 AN OVERVIEW OF THE FAST EYE MOVEMENT SYSTEM

The visual system is our most important sensory system. It provides a view of the world around us captured with receptors in the eyeball that is transmitted to the central nervous system (CNS). The eye movement or oculomotor system is responsible for movement of the

eyes so images are clearly seen. The oculomotor system also responds to auditory and vestibular sensory stimuli. A saccadic or fast eye movement involves quickly moving the eye from one image to another image. This type of eye movement is very common, and it is observed most easily while reading—that is, when the end of a line is reached, the eyes are moved quickly to the beginning of the next line. Saccades are also used to locate or acquire targets. Smooth pursuit is a slow eye movement used to track an object as it moves by keeping the eyes on the target. In addition to these two movements, the eye movement system also includes the vestibular ocular movement, optokinetic eye movement, and vergence movement. Vestibular ocular movements are used to maintain the eyes on the target during head movements. Optokinetic eye movements are reflex movements that occur when moving through a target-filled environment or to maintain the eyes on target during continuous head rotation. The optokinetic eye movement is a combination of saccadic and slow eye movements that keeps a full-field image stable on the retina during sustained head rotation. Each of these four eye movements is a conjugate eye movement—that is, movements of both eyes together driven by a common neural source. Vergence eye movements use nonconjugate eye movements to keep the eyes on the target. If the target moves closer, the eyes converge, and if the target moves farther away, they diverge. Each of these movements is controlled by a different neural system, and all of these controllers share the same final common pathway to the eye muscles.

Each eye can be moved within the orbit in three directions: vertically, horizontally, and torsionally. These movements are due to three pairs of agonist-antagonist muscles. These muscles are called antagonistic pairs because their activity opposes each other and follows the principle of reciprocal innervation. Figure 13.2 shows the muscles of the eye, optic nerve, and the eyeball. We refer to the three muscle pairs and the eyeball as the oculomotor plant, and the oculomotor system as the oculomotor plant and the neural system controlling the eye movement system.

At the rear of the eyeball is the retina, shown in Figure 13.3. Regardless of the input, the oculomotor system is responsible for movement of the eyes so images are focused on

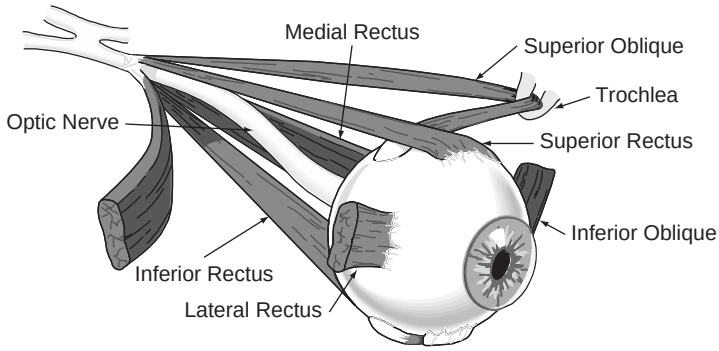


FIGURE 13.2 The muscles, eyeball, and optic nerve of the right eye. The left eye is similar except the lateral and medial rectus muscles are reversed. The lateral and medial rectus muscles are used to move the eyes in a horizontal motion. The superior rectus, inferior rectus, superior oblique, and inferior oblique are used to move the eyes vertically and torsionally. The contribution from each muscle depends on the position of the eye. When the eyes are looking straight ahead, called primary position, the muscles are stimulated and under tension.

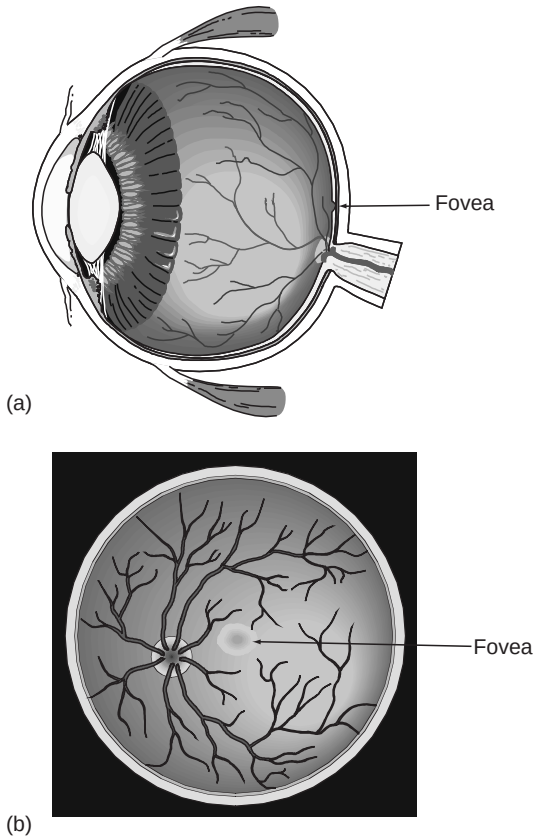


FIGURE 13.3 (A) Side view of the eye. The rear surface of the eye is called the retina. The retina is part of the central nervous system and consists of two photoreceptors: rods and cones. (B) Front view looking at the rear inside surface (retina) of the eye. The fovea is located centrally and is approximately 1 mm in diameter. The oculomotor system maintains targets centered on the fovea.

the central one-half degree region of the retina, known as the fovea. Lining the retina are photoreceptive cells that translate images into neural impulses. These impulses are then transmitted along the optic nerve to the central nervous system via parallel pathways to the superior colliculus and the cerebral cortex. The fovea is more densely packed with photoreceptive cells than the retinal periphery—thus a higher-resolution image (or higher visual acuity) is generated in the fovea than the retinal periphery. The purpose of the fovea is to allow us to *clearly* see an object, and the purpose of the retinal periphery is to allow us to *detect* a new object of interest. Once a new object of interest is detected in the periphery, the system redirects the eyes to the new object.

One of the most successfully studied systems in the human is the oculomotor or eye movement system. Some of the reasons for this success are the relative ease in obtaining data, the simplicity of the system, and the lack of feedback during dynamic changes in the system. A saccade is a fast eye movement that involves quickly moving the eyes from

one target or image to another. The word *saccade* originated from the French word *saquer*, which means to jerk the reins of a horse. A saccade is a very quick and jerky movement of the eye from one target to another.

The eye muscles are among the fastest in the human body, with a 10° saccade taking only 50 ms. The saccadic system can be thought of as a targeting system that is concerned only with accurate and swift eye movements from one target to another without concern for the information swept across the retina during the eye movement. During a saccade, the visual system is turned off. After the saccade is complete, the system operates in a closed-loop mode to ensure that the eyes reached the correct destination. Information from the retina and muscle proprioceptors is used to correct any error between the desired and current eye position. The saccade system operates in a closed-loop mode to reduce this error to zero with a corrective saccade. One possible explanation of the operation of the neural control of saccades is that the saccadic neural controller is an open-loop time-optimal system using an internal closed-loop controller [11, 13, 16, 50]. This system does not rely on muscle proprioceptors or real time visual feedback to ensure accuracy of movement because the eye movements occur too fast. Instead, a complex neural network involving the mesencephalon, cerebellum, brainstem, and cerebrum keeps track of the eye movement.

A typical experiment for recording saccades has the subject sitting before a horizontal target display of small light emitting diodes (LEDs), as shown in Figure 13.4 (left). The subject is instructed to maintain their eyes on the lit LED by moving their eyes as fast as possible to avoid errors. A saccade is made by the subject when the active LED is switched off and another LED is switched on. Eye movements can be recorded using a variety of techniques, including electrooculography, video oculography, scleral search coil, and

FIGURE 13.4 Experimental setup to record a saccade (left) and an infrared eye movement recorder (right). The eye movement recorder is based on the design by Engelken et al. [21].

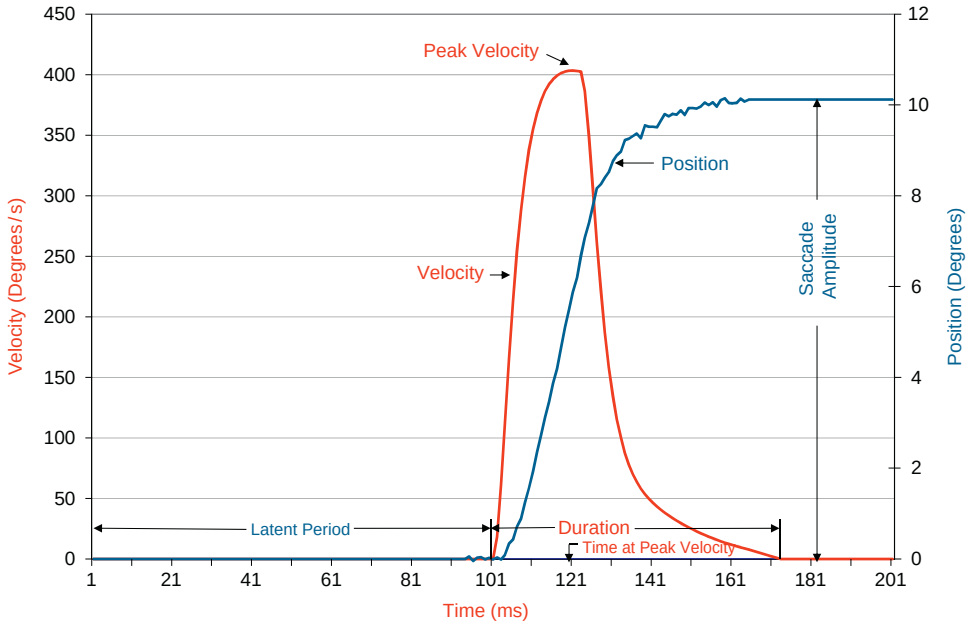


FIGURE 13.5 A 10° saccade with various indices labeled.

infrared oculography (shown in Figure 13.4 (right)). A typical saccade is shown in Figure 13.5, with a latent period of approximately 100 ms, amplitude of 10°, and a duration of approximately 60 ms. Saccadic eye movements are conjugate and ballistic, with a typical duration of 30–100 ms and a latency of 100–300 ms. The latent period is thought to be the time interval during which the CNS determines whether to make a saccade and, if so, calculates the distance the eyeball is to be moved, transforming retinal error into transient muscle activity. Also shown in this figure is the velocity of the saccade with a peak velocity of approximately $400°s^{-1}$.

Generally, saccades are extremely variable, with wide variations in the latent period, time to peak velocity, peak velocity, and saccade duration. Furthermore, variability is well coordinated for saccades of the same size; saccades with lower peak velocity are matched with longer saccade durations, and saccades with higher peak velocity are matched with shorter saccade durations. Thus, saccades driven to the same destination usually have different trajectories.

To appreciate differences in saccade dynamics, it is often helpful to describe them with saccade main sequence diagrams [3, 17, 27]. The main sequence diagrams plot saccade peak velocity–saccade magnitude, saccade duration–saccade magnitude, and saccade latent period–saccade magnitude. The saccade size or amplitude is the angular displacement from the initial position to its destination. The size of a saccade ranges from less than a degree (microsaccades) to 45° in both the nasal (toward the nose) and temporal (toward the temple) directions. Peak or maximum velocity occurs at approximately half the duration of the saccade for small saccades less than 15° [2]. The duration of a saccadic eye movement is the time from the start to the end of a saccade. Duration is usually hard to determine

from the saccade amplitude versus time graph, but it is more easily seen in the velocity versus time graph, as shown in Figure 13.5. Saccade durations can range from approximately 30 ms for saccades less than 5° and up to 100 ms for large saccades. For saccades greater than 7°, there is a linear relationship between saccade amplitude and duration. The latent period is the time interval from when a target appears until the eyes begin to move.

Figure 13.6 shows the main sequence characteristics for a subject executing 26 saccades. The subject actually executed 52 saccades in both the positive and negative directions, with only the results of the saccades in the positive direction displayed in Figure 13.6 for simplicity. Note that saccade characteristics moving to the left are different from those moving to the right. The solid lines in the figures include a fit to the data. Peak velocity–saccade magnitude is basically a linear function until approximately 15°, after which it levels off to a constant for larger saccades. Many researchers have fit this relationship to an exponential function. The line in graph (A) is fitted to the nonlinear equation

$$v_{\max} = \alpha\left(1 - e^{-\frac{x}{\beta}}\right) \tag{13.1}$$

where $v_{\max}$ is the maximum velocity, x the saccade size, and the constants α and β evaluated to minimize the summed error squared between the model and the data. Note that α is to represent the steady-state peak velocity–saccade magnitude curve, and β is to represent the "time constant" for the peak velocity–saccade magnitude curve. For this data set for positive eye movements, α equals 825, and β equals 9.3.

A similar pattern is observed with eye movements moving in the negative direction (not shown), but the parameters are $\alpha = 637$ and $\beta = 6.9$, which are typically different from the values computed for the positive direction. The exponential shape of the peak velocity–saccade amplitude relationship might suggest that the system is nonlinear if a step input to the system is assumed. A step input provides a linear peak velocity–saccade amplitude relationship. In fact, the saccade system is not driven by a step input but rather a more complex pulse-step waveform, as discussed later. Thus, the saccade system cannot be assumed to be nonlinear solely based on the peak velocity–saccade amplitude relationship.

Figure 13.6B shows the data depicting a linear relationship between saccade duration–saccade magnitude. If a step input is assumed, then the dependence between saccade duration and saccade magnitude also might suggest that the system is nonlinear. A linear system with a step input always has a constant duration. Since the input is not characterized by a step waveform, the saccade system cannot be assumed to be nonlinear solely based on the saccade duration–saccade magnitude relationship.

Figure 13.6C shows the latent period–saccade magnitude data. It is quite clear that the latent period does not show any linear relationship with saccade size—that is, the latent period's value appears independent of saccade size. However, some other investigators have proposed a linear relationship between the latent period and saccade magnitude. This feature is unimportant for the presentation in this book, since in the development of the oculomotor plant models, the latent period is implicitly assumed within the model.

Because of the complexity of the eye movement system, attention is restricted to horizontal fast eye movements. In reality, the eyeball is capable of moving horizontally, vertically, and torsionally. An appropriate model for this system would include a model for each muscle

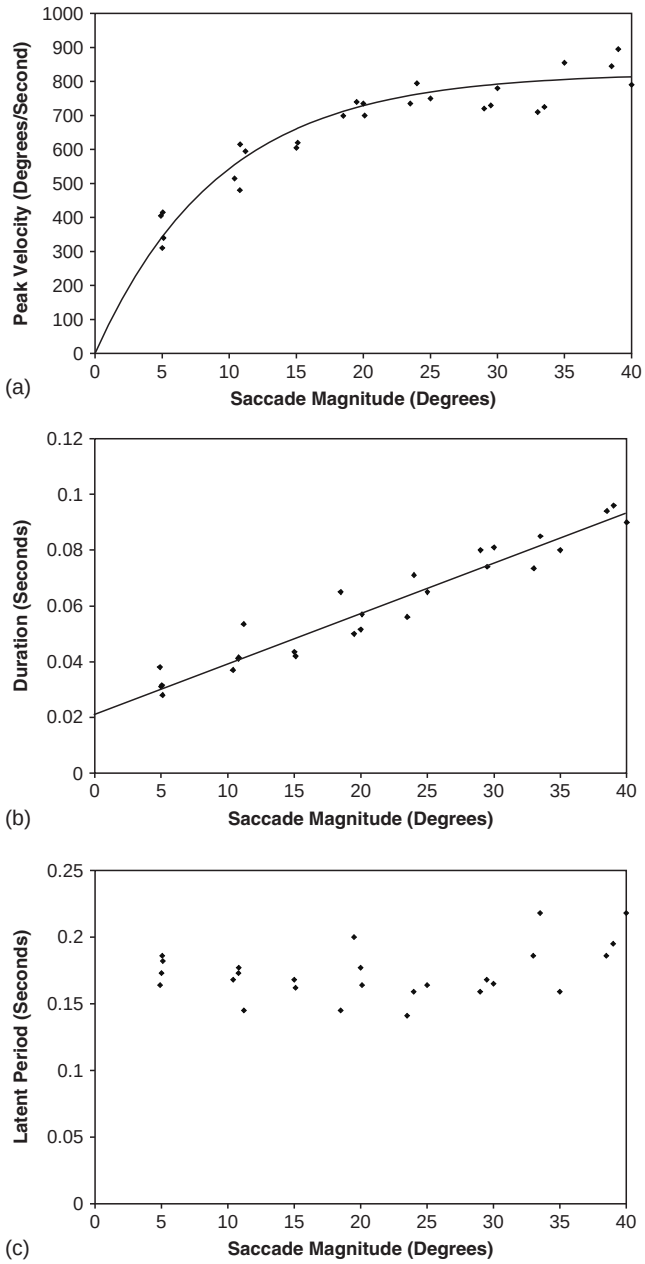


FIGURE 13.6 Main sequence diagrams for positive saccades. Similar shapes are observed for negative saccades. (a) Peak velocity–saccade magnitude, (b) saccade duration-saccade magnitude, and (c) latent period-saccade magnitude for 26 saccadic movements by a single subject. *Adapted from [9].*

and a separate controller for each muscle pair. The development of the horizontal saccadic eye movement models in this chapter are historical and are presented in increasing complexity with models of muscle introduced out of sequence so their importance is fully realized. Not every oculomotor model is discussed. A few are presented for illustrative purposes.

13.3 THE WESTHEIMER SACCADIC EYE MOVEMENT MODEL

The first quantitative saccadic eye movement model was published by Westheimer in 1954. In this model, he described horizontal saccades in response to a 20° target displacement. A mechanical description of the model is given in Figure 13.7, and a system description is given in Eq. (13.2).

$$J\ddot{\theta} + B\dot{\theta} + K\theta = \tau(t) \tag{13.2}$$

To analyze the characteristics of this model and compare it to data, Laplace variable analysis is used. Assuming zero initial conditions, the Laplace transform of Eq. (13.2) yields

$$\theta\left(s^2J + sB + K\right) = \tau(s) \tag{13.3}$$

and as a transfer function in standard form

$$H(s) = \frac{\theta}{\tau} = \frac{1}{s^2J + sB + K} = \frac{\dfrac{\omega_n^2}{K}}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{13.4}$$

where according to Westheimer's data for a 20° saccade,

$$\omega_n = \sqrt{\frac{K}{J}} = 120 \text{ and } \zeta = \frac{B}{2\sqrt{KJ}} = 0.7.$$

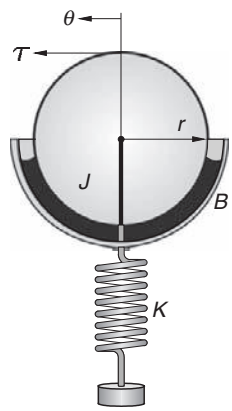


FIGURE 13.7 Westheimer's second-order model of the saccade system. The parameters J, B, and K are rotational elements for moment of inertia, friction, and stiffness, respectively, and represent the eyeball and its associated viscoelasticity. The torque applied to the eyeball by the lateral and medial rectus muscles is given by $\tau(t)$, and θ is the angular eye position. The radius of the eyeball is r.

According to the data, the roots are complex and given by

$$s_{1,2} = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2} = -84 \pm j85.7.$$

The input to this system is a step input $\tau(s) = \frac{\gamma}{s}$, and $\theta(t)$ is determined as

$$\theta(t) = \frac{\gamma}{K}\left[1 + \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}}\cos\left(\omega_n\sqrt{1-\zeta^2}t + \psi\right)\right] \tag{13.5}$$

where

$$\psi = \pi + \tan^{-1}\frac{-\zeta}{\sqrt{1-\zeta^2}}.$$

EXAMPLE PROBLEM 13.1.

Show the intermediate steps in going from Eq. (13.4) to Eq. (13.5).

Solution

Substituting the input, $\tau(s) = \frac{\gamma}{s}$, into Eq. (13.4) yields

$$\theta(s) = \frac{\gamma\frac{\omega_n^2}{K}}{s\left(s^2 + 2\zeta\omega_n s + \omega_n^2\right)}$$

Assuming a set of complex roots based on the estimates from Westheimer, a partial fraction expansion gives

$$\theta(s) = \frac{\frac{\gamma}{K}}{s} + \frac{\frac{\gamma}{2K\left[(\zeta^2-1) - j\zeta\sqrt{1-\zeta^2}\right]}}{\left(s + \zeta\omega_n - j\omega_n\sqrt{1-\zeta^2}\right)} + \frac{\frac{\gamma}{2K\left[(\zeta^2-1) + j\zeta\sqrt{1-\zeta^2}\right]}}{\left(s + \zeta\omega_n + j\omega_n\sqrt{1-\zeta^2}\right)}$$

where the magnitude numerator of the complex terms is

$$M = \left|\frac{\gamma}{2K\left[(\zeta^2-1) - j\zeta\sqrt{1-\zeta^2}\right]}\right| = \frac{\gamma}{2K\left((\zeta^2-1)^2 + \zeta^2(1-\zeta^2)\right)^{\frac{1}{2}}} = \frac{\gamma}{2K(1-\zeta^2)^{\frac{1}{2}}}$$

and the phase angle is

$$\phi = \tan^{-1}\frac{-\zeta\sqrt{1-\zeta^2}}{(\zeta^2-1)} = \tan^{-1}\frac{-\zeta\sqrt{1-\zeta^2}}{-\left(\sqrt{1-\zeta^2}\right)^2} = \tan^{-1}\frac{-\zeta}{-\sqrt{1-\zeta^2}}$$

The solution is given by

$$\theta(t) = \frac{\gamma}{K}\left[1 + \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}}\cos\left(\omega_n\sqrt{1-\zeta^2}t + \psi\right)\right]$$

Continued

where $\psi = \pi + \tan^{-1}\frac{-\zeta}{\sqrt{1-\zeta^2}}$. It is always helpful to check analysis results and one easy point to check is usually at time zero.

$$\theta(0) = \frac{\gamma}{K}\left[1 + \frac{\cos(\phi)}{\sqrt{1-\zeta^2}}\right] = \frac{\gamma}{K}\left[1 + \frac{-\sqrt{1-\zeta^2}}{\sqrt{1-\zeta^2}}\right] = 0$$

as it should, since the saccade starts at primary position or $\theta(0) = 0$.

To fully explore the quality of a model, it is necessary to compare its performance against the data. For a saccade, convenient metrics are time to peak overshoot, which gives an indication of saccade duration and peak velocity. These metrics were discussed previously when the main sequence diagram was described in Section 13.2.

The time to peak overshoot of saccade model, T_p, is found by first calculating

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial t}\left[\frac{\gamma}{K}\left(1 + \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}}\cos\left(\omega_n\sqrt{1-\zeta^2}t + \psi\right)\right)\right]\Bigg|_{t=T_p} = 0 \tag{13.6}$$

Using the chain rule to evaluate Eq. (13.6) and substituting $t = T_p$, yields

$$T_p = \frac{\pi}{\omega_d} = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}}. \tag{13.7}$$

With Westheimer's constants of $\zeta = 0.7$ and $\omega_n = 120$, we find that $T_p = \frac{\pi}{120\times\sqrt{1-0.7^2}} = 37$ ms. T_p in the Westheimer model is independent of saccade size and not in agreement with experimental data presented in the main sequence diagram. The data show a saccade duration that increases with saccade amplitude, where this model has a constant duration.

EXAMPLE PROBLEM 13.2

Show that Eq. (13.7) follows from Eq. (13.6). Find the value of $\theta(T_p)$.

Solution

From Eq. (13.6), we have

$$\frac{\gamma}{K\sqrt{1-\zeta^2}}\left[-\zeta\omega_n\cos\left(\omega_d T_p + \psi\right)e^{-\zeta\omega_n t} - \omega_d e^{-\zeta\omega_n t}\sin\left(\omega_d T_p + \psi\right)\right] = 0 \tag{13.8}$$

Eq. (13.8) is rewritten as

$$-\zeta\cos\left(\omega_d T_p + \psi\right) = \sqrt{1-\zeta^2}\sin\left(\omega_d T_p + \psi\right)$$

which yields

$$\tan\left(\omega_d T_p + \psi\right) = \frac{-\zeta}{\sqrt{1-\zeta^2}} = \tan\psi \tag{13.9}$$

Eq. (13.9) is true whenever $\omega_d T_p = n\pi$. The time to peak amplitude is the smallest value that satisfies Eq. (13.9), which is $n = 1$. Thus, $T_p = \dfrac{\pi}{\omega_d} = \dfrac{\pi}{\omega_n\sqrt{1-\zeta^2}}$. To determine $\theta(T_p)$, note that $\omega_{\mathrm{d}} t + \psi = \pi + \psi$, and

$$\begin{aligned}\theta(T_p) &= \frac{\gamma}{K}\left[1 + \frac{e^{\frac{-\zeta\pi}{\sqrt{1-\zeta^2}}}}{\sqrt{1-\zeta^2}}\cos(\pi+\psi)\right] \\ &= \frac{\gamma}{K}\left[1 + \frac{e^{-\frac{\zeta\pi}{\sqrt{1-\zeta^2}}}}{\sqrt{1-\zeta^2}} \times \sqrt{1-\zeta^2}\right] \\ &= \frac{\gamma}{K}\left(1 + e^{\frac{-\pi\zeta}{\sqrt{1-\zeta^2}}}\right)\end{aligned} \tag{13.10}$$

An important aid in examining the suitability of a model is to study the model predictions and data estimates of higher-order derivatives. If there are problems with the model, these problems are amplified when comparing estimates of the higher-order derivatives with model predictions. For the Westheimer model, maximum velocity is found from $\left.\dfrac{\partial^2\theta}{\partial t^2}\right|_{t=T_{mv}} = 0$, where T_{mv} is the time at peak velocity. Using the solution given by Eq. (13.6), we next compute

$$\begin{aligned}\frac{\partial^2\theta}{\partial t^2} &= \frac{\partial}{\partial t}\left[\frac{-\gamma}{K}\frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}}\{\zeta\omega_n\cos(\omega_d t+\psi) + \omega_d\sin(\omega_d t+\psi)\}\right] \\ &= \frac{-\gamma}{K\sqrt{1-\zeta^2}}[-\zeta\omega_n e^{-\zeta\omega_n t}(\zeta\omega_n\cos(\omega_d t+\psi) + \omega_d\sin(\omega_d t+\psi)) \\ &\quad + e^{-\zeta\omega_n t}(-\zeta\omega_n\omega_d\sin(\omega_d t+\psi) + \omega_d^2\cos(\omega_d t+\psi))] = 0\end{aligned} \tag{13.11}$$

After taking the second derivative in Eq. (13.11), we find that

$$T_{mv} = \frac{1}{w_d}\tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \tag{13.12}$$

$\theta(T_{\mathrm{m}v})$ can be evaluated using T_{mv}.

Using Westheimer's parameter values with any arbitrary saccade magnitude given by $\Delta\theta = \frac{\gamma}{K}$ and Eq. (13.10) gives

$$\dot{\theta}(T_{mv}) = 55.02 \times \Delta\theta \tag{13.13}$$

Equation (13.13) indicates that peak velocity is directly proportional to saccade magnitude. As illustrated in the main sequence diagram shown in Figure 13.6, experimental peak velocity data have an exponential form and do not represent a linear function as predicted by the Westheimer model. Both Eqs. (13.7) and (13.13) are consistent with linear systems theory. That is, for a step input to a linear system, the duration (and time to peak overshoot) stays constant regardless of the size of the input, and the peak velocity increases with the size of the input.

EXAMPLE PROBLEM 13.3

Show that Eq. (13.12) follows from (13.11).

Solution

Beginning with Eq. (13.11), we have

$$\frac{-\gamma}{K\sqrt{1-\zeta^2}}\left[-\zeta\omega_n e^{-\zeta\omega_n t}(\zeta\omega_n \cos(\omega_d t+\psi) + \omega_d \sin(\omega_d t+\psi)) + e^{-\zeta\omega_n t}\left(-\zeta\omega_n\omega_d \sin(\omega_d t+\psi) + \omega_d^2 \cos(\omega_d t+\psi)\right)\right] = 0$$

The terms multiplying the sinusoids are removed, since they do not equal zero. Therefore,

$$\left(\omega_d^2 - \zeta^2\omega_n^2\right)\cos(\omega_d t+\psi) - 2\zeta\omega_n\omega_d \sin(\omega_d t+\psi) = 0$$

which reduces to

$$\frac{\omega_d^2 - \zeta^2\omega_n^2}{2\zeta\omega_n\omega_d} = \frac{\sin(\omega_d t+\psi)}{\cos(\omega_d t+\psi)} = \tan(\omega_d t+\psi) \tag{13.14}$$

Substituting $\omega_d = \omega_n\left(1-\zeta^2\right)$ into Eq. (13.14), we have

$$\frac{\omega_n^2\left(1-\zeta^2\right) - \zeta^2\omega_n^2}{2\zeta\omega_n\omega_n\left(1-\zeta^2\right)} = \frac{1-2\zeta^2}{2\zeta\sqrt{1-\zeta^2}} = \tan(\omega_d t+\psi) \tag{13.15}$$

With $\tan\psi = \frac{-\zeta}{\sqrt{1-\zeta^2}}$, we factor out $\frac{-\zeta}{\sqrt{1-\zeta^2}}$ in Eq. (13.15), substitute $\tan\psi$, giving

$$\frac{1-2\zeta^2}{2\zeta\sqrt{1-\zeta^2}} = \frac{-\zeta}{\sqrt{1-\zeta^2}}\left(1-\frac{1}{2\zeta^2}\right) = \left(1-\frac{1}{2\zeta^2}\right)\tan\psi = \tan(\omega_d t+\psi) \tag{13.16}$$

Now

$$\tan(\omega_d t+\psi) = \frac{\tan(\omega_d t) + \tan\psi}{1 - \tan(\omega_d t)\tan\psi} \tag{13.17}$$

and

$$\tan\phi = \tan\psi$$

Substituting for $\tan(\omega_d t + \psi)$ from Eq. (13.17) into Eq. (13.16) gives

$$\left(1 - \frac{1}{2\zeta^2}\right)\tan\psi = \frac{\tan(\omega_d t) + \tan\psi}{1 - \tan(\omega_d t)\tan\psi} \tag{13.18}$$

Multiplying both sides of Eq. (13.18) by $(1 - \tan\omega_d t \tan\psi)$ gives

$$\left(1 - \frac{1}{2\zeta^2}\right)\tan\psi - \left(1 - \frac{1}{2\zeta^2}\right)\tan(\omega_d t)(\tan\psi)^2 = \tan(\omega_d t) + \tan\psi \tag{13.19}$$

Collecting like terms in Eq. (13.19) gives

$$\tan(\omega_d t)\left(1 + (\tan\psi)^2\left(1 - \frac{1}{2\zeta^2}\right)\right) = \frac{-\tan\psi}{2\zeta^2} \tag{13.20}$$

Dividing both sides of Eq. (13.20) by the term multiplying $\tan(\omega_d t)$ gives

$$\tan(\omega_d t) = \frac{-\tan\psi}{2\zeta^2\left(1 + (\tan\psi)^2\left(\frac{2\zeta^2 - 1}{2\zeta^2}\right)\right)} = \frac{-\tan\psi}{2\zeta^2 + (\tan\psi)^2(2\zeta^2 - 1)} \tag{13.21}$$

With $\tan\psi = \tan\phi = \dfrac{-\zeta}{\sqrt{1-\zeta^2}}$, we have

$$\tan(\omega_d t) = \frac{\dfrac{\zeta}{\sqrt{1-\zeta^2}}}{2\zeta^2 + \dfrac{\zeta^2}{1-\zeta^2}(2\zeta^2 - 1)} = \frac{\dfrac{\zeta}{\sqrt{1-\zeta^2}}}{\zeta^2\left(2 + \dfrac{(2\zeta^2 - 1)}{1-\zeta^2}\right)} = \left(\frac{1-\zeta^2}{1-\zeta^2}\right)\frac{\dfrac{\zeta}{\sqrt{1-\zeta^2}}}{\zeta^2\left(2 + \dfrac{(2\zeta^2 - 1)}{1-\zeta^2}\right)}$$

$$= \frac{\sqrt{1-\zeta^2}}{\zeta(2(1-\zeta^2) + (2\zeta^2 - 1))} = \frac{\sqrt{1-\zeta^2}}{\zeta}$$

Taking the inverse tangent of the previous equation gives $T_{mv} = \dfrac{1}{w_d}\tan^{-1}\left(\dfrac{\sqrt{1-\zeta^2}}{\zeta}\right)$.

Westheimer noted the differences between saccade duration–saccade magnitude and peak velocity–saccade magnitude in the model and the experimental data and inferred that the saccade system was not linear because the peak velocity–saccade magnitude plot was nonlinear. He also noted that the input was not an abrupt step function. Overall, this model provided a satisfactory fit to the eye position data for a saccade of 20° but not for saccades of other magnitudes. Interestingly, Westheimer's second-order model proves to be an adequate model for saccades of all sizes if a different input function, as described in the next section, is assumed. Due to its simplicity, the Westheimer model of the oculomotor plant is still popular today.

EXAMPLE PROBLEM 13.4

Simulate a 20° saccade using the Westheimer saccade model. Plot position versus time.

Solution

The Laplace transform of the Westheimer saccade model with a step input is

$$\theta(s) = H(s)\tau(s) = \frac{\frac{\omega_n^2}{K}}{s^2 + 2\zeta\omega_n s + \omega_n^2} \times \frac{\gamma}{s}$$

and based on the model, the SIMULINK program is shown in Figure 13.8 (top). Using the parameters given by Westheimer, $\zeta = 0.7$ and $\omega_n = 120$, and the input $\gamma = 20$, the response is shown in Figure 13.8 (bottom). Notice that the response is underdamped as expected and that $T_p = 37$ ms.

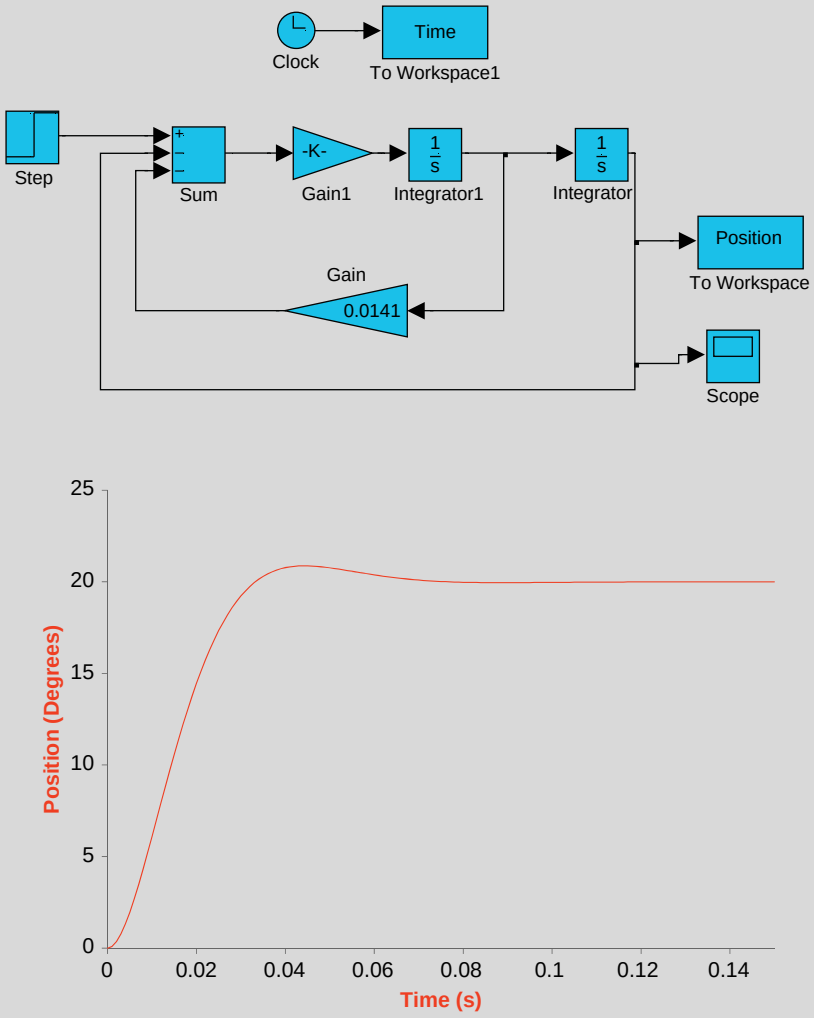


FIGURE 13.8 SIMULINK program and response for Example Problem 13.4.

13.4 THE SACCADE CONTROLLER

One of the challenges in modeling physiological systems is the lack of data or information about the input to the system. For instance, in the fast eye movement system the input is the neurological signal from the CNS to the muscles connected to the eyeball. Information about the input is not available in this system, since it involves thousands of neurons firing at a very high rate. Recording the signal would involve invasive surgery and instrumentation that was not available in the 1960s. Often, however, it is possible to obtain information about the input via indirect means, as described in this section.

In 1964, Robinson performed an experiment in an attempt to measure the input to the eyeballs during a saccade. To record the input, one eye was held fixed using a suction contact lens, while the other eye performed a saccade from target to target. Since the same innervation signal is sent to both eyes during a saccade, Robinson inferred that the input, recorded through the transducer attached to the fixed eyeball, was the same input driving the other eyeball. He proposed that muscle tension driving the eyeballs during a saccade is a pulse plus a step or, simply, a pulse-step input (Figure 13.9).

Today, microelectrode studies are carried out to record the electrical activity in oculomotor neurons in monkeys. Figure 13.10 shows a micropipette being used to record the activity in the oculomotor nucleus, an important neuron population responsible for driving a saccade. Additional experiments on oculomotor muscle have been carried out to learn more about the saccade controller since Robinson's initial study. For instance in 1975, Collins and his coworkers reported using a miniature "C" gauge force transducer to measure muscle tension in vivo at the muscle tendon during unrestrained human eye movements. This type of study has allowed a better understanding of the tensions exerted by each muscle, rather than the combined effect of both muscles, as shown in Figure 13.11.

It is important to distinguish between the tension or force generated by a muscle, called muscle tension, and the force generator within the muscle, called the *active-state tension generator*. The active-state tension generator creates a force within the muscle that is transformed through the internal elements of the muscle into the muscle tension. Muscle tension is external and measurable. Active-state tension is internal and not measurable. Active-state tension

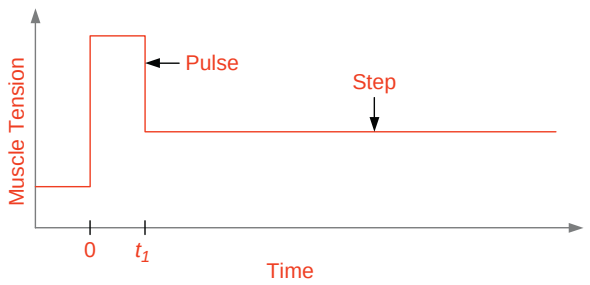


FIGURE 13.9 The muscle tension recorded during a saccade.

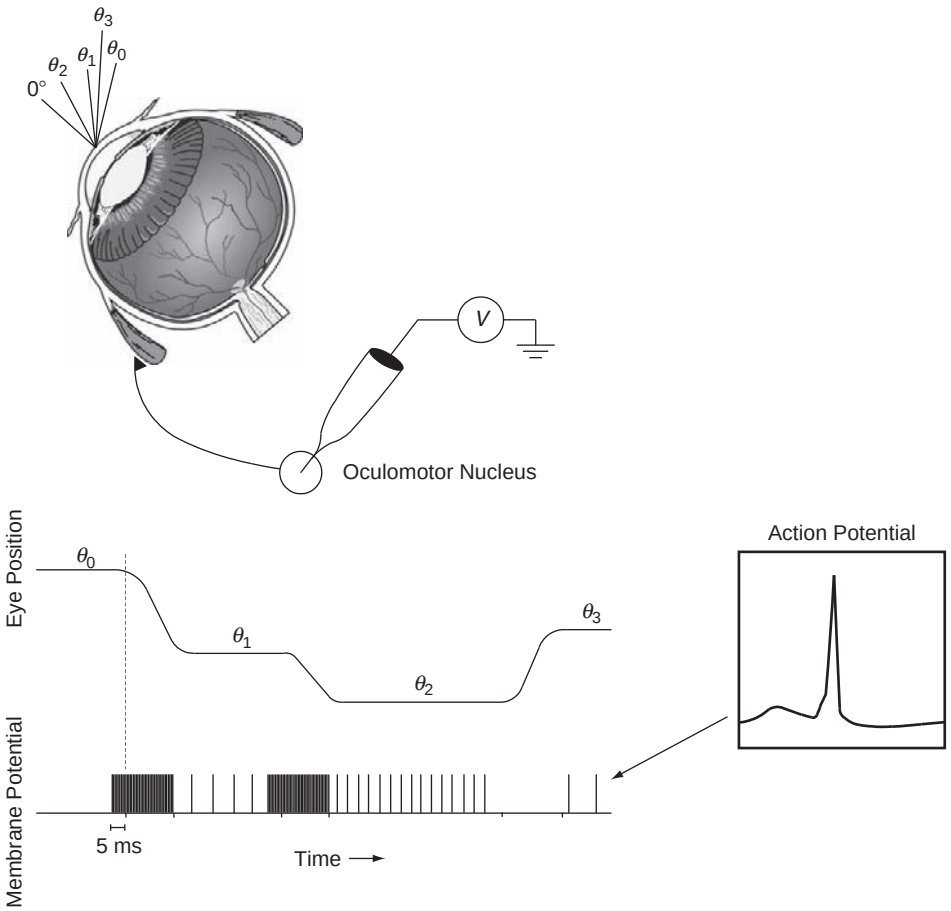


FIGURE 13.10 Recording of a series of saccades using a micropipette and the resultant electrical activity in a single neuron. Spikes in the membrane potential indicate an action potential occurred. Saccade neural activity initiates with a burst of neural firing approximately 5 ms before the eye begins to move and continues until the eye has almost reached its destination. Relative position of the eye is shown at the top with angles θ_0 through θ_3. Initially, the eye starts in position θ_0, a position in the extremity in which the muscle is completely stretched with zero input. To move the eye from θ_0 to θ_1, neural burst firing occurs. To maintain the eye at θ_1, a steady firing occurs in the neuron. The firing rate for fixation is in proportion to the shortness of the muscle. Next, the eye moves from θ_1 to θ_2. This saccade moves much more slowly than the first saccade with approximately the same duration as the first. The firing level is also approximately at the same level as the first. The difference in input corresponds to fewer neurons firing to drive the eye to its destination, which means a smaller input than the first saccade. Because the muscle is shorter after completing this saccade, the fixation firing rate is higher than the previous position at θ_1. Next, the eye moves in the opposite direction to θ_3. Since the muscle is lengthening, the input to the muscle is zero—that is, no action potentials are used to stimulate the muscle. The fixation firing level θ_3 is less than that for θ_1 because the muscle is longer.

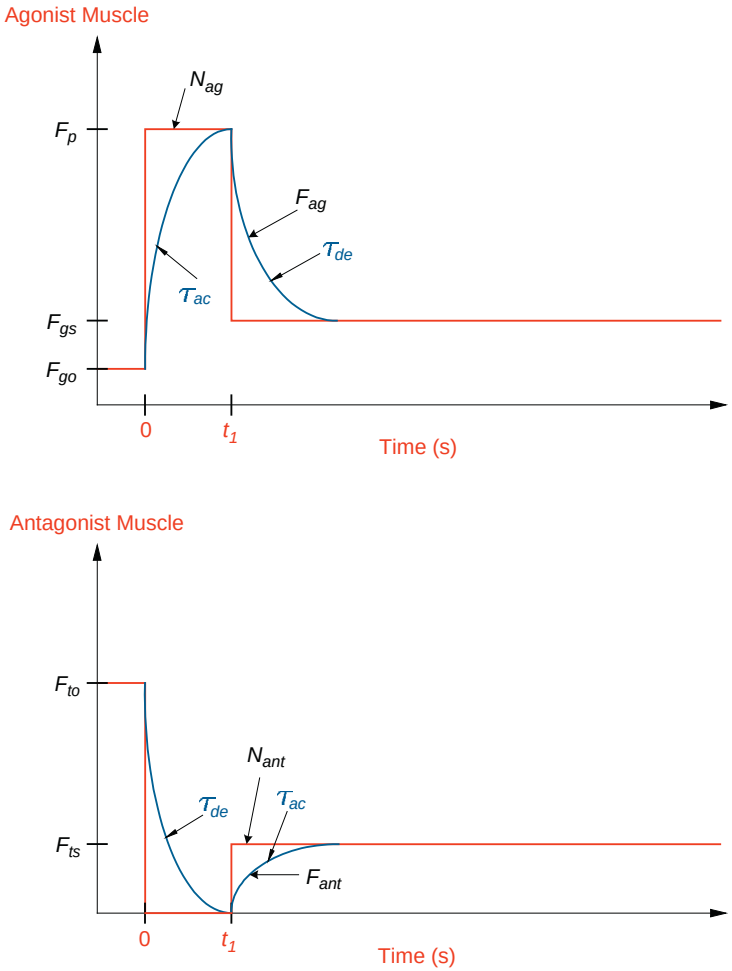


FIGURE 13.11 Agonist, N_{ag}, and antagonist, N_{ant}, control signals and the agonist, F_{ag}, and antagonist, F_{ant}, active-state tensions. Note that the time constant for activation, τ_{ac}, is different from the time constant for deactivation, τ_{de}. The time interval, t_1, is the duration of the pulse.

follows most closely the neural input to the muscle. From Figure 13.11, a pattern of neural activity is observed as follows:

1. The muscle that is being contracted (agonist) is stimulated by a pulse, followed by a step to maintain the eyeball at its destination.
2. The muscle that is being stretched (antagonist) is unstimulated during the saccade (stimulated by a pause or a negative pulse to zero), followed by a step to maintain the eyeball at its destination.

Figure 13.11 quantifies these relationships for the agonist neural input, N_{ag}, and the antagonist neural input, N_{ant}. The pulse input is required to get the eye to the target as soon

as possible, and the step is required to keep the eye at that location. It has been reported that the active-state tensions are not identical to the neural controllers but are described by low-pass filtered pulse-step waveforms. The active-state tensions, F_{ag} and F_{ant}, are shown in Figure 13.11 with time-varying time constants τ_{ac} and τ_{de}. It is thought that the low-pass filtering involves the movement of Ca^{++} across the cell membrane.

Some investigators have reported a different set of time constants for the agonist and antagonist activity, and others have noted a firing frequency–dependent agonist activation time constant. Others suggest that the agonist activation time constant is a function of saccade magnitude. For simplicity in this chapter, activation and deactivation time constants are assumed to be identical for both agonist and antagonist activity. The parameters are defined as follows:

F_{go} is the initial agonist active-state tension before the saccade starts
F_P is the maximum agonist active-state tension
F_{gs} is the steady-state agonist active-state tension after the saccade ends
F_{to} is the initial antagonist active-state tension before the saccade starts
F_{ts} is the steady-state antagonist active-state tension after the saccade ends

Generally, the pulse is used to get the eyeball to the target quickly, and the step is required to keep the eye at that location. The same innervation signal is sent to both eyes, and as a result, they move together. We call this a conjugate eye movement.

13.5 DEVELOPMENT OF AN OCULOMOTOR MUSCLE MODEL

It is clear that an accurate model of muscle is essential for the development of a model of the horizontal fast eye movement system that is driven by a pair of muscles (lateral and medial rectus muscles). In fact, the Westheimer model does not include any muscle model and relies solely on the inertia of the eyeball, friction between the eyeball and socket, and elasticity due to the optic nerve and other attachments as the elements of the model. In this section, the historical development of a muscle model is discussed as it relates to the oculomotor system. Muscle model research involves a broad spectrum of topics, ranging from the nano models that deal with the sarcomeres to macro models, in which collections of cells are grouped into a lumped parameter system and described with ordinary mechanical elements. Here the focus is on a macro, or lumped parameter model of the oculomotor muscle based on physiological evidence from experimental testing. The model elements, as presented, consist of an active-state tension generator (input), elastic elements, and viscous elements. Each element is introduced separately and the muscle model is incremented in each subsection. It should be noted that the linear muscle model presented in Sections 13.7 and 13.8 completely revises the subsections before it. The earlier subsections are presented because of their historical significance and to appreciate the current muscle model.

13.5.1 Muscle Model Passive Elasticity

Consider the experiment of stretching an unexcited muscle and recording tension to investigate the passive elastic properties of muscle. The data curve shown in Figure 13.12 is a typical recording of the tension observed in an eye rectus muscle. The tension required

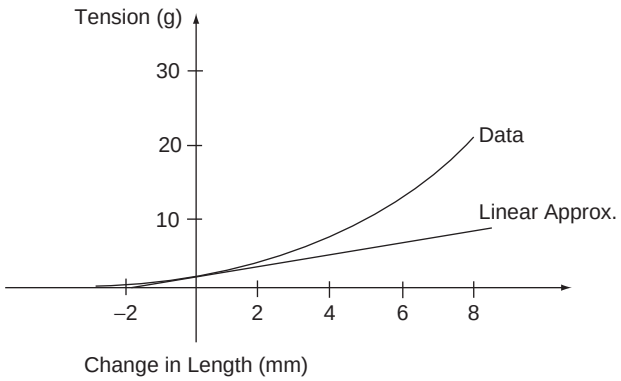


FIGURE 13.12 The tension-displacement curve for unexcited muscle. The slope of the linear approximation to the data is muscle passive elasticity, K_{pe}.

to stretch a muscle is a *nonlinear* function of distance. Thus, in order to precisely model this element, a nonlinear spring element should be used. Note that the change in length at 0 refers to the length of the muscle at primary position (looking straight ahead). Thus, the eye muscles are stretched, approximately 3 mm, when the eye movement system is at rest in primary position. At rest, the muscle length is approximately 37 mm.

To be useful in a linear model of muscle, Figure 13.12 should be linearized in the vicinity of an operating point. The operating point should be somewhat centered in the region in which the spring operates. In Figure 13.12, a line tangent to the curve at primary position provides a linear approximation to the spring's behavior in this region as done historically. For ease in analysis, the following relationships hold for a sphere representing the eyeball radius of 11 mm.

$$1\ \text{g} = 9.806 \times 10^{-3}\ \text{N}$$

$$1^\circ = 0.192\ \text{mm} = 1.92 \times 10^{-4}\ \text{m}$$

The slope of the line, K_{pe}, is approximately

$$K_{pe} = 0.2\ \frac{\text{g}}{^\circ} = 0.2\ \frac{\text{g}}{^\circ} \times \frac{9.806 \times 10^{-3}\ \text{N}}{1\ \text{g}} \times \frac{1^\circ}{1.92 \times 10^{-4\text{m}}} = 10.2\ \frac{\text{N}}{\text{m}}$$

and represents the elasticity of the passive elastic element.

The choice of the operating region is of vital importance concerning the slope of the curve. At this time, a point in the historical operating region of rectus muscle is used. In most of the oculomotor literature, the term K_{pe} is typically subtracted out of the analysis and is not used. The operating point will be revisited in Section 13.7, and this element will be completely removed from the model.

13.5.2 Active-State Tension Generator

In general, a muscle produces a force in proportion to the amount of stimulation. The element responsible for the creation of force is the active-state tension generator. Note that this

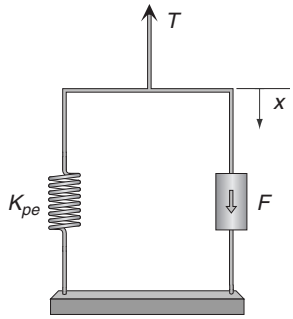


FIGURE 13.13 A muscle model consisting of an active-state tension generator, F, and a passive elastic element, K_{pe}. Upon stimulation of the active-state tension generator, a tension, T, is exerted by the muscle.

terminology is used so there is no confusion with the force created within the muscle when the tension created by the muscle is discussed. The active tension generator is included along with the passive elastic element in the muscle model as shown in Figure 13.13. The relationship between tension, T, active-state tension, F, and elasticity is given by

$$T = F - K_{pe}x \tag{13.21}$$

Isometric (constant length) experiments have been performed on humans over the years to estimate the active tension generator at different levels of stimulation. These experiments were usually performed in conjunction with strabismus surgery when muscles were detached and reattached to correct crossed eyes. Consider the tension created by a muscle when stimulated as a function of length as shown in Figure 13.14. The data were collected from the lateral rectus muscle that was detached from one eyeball, while the other unoperated eyeball fixated at different locations in the nasal (N) and temporal (T) directions from $-45°$ to $45°$. This experiment was carried out under the assumption that the same neural input is sent to each eyeball (Hering's Law of equal innervation), so the active-state tension in the freely moving eyeball should be the same as that in the detached lateral rectus muscle. At each fixation point, the detached lateral rectus muscle was stretched and tension data was recorded at each of the points indicated on the graph. The thick line represents the muscle tension at that particular eye position under normal conditions. The curve for $45°$ T is the zero stimulation case and represents the passive mechanical properties of muscle. Note that the tension generated is a *nonlinear* function of the length of the muscle.

To compare the model in Figure 13.13 against the data in Figure 13.14, it is convenient to subtract the passive elasticity in the data (represented by the $45°$ T curve) from each of the data curves $30°$ T through $45°$ N, leaving only the hypothetical active-state tension. Shown in the graph on the left in Figure 13.15 is one such calculation for $15°$ N with the active-state tension given by the dashed line. The other curves in Figure 13.14 give similar results and have been omitted because of the clutter. The dashed line should represent the active-state tension, which appears to be a function of length. If this was a *pure* active-state tension element, the subtracted curve should be a horizontal line indicative of the size of the input.

One such input is shown for the active-state tension in the graph on the right in Figure 13.15. The result in Figure 13.15 implies that either the active-state tension's effect

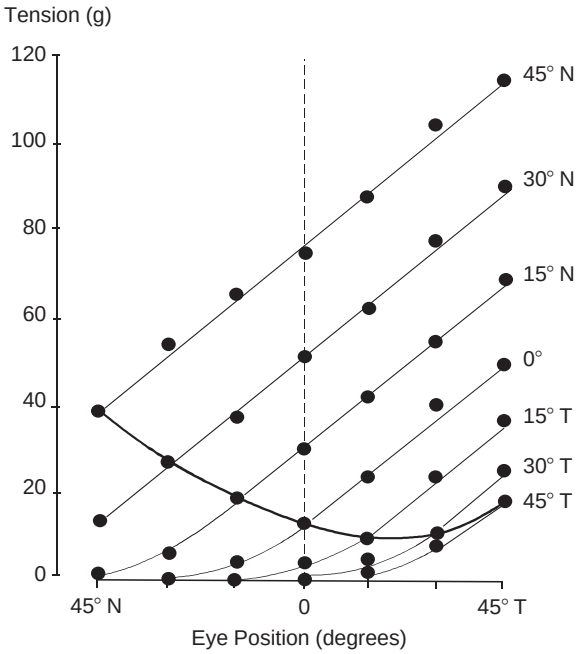


FIGURE 13.14 Length-tension curves for lateral rectus muscle at different levels of activation. Dots represent tension data recorded from the detached lateral rectus muscle during strabismus surgery, while the unoperated eyeball fixated at targets from –45° to 45°. *Adapted from Collins et al. [8].*

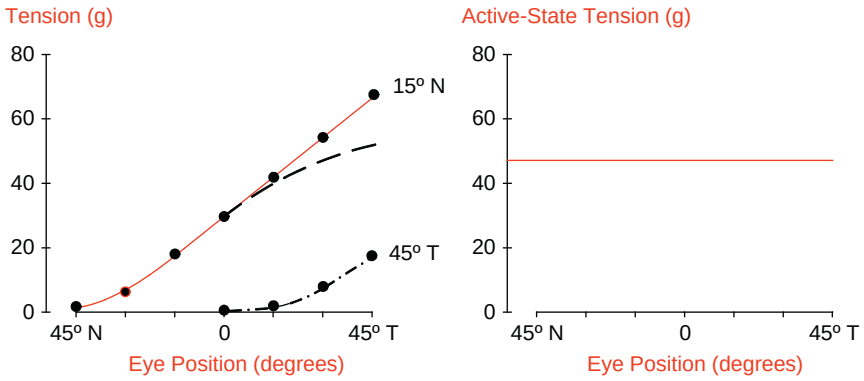


FIGURE 13.15 (Left) Length-tension curves for extraocular muscle at two levels of activation corresponding to the 15° N and 45° T positions. The solid line represents tension data recorded from the detached lateral rectus muscle during strabismus surgery, while the unoperated eyeball fixated at targets. The dashed curve is the 15° curve with the 45° curve (dashed-dot line) subtracted from it. The resultant dashed curve represents the active-state tension as a function of eye position. (Right) The theoretical graph for active-state tension versus eye position as given by Eq. (3.4.1). *Adapted from Collins et al. [8].*

is a nonlinear element—that is, there may be other nonlinear or linear elements missing in the model—or, perhaps, some of the assumptions made in the development of the model are wrong. For the moment, consider that the analysis is correct and assume that some elements are missing. This topic will be revisited in Section 13.7.

13.5.3 Elasticity

The normal operating point (at primary position), L_p, is much shorter than the length at which maximum force occurs at approximately 30°. Even when the effects of the passive muscle are removed, a relationship between length and tension is still evident, as previously described. Let us introduce a new elastic element into the model to account for the relationship between length and tension as shown in Figure 13.16, which is described by the following equation.

$$T = F - K_{pe}x - Kx \tag{13.22}$$

The new elastic element, K, accounts for the slope of the subtracted curve shown as the dashed line in the graph on the left in Figure 13.15. The slope of the line, K, at primary position, is approximately $0.8\ \text{g}/^\circ = 40.86\ \text{N/m}$ (a value typically reported in the literature). The slope for each of the curves in Figure 13.14 can be calculated in the same manner at primary position with the resultant slopes all approximately equal to the same value as the one for 15° N. At this time, the introduction of additional experiments will provide further insight to the development of the muscle model.

Series Elastic Element

Experiments carried out by Levin and Wyman in 1927 and Collins in 1975 indicated the need for a series elasticity element, in addition to the other elements previously presented in the muscle model. The experimental setup and typical data from the experiment are shown in Figure 13.17. The protocol for this experiment, called the quick release experiment, is as follows. (1) A weight is hung onto the muscle. (2) The muscle is fully stimulated at time t_1. (3) The weight is released at time t_2. At time t_2, the muscle changes length almost instantaneously when the weight is released. The only element that can instantaneously change its length is a spring. Thus, to account for this behavior, a spring, called the series elastic element, K_{se}, is connected in series to the active-state tension element. While some

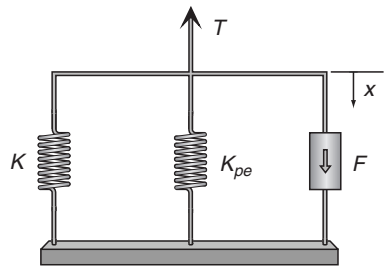


FIGURE 13.16 A muscle model consisting of an active-state tension generator, F, passive elastic element, K_{pe}, and elastic element, K. Upon stimulation of the active-state tension generator, a tension T is exerted by the muscle.

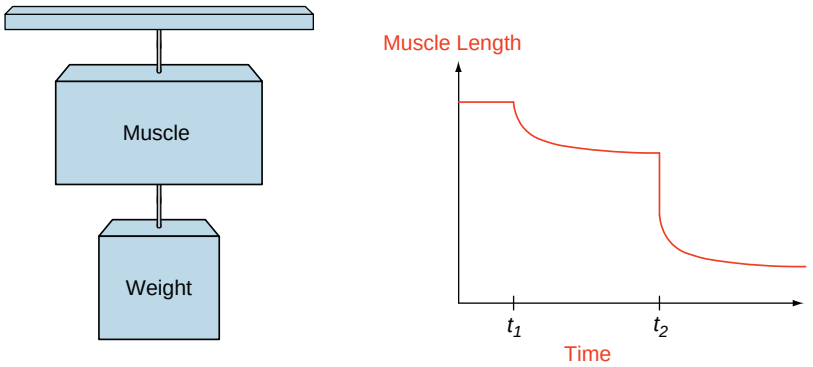


FIGURE 13.17 The quick release experiment. The figure on the left depicts the physical setup of the experiment. The figure on the right shows typical data from the experiment. At time t_1 the muscle is fully stimulated, and at time t_2 the weight is released.

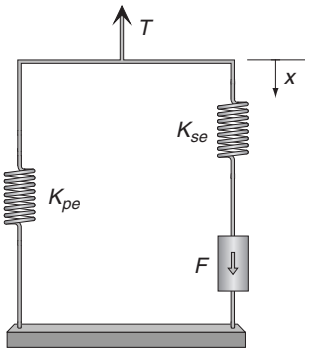


FIGURE 13.18 A muscle model consisting of an active-state tension generator F, passive elastic element K_{pe}, and series elastic element K_{se}. Upon stimulation of the active-state tension generator F, a tension T is exerted by the muscle.

investigators argue that this element is nonlinear, we will assume that it is linear for simplicity. The updated muscle model is shown in Figure 13.18.

Based on the experiment carried out by Collins in 1975 on rectus eye muscle, an estimate for K_{se} was given as 125 N/m (2.5 gm/°). Since the value of K_{se} does not equal the value of K, another elastic element is needed to account for this behavior.

Length Tension Elastic Element

Given the inequality between K_{se} and K, another elastic element, called the length-tension elastic element, K_{lt}, is placed in parallel with the active-state tension element as shown in the illustration on the left in Figure 13.19. For ease of analysis, K_{pe} is subtracted out (removed) using the graphical technique shown in Figure 13.15. To estimate a value for K_{lt}, the muscle model shown on the right in Figure 13.19 is analyzed and reduced to an expression involving K_{lt}. Analysis begins by summing the forces acting on nodes 1 and 2.

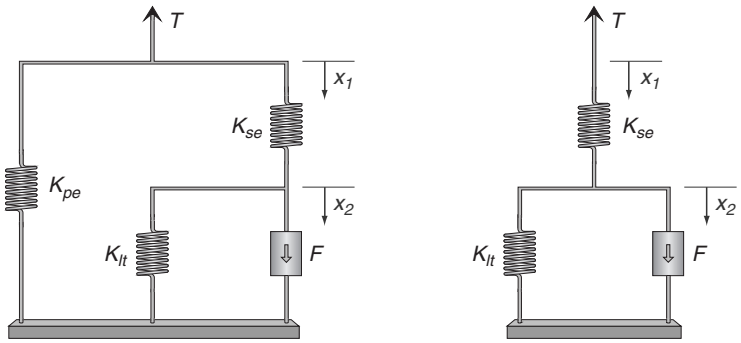


FIGURE 13.19 Diagram on left illustrates a muscle model consisting of an active-state tension generator F in parallel to a length-tension elastic element K_{lt}, connected to a series elastic element $K_{se,}$ all in parallel with the passive elastic element K_{pe}. Upon stimulation of the active-state tension generator F, a tension T is exerted by the muscle. The diagram on the right is the same muscle model except that K_{pe} has been removed.

$$T = K_{se}(x_2 - x_1) \tag{13.23}$$

$$F = K_{lt}x_2 + K_{se}(x_2 - x_1) \quad \rightarrow \quad x_2 = \frac{F + K_{se}x_1}{K_{se} + K_{lt}} \tag{13.24}$$

Substituting x_2 from Eq. (13.24) into (13.23) gives

$$T = \frac{K_{se}}{K_{se} + K_{lt}}(F + K_{se}x_1) - K_{se}x_1 = \frac{K_{se}}{K_{se} + K_{lt}}F - \frac{K_{se}K_{lt}}{K_{se} + K_{lt}}x_1 \tag{13.25}$$

Equation (13.25) is an equation for a straight line with y-intercept $\dfrac{K_{se}}{K_{se} + K_{lt}}F$ and slope $\dfrac{K_{se}K_{lt}}{K_{se} + K_{lt}}$. The slope of the length-tension curve in Figure 13.14 is given by $K = 0.8\ \text{g}/^\circ = 40.86\ \text{N/m}$. Therefore,

$$K = \frac{K_{se}K_{lt}}{K_{se} + K_{lt}} = 40.86\ \frac{N}{m} \tag{13.26}$$

Solving Eq. (13.26) for K_{lt} yields

$$K_{lt} = \frac{K_{se}K}{K_{se} - K} = 60.7\ \frac{N}{m} \tag{13.27}$$

13.5.4 The Force-Velocity Relationship

Early experiments indicated that muscle had elastic as well as viscous properties. Muscle was tested under isotonic (constant force) experimental conditions as shown in Figure 13.20 to investigate muscle viscosity. The muscle and load were attached to a lever with a high lever ratio. The lever reduced the gravity force (mass $\times$ gravity) of the load at the muscle by one over the lever ratio, and the inertial force (mass $\times$ acceleration) of the load by one over the lever ratio squared. With this arrangement, it was assumed that the inertial force exerted by the load during isotonic shortening could be ignored. The second assumption

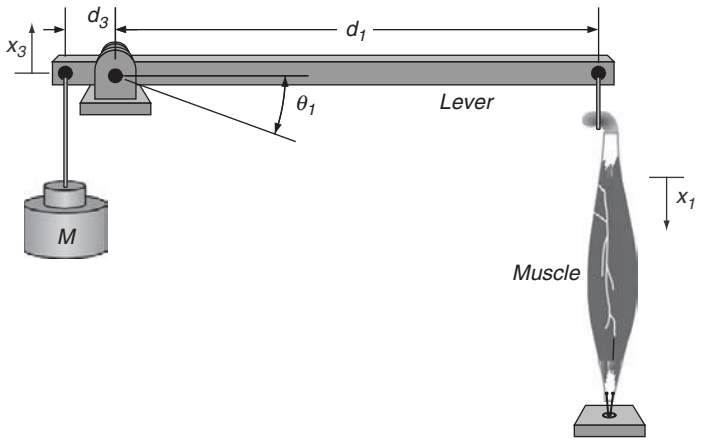


FIGURE 13.20 Drawing of the classical isotonic experiment with inertial load and muscle attached to the lever. The muscle is stretched to its optimal length according to experimental conditions and attached to the ground.

was that if mass was not reduced enough by the lever ratio (enough to be ignored), then taking measurements at maximum velocity provided a measurement at a time when acceleration is zero, and, therefore, inertial force equals zero. If these two assumptions are valid, then the experiment would provide data free of the effect of inertial force as the gravity force is varied.

According to the experimental conditions, the muscle is stretched to its optimal length at the start of the isotonic experiment. The isotonic experiment begins by attaching a load M, stimulating the muscle, and recording position. The two curves in Figure 13.21 depict the time course for the isotonic experiment for a small and large load. Notice that the durations of both responses are approximately equal regardless of the load, in spite of the apparent much longer time delay associated with the large load. Next, notice that the heavier the load, the less the total shortening. Maximum velocity is calculated numerically from the position data. To estimate muscle viscosity, this experiment is repeated with many loads at the same stimulation level, and maximum velocity is calculated. Figure 13.22 illustrates the typical relationship between load ratio (P/P_o) and maximum velocity, where $P = Mg$

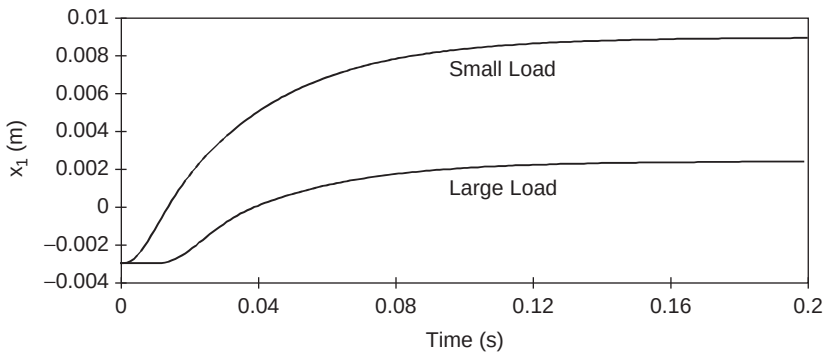


FIGURE 13.21 The typical response of a muscle stimulated with a large and small load.

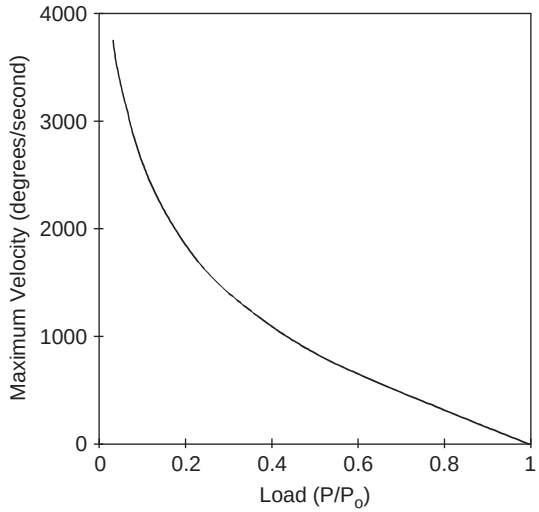


FIGURE 13.22 The force-velocity curve.

and P_o is the isometric tension (the largest weight that the muscle can move) for maximally stimulated muscle. This curve is usually referred to as the *force-velocity curve*.

Clearly, the force-velocity curve is nonlinear and follows a hyperbolic shape. If a smaller stimulus than maximum is used to stimulate the muscle, then a family of force-velocity curves results, as shown in Figure 13.23. Each curve is generated with a different active-state

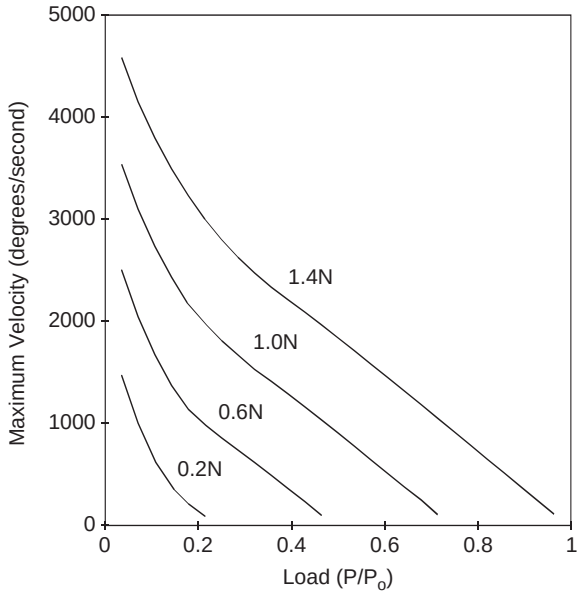


FIGURE 13.23 Family of force-velocity curves for active-state tensions ranging from 1.4 to 0.2 N.

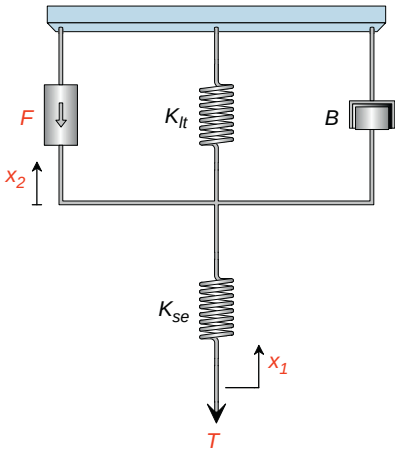


FIGURE 13.24 A muscle model consisting of an active-state tension generator F, in parallel with a length-tension elastic element K_{lt}, and viscous element B, connected to a series elastic element K_{se}. The passive elastic element K_{pe} has been removed from the model for simplicity. Upon stimulation of the active-state tension generator F, a tension T is exerted by the muscle.

tension as indicated. The force-velocity characteristics in Figure 13.23 are similar to those shown in Figure 13.22. In particular, the slope of the force-velocity curve for a small value of active-state tension is quite different from that for a large value of the active-state tension in the operating region of the eye muscle (i.e., approximately 800°/s).

To include the effects of viscosity from the isotonic experiment in the muscle model, a viscous element is placed in parallel with the active-state tension generator and the length tension elastic element as shown in Figure 13.24. The impact of this element is examined by analyzing the behavior of the model in Example Problem 13.5 by simulating the conditions of the isotonic experiment. At this stage, it is assumed that the viscous element is linear in this example. For simplicity, the lever is removed along with the virtual acceleration term $M\ddot{x}_1$. A more thorough analysis including the lever is considered later in this chapter. For simplicity, the passive elastic element K_{pe} is removed from the diagram and analysis.

EXAMPLE PROBLEM 13.5

Consider the system shown in Figure 13.25 that represents a model of the isotonic experiment. Assume that the virtual acceleration term $M\ddot{x}_1$ can be ignored. Calculate and plot maximum velocity as a function of load.

Solution

Assume that $\dot{x}_2 > \dot{x}_1$, and that the mass is supported so $x_1 > 0$. Summing the forces acting on nodes 1 and 2 gives

Continued

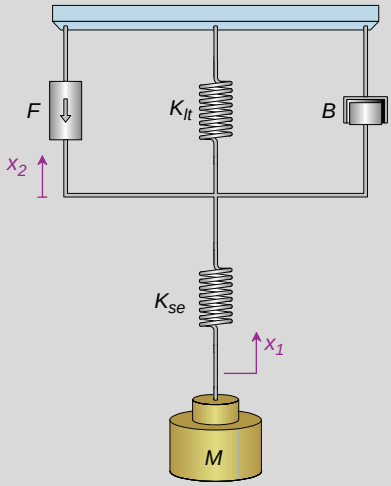


FIGURE 13.25 System for Example Problem 13.5.

$$Mg = K_{se}(x_2 - x_1) \quad \rightarrow \quad x_1 = x_2 - \frac{Mg}{K_{se}}$$

$$F = B\dot{x}_2 + K_{lt}x_2 + K_{se}(x_2 - x_1)$$

Substituting x_1 into the second equation yields

$$F = B\dot{x}_2 + K_{lt}x_2 + Mg$$

Solving the previous equation for x_2 and $\dot{x}_2$ gives

$$x_2(t) = \frac{F - Mg}{K_{lt}}\left(1 - e^{-\frac{K_{lt}t}{B}}\right)$$

$$\dot{x}_2(t) = \frac{F - Mg}{B}e^{-\frac{K_{lt}t}{B}}$$

Maximum velocity, $V_{\max}$, for all loads is given by $V_{\max} = \frac{F-Mg}{B}$ and $\dot{x}_1 = \dot{x}_2$, since $\dot{x}_1 = \frac{d}{dt}\left(x_2 - \frac{Mg}{K_{se}}\right)$. Figure 13.26 depicts a linear relationship between maximum velocity and load.

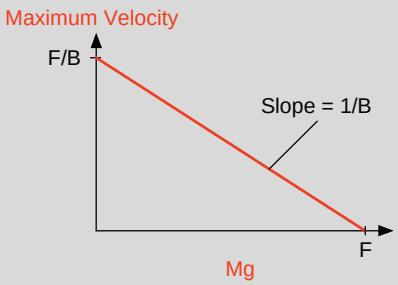


FIGURE 13.26 Result for Example Problem 13.5.

The assumption of a linear viscosity element appears to be in error, since the analysis in Example Problem 13.5 predicts a linear relationship between load and maximum velocity (according to the assumptions of the solution), and the data from the isotonic experiment shown in Figure 13.22 is clearly nonlinear. Thus, a reasonable assumption is that the viscosity element is nonlinear.

Traditionally, muscle viscosity is characterized by the nonlinear Hill hyperbola, given by

$$V_{\max}(P + a) = b(P_0 - P) \tag{13.28}$$

where $V_{\max}$ is the maximum velocity, P is the external force, P_o the isometric tension, and a and b are the empirical constants representing the asymptotes of the hyperbola. As described previously, P_o represents the isometric tension, which is the largest weight that the muscle can move, and P is the weight Mg. Hill's data suggests that

$$a = \frac{P_0}{4} \quad \text{and} \quad b = \frac{V_{\max}}{4}$$

Therefore, with these values for a and b, the Hill equation is rewritten from Eq. (13.28) as

$$P = P_0 - \frac{V_{\max}(P_0 + a)}{b + V_{\max}} = P_0 - BV_{\max} \tag{13.29}$$

where

$$B = \frac{P_0 + a}{b + V_{\max}} \tag{13.30}$$

The term B represents the viscosity of the element. Clearly, the force due to viscosity is nonlinear due to the velocity term, $V_{\max}$, in the denominator of Eq. (13.29).

In oculomotor models, $V_{\max}$ is usually replaced by $\dot{x}_2$, P is replaced by muscle tension, T, and P_o is replaced by the active-state tension, F, as defined from Figure 13.24. Therefore, Eqs. (13.29) and (13.30) are rewritten as

$$T = F - BV \tag{13.31}$$

where

$$B = \frac{F + a}{b + \dot{x}_2} \tag{13.32}$$

Some oculomotor investigators have reported values for a and b in the Hill equation that depend on whether the muscle is being stretched or contracted. There is some evidence to suggest that stretch dynamics are different from contraction dynamics. However, the form of the viscosity expression for muscle shortening or lengthening is given by Eq. (13.32), with values for a and b parameterized appropriately. For instance, Hsu and coworkers [33] described the viscosity for shortening and lengthening for oculomotor muscles as

$$B_{ag} = \frac{F_{ag} + AG_a}{\dot{x}_2 + AG_b} \tag{13.33}$$

$$B_{ant} = \frac{F_{ant} - ANT_a}{-\dot{x}_2 - ANT_b} \tag{13.34}$$

where AG_a, AG_b, ANT_a, and ANT_b are parameters based on the asymptotes for contracting (agonist) or stretching (antagonist), respectively.

13.5.5 A Linear Muscle Model

At this time, we will put all of the elements that have been discussed into a model of muscle as shown in Figure 13.27 (left), and then we will analyze this model to determine the tension created by the muscle. Note that in the muscle model, we have subtracted out the effects of passive elasticity and assumed that $\dot{x}_2 > \dot{x}_1$.

Thus, starting with the free body diagram in Figure 13.27 (right), we have our two node equations as

$$T = K_{se}(x_2 - x_1)$$
$$F = B\dot{x}_2 + K_{lt}x_2 + K_{se}(x_2 - x_1)$$

We solve for x_2 from the node 1 equation as $x_2 = \frac{T}{K_{se}} + x_1$, and we substitute it into the node 2 equation, giving us

$$F = B\dot{x}_2 + (K_{se} + K_{lt})x_2 - K_{se}x_1 = B\dot{x}_2 + K_{st}\left(\frac{T}{K_{se}} + x_1\right) - K_{se}x_1$$

For convenience, we will let $K_{st} = K_{se} + K_{lt}$, and we will multiply the previous equation by K_{se} and rearrange terms, so we have

$$K_{se}F = K_{se}B\dot{x}_2 + K_{st}T + K_{se}K_{lt}x_1$$

or

$$T = \frac{K_{se}F}{K_{st}} - \frac{K_{se}K_{lt}}{K_{st}}x_1 - \frac{K_{se}B}{K_{st}}\dot{x}_2$$

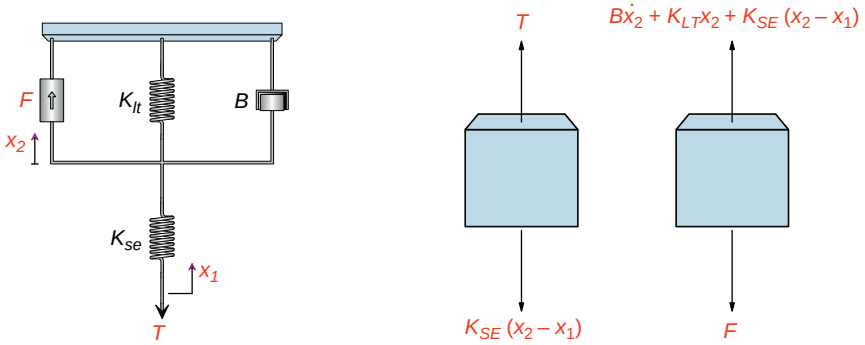


FIGURE 13.27 (Left) Updated model of muscle with active-state tension generator, length-tension elastic element, series elastic element, and viscosity element. (Right) Free body diagram of the system on the left.

Equation (13.32) or Eqs. (13.33) and (13.34), depending on whether the muscle is contracting or lengthening, can be substituted for parameter B in the preceding equation to give a nonlinear model of oculomotor muscle.

13.5.6 Passive Tissues of the Eyeball

At this point, we return to modeling the eyeball. As previously discussed, Robinson not only described the passive properties of muscle, but he also determined the elasticity, viscosity, and inertia of the eyeball from his experiments during strabismus surgery. With the two horizontal muscles, Figure 13.28 describes the passive tissues of the eyeball. Note that the passive elasticity of the eyeball, K_p, is a combination of the effects due to the four other muscles, the optic nerve, and so on. The viscous element of the eyeball, B_p, is due to the friction of the eyeball within the eye socket.

13.5.7 Activation and Deactivation Time Constants

The control signal that the central nervous system sends to the oculomotor system during a saccade is a pulse-step signal as described in Figure 13.11. The signal the oculomotor system actually experiences is a low-pass filtered version of this signal. If we let $C(s)$ = control signal, $F(s)$ = active-state tension, and $H(s)$ = low-pass filter, then

$$F(s) = C(s)H(s) = \frac{C(s)}{(s\tau + 1)}$$

where τ is the low-pass filter time constant.

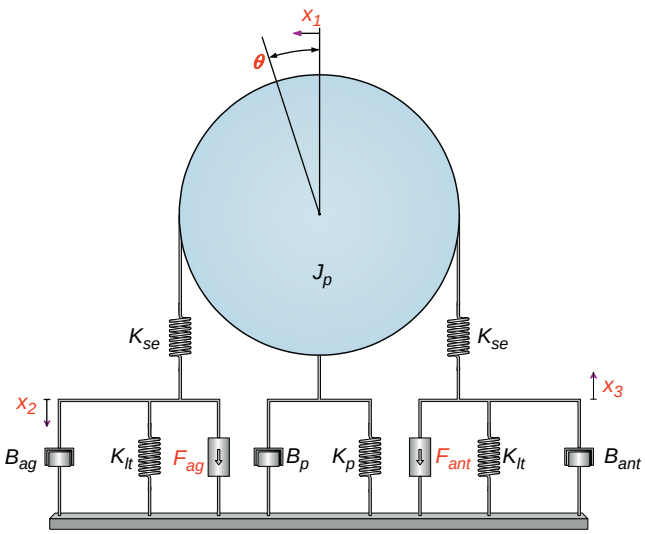


FIGURE 13.28 This model describes the two rectus muscles (agonist (ag) and antagonist (ant)), connected to the eyeball. θ represents the angle that the eyeball is rotated, and x_1 represents the length of arc rotated.

The agonist time constant reported by Bahill and coworkers [4] is a function of motoneuronal firing frequency; the higher the rate, the shorter the time constant. As explained, this is because large saccadic eye movements utilize fast muscle fibers and smaller saccadic eye movements utilize slow muscle fibers. There are two muscles involved with a horizontal eye movement: the agonist and the antagonist muscles. The agonist muscle forcibly contracts and moves the eyeball (fovea) to the target location. The antagonist muscle is completely inhibited during the pulse phase of the trajectory. Keep in mind that muscles are always under tension (tonic state at primary position) to avoid slack.

The control signal from the CNS to each muscle is a series of pulses or spikes due to the action potentials of the motoneurons, as illustrated in Figure 13.10. This diagram illustrates a typical pattern observed during a series of fast eye movements in both horizontal directions. Notice that during a movement in the "on" direction (lateral), the rate of firing increases greatly; in the "off" direction (medial), the firing rate is zero. Also, notice that the burst firing starts approximately 5 ms before the saccade begins and that the longer the neurons fire, the larger the saccade. There is a large, nonconstant time delay between the time a target moves and when the eye actually starts to move. This is due to the *CNS* system calculating the forces necessary to bring the fovea to the target location. This movement is *ballistic* (not guided) to the extent that there are no known stretch receptors indicating muscle activity.

13.6 THE 1984 LINEAR RECIPROCAL INNERVATION SACCADIC EYE MOVEMENT MODEL

Based on physiological evidence, Bahill and coworkers in 1980 presented a linear fourth-order model of the horizontal oculomotor plant that provides an excellent match between model predictions and horizontal eye movement data. This model eliminates the differences seen between velocity and acceleration predictions of the Westheimer and Robinson models, and the data. For ease in this presentation, the 1984 modification of this model by Enderle and coworkers is used. (A more thorough presentation of this model is given in Enderle, 2010a.)

In the previous analysis, B_{ag} and B_{ant} are nonlinear functions of velocity. We can linearize these functions by approximating the force-velocity family of curves with straight line segments as illustrated in Figure 13.29. Antagonist activity is typically at the 5 percent level, and agonist activity is at the 100 percent level. Thus, we can assume that B_{ag} and B_{ant} are constants with different values, since the slopes are different in the linearization.

Using the linearized viscosity elements in our model of the eye movement system, we write the linear differential equation that describes saccades as a function of θ. The updated model is shown in Figure 13.30. The material presented here is based on the work published by Bahill and coworkers [4], and Enderle and coworkers [20].

To begin the analysis, we first draw the free body diagrams and write the node equations as shown in Figure 13.31.

Node 1: $rK_{se}(x_2 - x_1) - rK_{se}(x_4 - x_3) = J_p\,\ddot{\theta} + B_p\dot{\theta} + K_p\theta$
Node 2: $F_{ag} = B_{ag}\dot{x}_2 + K_{se}(x_2 - x_1) + K_{lt}x_2$
Node 3: $K_{se}(x_4 - x_3) = F_{ant} + K_{lt}x_3 + B_{ant}\dot{x}_3$

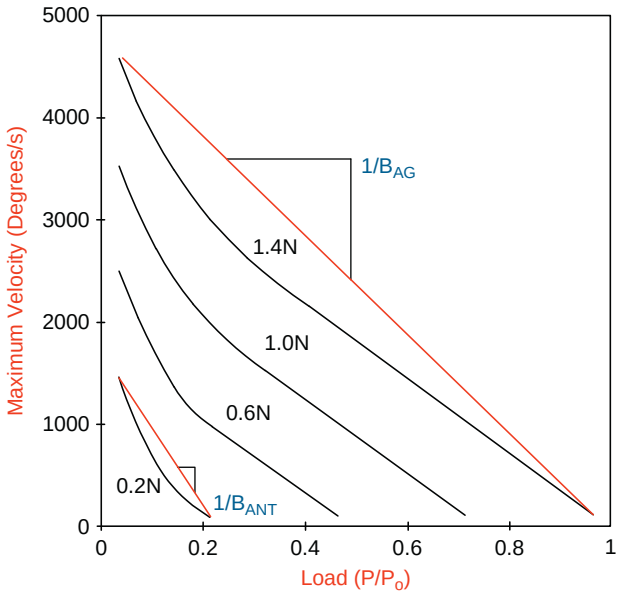


FIGURE 13.29 Linearization of nonlinear force-velocity curves.

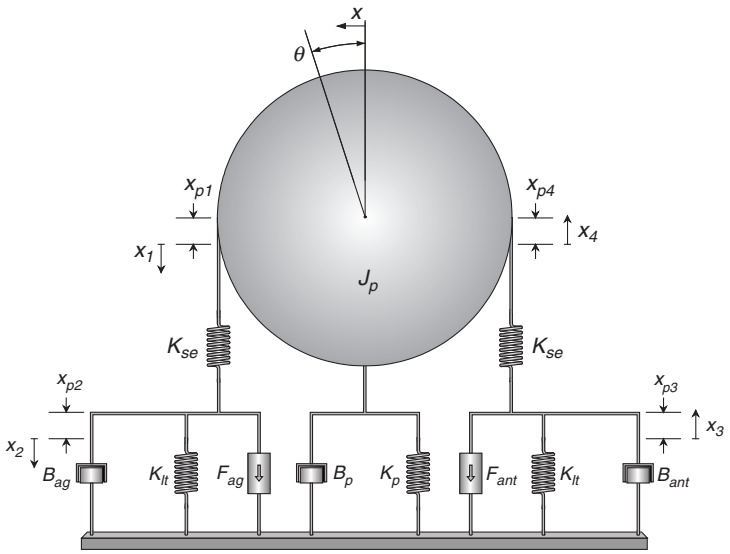


FIGURE 13.30 Linear eye movement model.

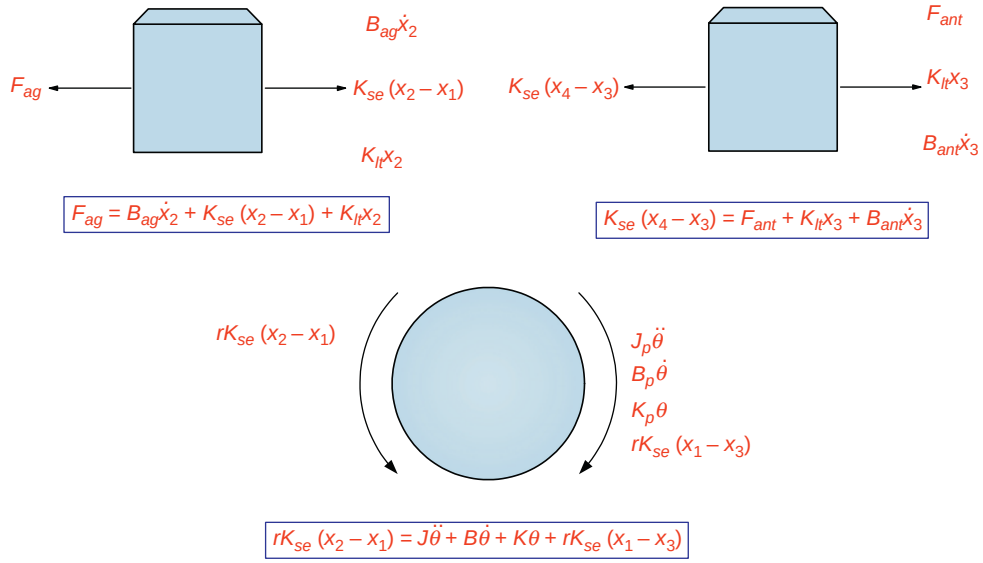


FIGURE 13.31 Free body diagrams for the system in Figure 13.30.

To solve for θ, we use the Laplace transform analysis (details omitted here; see Enderle, 1984, or Enderle, 2010a).

$$\delta\left(K_{st}\left(F_{ag} - F_{ant}\right) + B_{ant}\dot{F}_{ag} - B_{ag}\dot{F}_{ant}\right) = \dddot{\theta} + C_3\ddot{\theta} + C_2\ddot{\theta} + C_1\dot{\theta} + C_0\theta \tag{13.35}$$

where

$$J = \frac{J_p}{r} \times 5.2087 \times 10^3,\ B = \frac{B_p}{r} \times 5.2087 \times 10^3,\ K = \frac{K_p}{r} \times 5.2087 \times 10^3,$$

$$\theta = \frac{x}{r} \times \frac{180}{\pi} = 5.2087 \times 10^3 x, \text{ and } \delta = \frac{57.296K_{se}}{rJB_{ant}B_{ag}}$$

$$C_3 = \frac{JK_{st}\left(B_{ag} + B_{ant}\right) + BB_{ant}B_{ag}}{JB_{ant}B_{ag}}$$

$$C_2 = \frac{JK_{st}^2 + BK_{st}\left(B_{ag} + B_{ant}\right) + B_{ag}B_{ant}(K + 2K_{se})}{JB_{ant}B_{ag}}$$

$$C_1 = \frac{BK_{st}^2 + \left(B_{ag} + B_{ant}\right)\left(KK_{st} + 2K_{se}K_{st} - K_{se}^2\right)}{JB_{ant}B_{ag}}$$

$$C_0 = \frac{KK_{st}^2 + 2K_{se}K_{st}K_{lt}}{JB_{ant}B_{ag}}$$

The agonist and antagonist active-state tensions are given by the following low-pass filtered pulse-step waveforms and are illustrated along with their corresponding neurological control signals in Figure 13.11.

$$\begin{aligned} F_{ag} &= F_{g0}u(-t) + \left(F_p + \left(F_{g0} - F_p\right)e^{\frac{-t}{\tau_{ac}}}\right)\left(u(t) - u(t-t_1)\right) \\ &+ \left(F_{gs} + \left(F_p + \left(F_{g0} - F_p\right)\right)e^{\frac{-t_1}{\tau_{ac}}} - F_{gs}\right)e^{\frac{-(t-t_1)}{\tau_{de}}}u(t-t_1) \\ F_{ant} &= F_{t0}u(-t) + F_{t0}e^{\frac{-t}{\tau_{de}}}\left(u(t) - u(t-t_1)\right) \\ &+ \left(F_{ts} + \left(F_{t0}e^{\frac{-t_1}{\tau_{de}}} - F_{ts}\right)e^{\frac{-(t-t_1)}{\tau_{ac}}}\right)u(t-t_1) \end{aligned} \tag{13.36}$$

where

F_{g0} = initial magnitude of the agonist active-state tension
F_p = pulse magnitude of the agonist active-state tension
F_{gs} = step magnitude of the agonist active-state tension
F_{t0} = initial magnitude of the antagonist active-state tension
F_{ts} = step magnitude of the antagonist active-state tension
τ_{ac} = activation time constant
τ_{de} = deactivation time constant
t_1 = duration of the agonist pulse active-state tension

Final parameter estimates for the saccadic eye movement model are found using the system identification technique, a frequency response method. The oculomotor system operates in an open-loop mode while executing a saccade. After completing the saccade, the central nervous system operates in a closed-loop mode and compares eye and target position [5]. Figure 13.32 presents a block diagram illustrating the open-loop, closed-loop operation of the oculomotor system, with the feedback element H operating only during discrete time intervals when a saccade is not being executed. The system identification is used to estimate the parameters and the input during the open-loop mode of the saccade.

The steady-state active-state tensions are given by

$$\begin{aligned} F_{ag} &= \begin{cases} 0.14 + 0.0185\theta & \text{N for } \theta < 14.23^\circ \\ 0.0283\theta & \text{N for } \theta \geq 14.23^\circ \end{cases} \\ F_{ant} &= \begin{cases} 0.14 - 0.0098\theta & \text{N for } \theta < 14.23^\circ \\ 0 & \text{N for } \theta \geq 14.23^\circ \end{cases} \end{aligned} \tag{13.37}$$

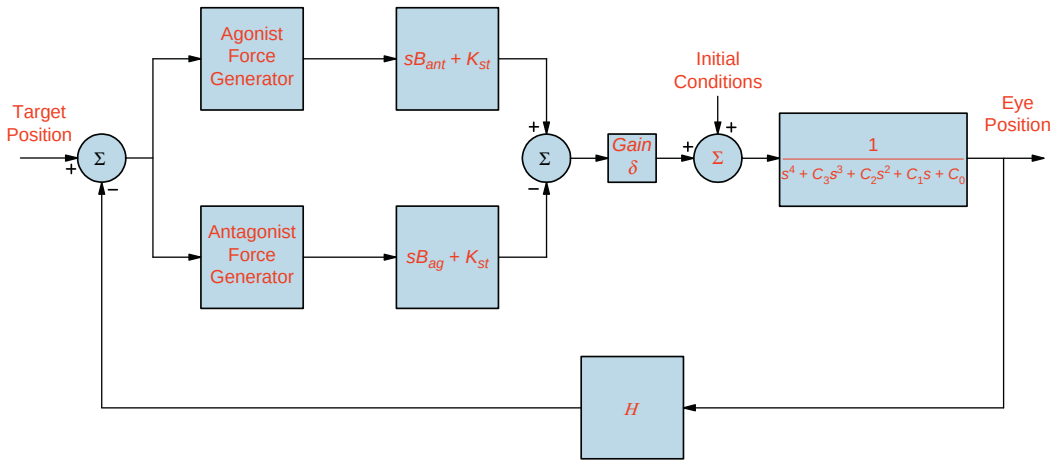


FIGURE 13.32 Block diagram of the modified linear homeomorphic eye movement model in Eq. (13.35).

A set of parameter estimates for the oculomotor plant from the system identification technique are

$$K_{se} = 125\ \text{Nm}^{-1}$$
$$K_{lt} = 32\ \text{Nm}^{-1}$$
$$K = 66.4\ \text{Nm}^{-1}$$
$$B = 3.1\ \text{Nsm}^{-1}$$
$$J = 2.2 \times 10^{-3}\ \text{Ns}^2 m^{-1}$$
$$B_{ag} = 3.4\ \text{Nsm}^{-1}$$
$$B_{ant} = 1.2\ \text{Nsm}^{-1}$$
$$\tau_{ac} = 0.009\ \text{s}$$
$$\tau_{de} = 0.0054\ \text{s}$$
$$\delta = 5.80288 \times 10^5$$

The eigenvalues for the oculomotor plant using the parameter values above are -15, -66, -173, and $-1{,}293$.

Figure 13.33 shows estimation routine results for a 15° target displacement. The accuracy of these results is typical for all target displacements with all subjects tested, except for saccades with glissadic or dynamic overshoot (saccades with glissadic or dynamic overshoot are discussed later in this chapter). Figures 13.34 and 13.35 further illustrate the accuracy of the system identification technique parameter estimation routine by the close agreement of the velocity and acceleration estimates with the two-point central differences estimates.

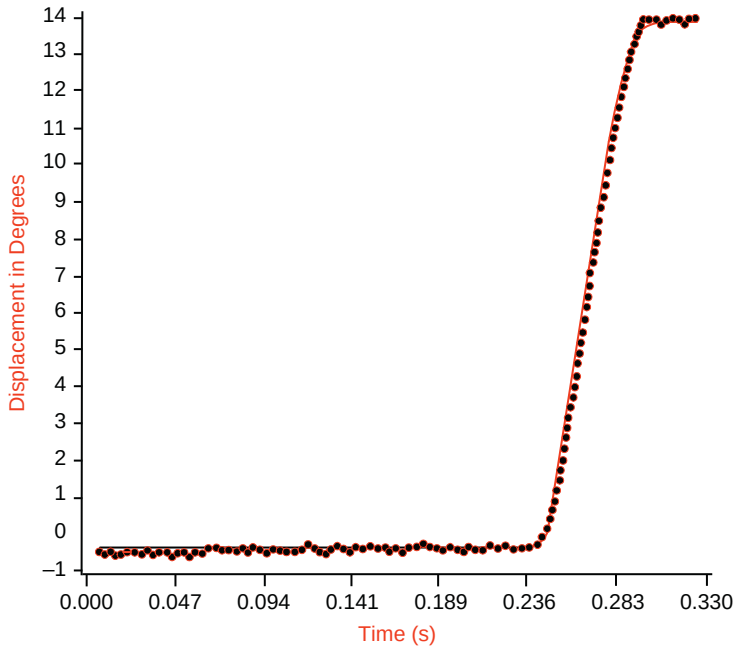


FIGURE 13.33 Response for a 15-degree saccadic eye movement. The solid line shows the predictions of the saccadic eye movement model, with final parameter estimates computed using the system identification technique. Dots are the data.

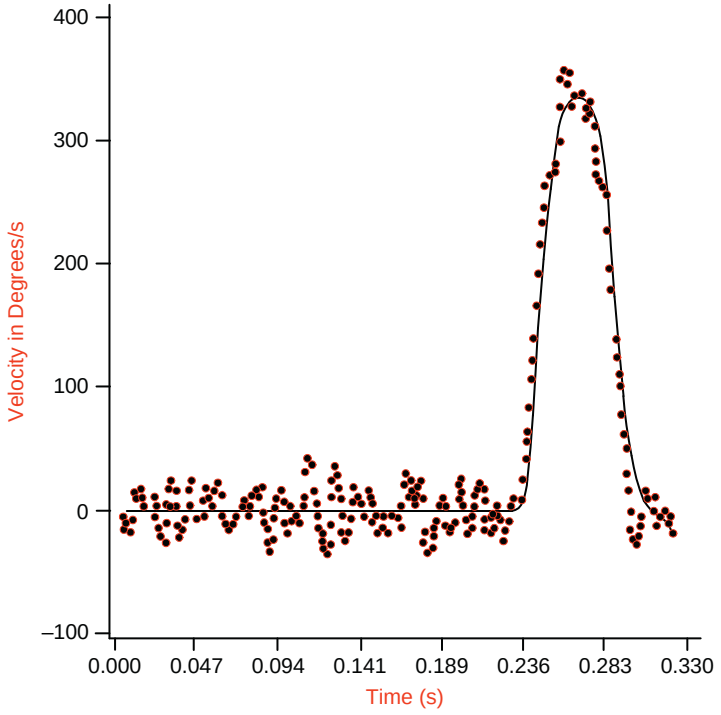


FIGURE 13.34 A plot of the velocity estimates (dots) and velocity simulation results (solid line) from the modified linear homeomorphic model.

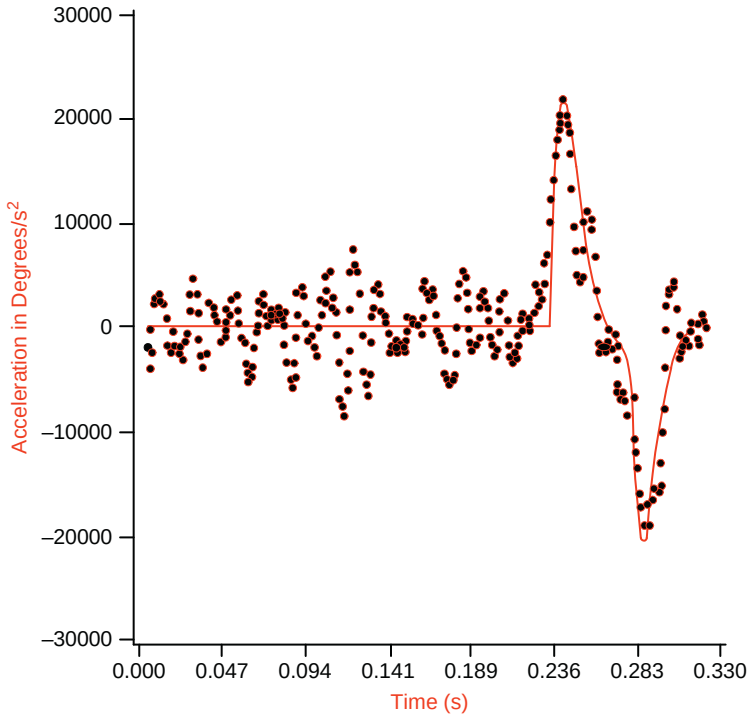


FIGURE 13.35 A plot of the acceleration estimates (dots) and acceleration simulation results (solid line) from the modified linear homeomorphic model.

EXAMPLE PROBLEM 13.6

Using the oculomotor plant model described with Eq. (13.35) and the following parameters, simulate a 20° saccade.

$K_{se} = 125\ \text{Nm}^{-1}, K_{lt} = 32\ \text{Nm}^{-1}, K = 66.4\ \text{Nm}^{-1}, B = 3.1\ \text{Nsm}^{-1}, J = 2.2 \times 10^{-3}\ \text{Ns}^2\text{m}^{-1}$
$B_{ag} = 3.4\ \text{Nsm}^{-1}, B_{ant} = 1.2\ \text{Nsm}^{-1}, \tau_{ac} = 0.009\ \text{s}, \tau_{de} = 0.0054\ \text{s}, \delta = 5.80288 \times 10^5,$
$F_p = 1.3\ \text{N}, t_1 = 31\ \text{ms}, \text{Latent Period} = 150\ \text{ms}.$

Plot the neural inputs, agonist and antagonist active-state tension, position, velocity, and acceleration versus time.

Solution

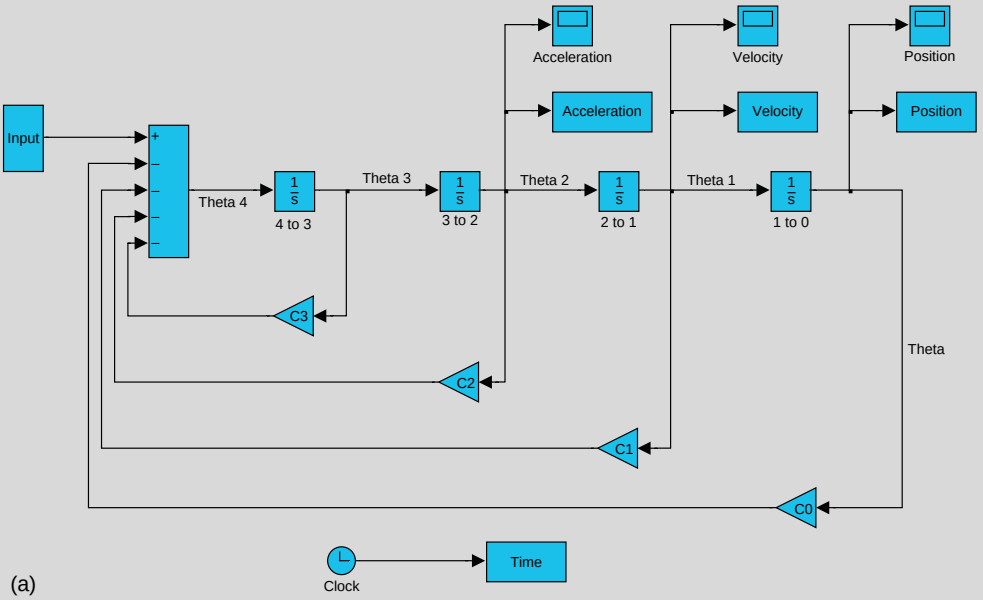
Using the given parameter values, the following m-file provides the parameter values for the SIMULINK program.

```
Fp=1.3
t1=.031
theta=20
```

```
theta0=0
TDE=.0054
TAC=.009
tdeinv=1/TDE
tacinv=1/TAC
KSE=125
KLT=32
K=66.4
B=3.1
J=2.2 * 10^-3
BAG=3.4
BANT=1.2
DELTA=5.80261*10^5
KST=KLT+KSE
C0=((K*KST^2)+(2*KSE*KST*KLT))/(J*BANT*BAG)
C1=((B*KST^2)+(BAG+BANT)*((K*KST)+(2*KSE*KST)-KSE^2))/
(J*BANT*BAG)
C2=((J*KST^2)+((B*KST)*(BAG+BANT))+((BANT*BAG)*(K+(2*KSE))))/
(J*BANT*BAG)
C3=(((J*KST)*(BAG+BANT))+(B*BANT*BAG))/(J*BANT*BAG)
if theta0 < 14.23
   Fag0=0.14+0.0185*theta0
else
   Fag0=0.0283*theta0
end
if theta0 < 14.23
   Fant0=0.14-0.0098*theta0
else
   Fant0=0
end
if theta < 14.23
   Fagss=0.14+0.0185*theta
else
    Fagss=0.0283*theta
end
if theta < 14.23
    Fantss=0.14-0.0098*theta
else
   Fantss=0
end
latent=.15
sstart=latent+t1
agstep=Fp-Fagss
```

Continued

Figure 13.36 shows the SIMULINK program. The main program is shown in Figure 13.36 (a), based on Eq. (13.35). The input to the system is shown in Figure 13.36 (b), and the agonist and antagonist inputs are shown in Figures 13.36(c) and (d). Figure 13.37 shows the plots of position, velocity, acceleration, agonist neural input and active-state tension, and antagonist neural input and active-state tension.



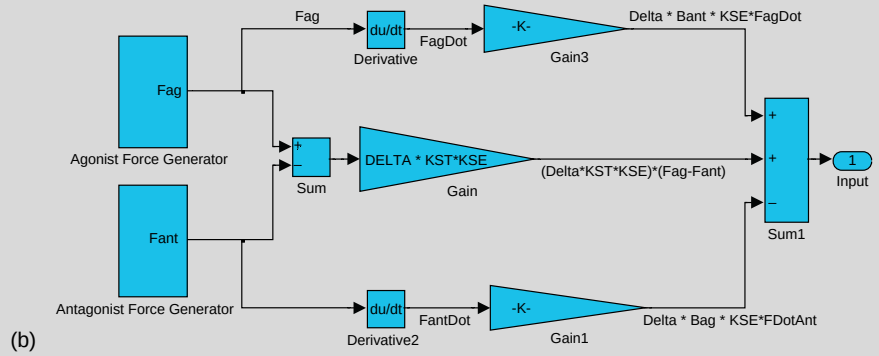
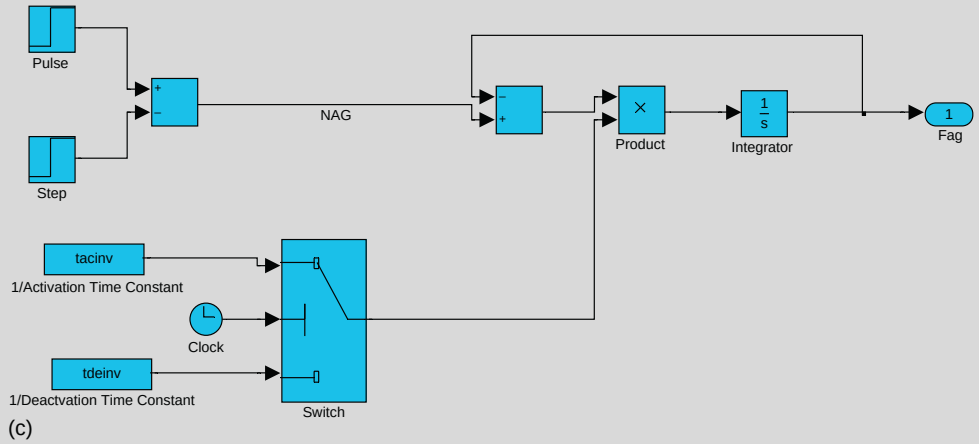
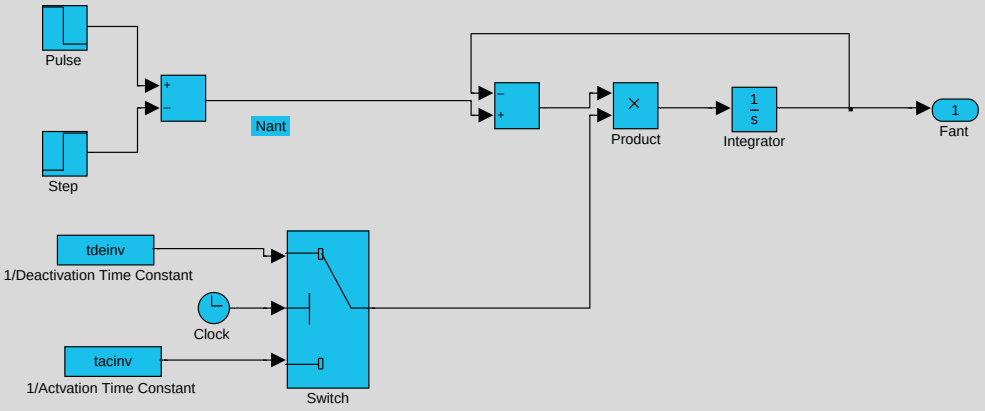


FIGURE 13.36 SIMULINK program for Example Problem 13.6.

Continued



(c)



(d)

FIGURE 13.36, cont'd

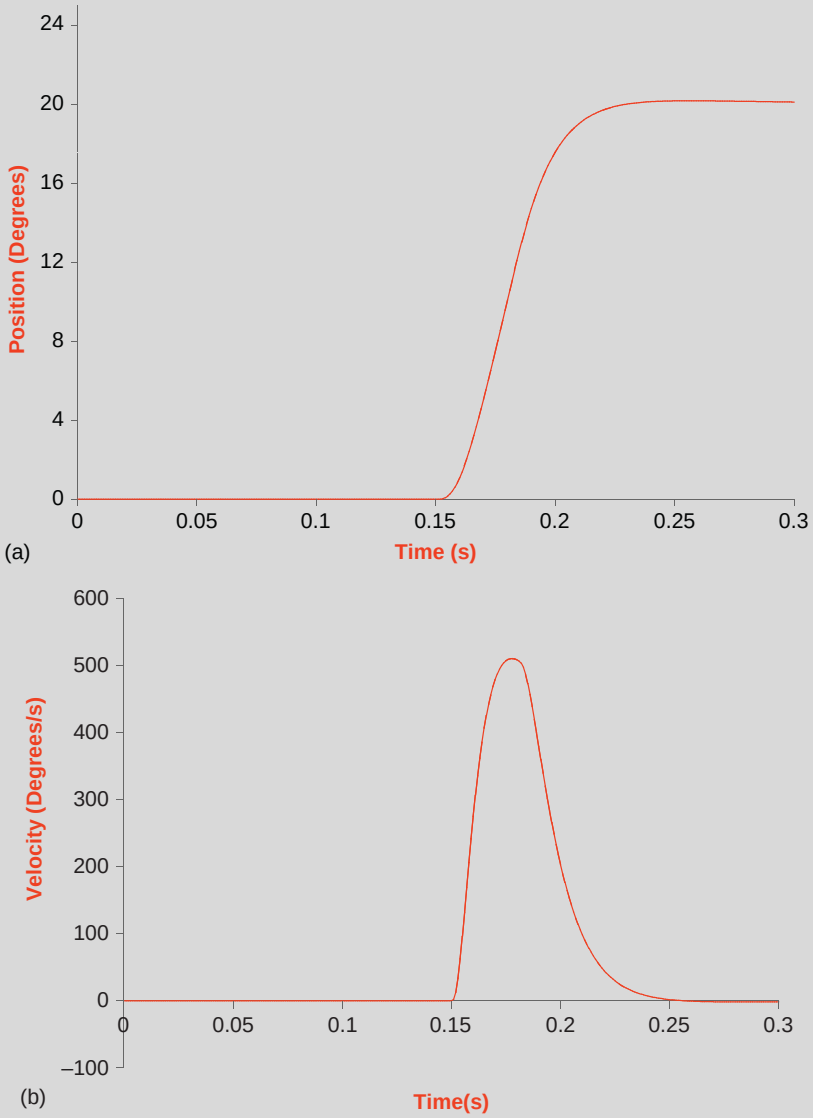
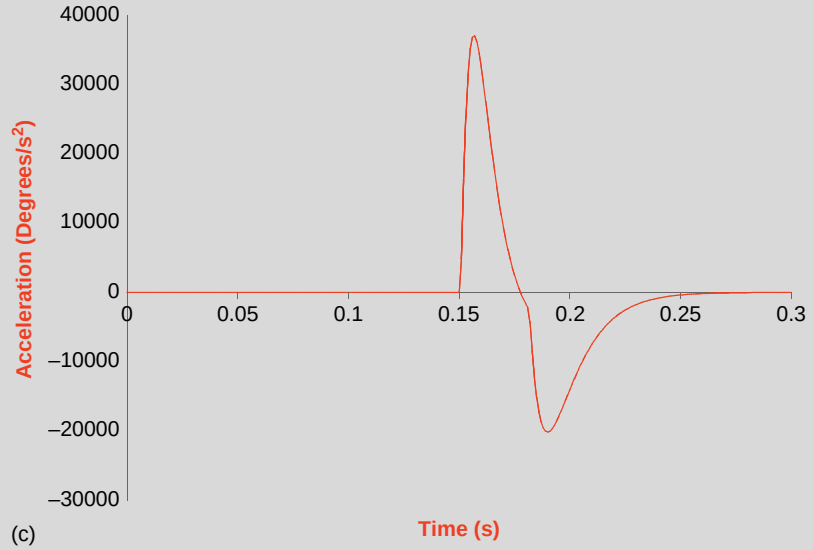
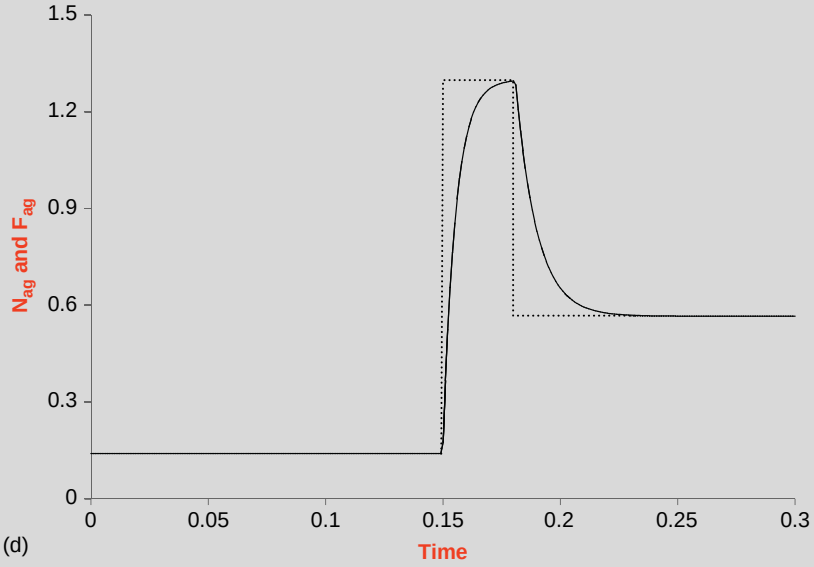


FIGURE 13.37 Plots of position, velocity, acceleration, agonist neural input (dotted line) and active-state tension (solid line), and antagonist neural input and active-state tension for Example Problem 13.6.

Continued



(c)



(d)

FIGURE 13.37, cont'd

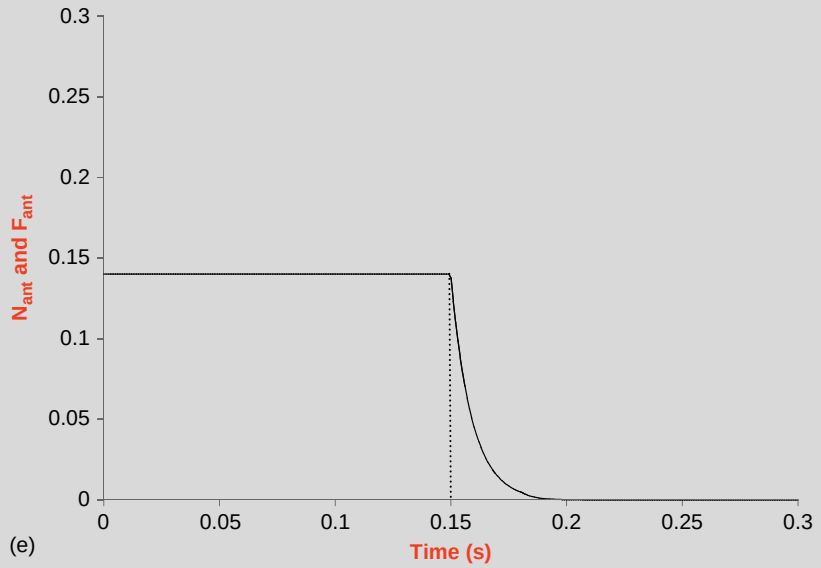


FIGURE 13.37, cont'd

13.7 THE 1995 LINEAR HOMEOMORPHIC SACCADIC EYE MOVEMENT MODEL

In the previous section, we presented a linear model of the oculomotor plant developed by linearizing the force-velocity curve. We then derived a linear differential equation to describe the system. Here we reexamine the static and dynamic properties of muscle in the development of a new linear model of oculomotor muscle. With the updated linear model of oculomotor muscle, the model of the oculomotor system will also be updated.

13.7.1 The Linear Muscle Model

The updated linear model for oculomotor muscle is shown in Figure 13.38. Each of the elements in the model is linear and supported with physiological evidence. The muscle is modeled as a parallel combination of viscosity B_2 and series elasticity K_{se}, connected to the parallel combination of active-state tension generator F, viscosity element B_1, and length tension elastic element K_{lt}. Variables x_1 and x_2 describe the displacement from the equilibrium for the stiffness elements in the muscle model. The only structural difference between this model and the previous oculomotor muscle model is the addition of viscous element B_2 and the removal of passive elasticity K_{pe}. As will be described, the viscous element B_2 is vitally important to describe the nonlinear force-velocity characteristics of the muscle, and the elastic element K_{pe} is unnecessary.

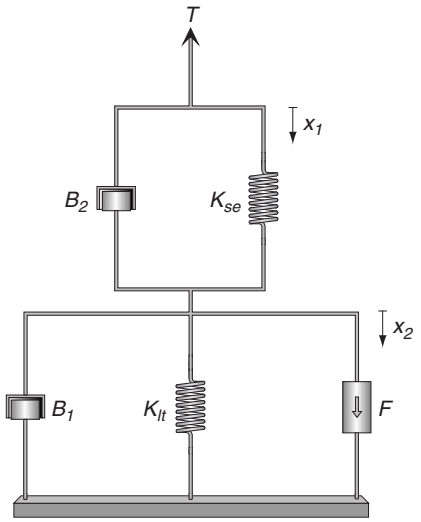


FIGURE 13.38 An updated linear muscle model consisting of an active-state tension generator F in parallel with a length-tension elastic element K_{lt} and viscous element B_1, connected to a series elastic element K_{se} in parallel with a viscous element B_2. Upon stimulation of the active-state tension generator F, a tension T is exerted by the muscle.

The need for two elastic elements in the linear oculomotor muscle model is supported through physiological evidence. As described previously, the use and value of the series elasticity K_{se} were determined from the isotonic-isometric quick release experiment by Collins. Length-tension elasticity K_{lt} was estimated in a slightly different fashion than before from the slope of the length-tension curve. Support for the two linear viscous elements is based on the isotonic experiment and estimated from simulation results presented in this chapter.

13.7.2 The Length-Tension Curve

The basis for assuming nonlinear elasticity is the nonlinear length-tension relation for excited and unexcited muscle for tensions below 10 g, as shown in Figure 13.14. Using a miniature "C" gauge force transducer, Collins in 1975 measured muscle tension in vivo at the muscle tendon during unrestrained human eye movements. Data shown in Figure 13.14 were recorded from the rectus muscle of the left eye by measuring the isometric tensions at different muscle lengths, ranging from eye positions of $-45°$ to $45°$, and different levels of innervation, established by directing the subject to look at the corresponding targets with the unhampered right eye from $-45°$ temporal (T) to $45°$ nasal (N). The change in eye position during this experiment corresponds to a change in muscle length of approximately 18 mm. Collins described the length-tension curves as "straight, parallel lines above about 10 g. Below the 10 g level, the oculorotary muscles begin to go slack." He also reported that the normal range of tensions for the rectus muscle during all eye movements never falls below 10 g into the slack region when the in vivo force transducer is used.

In developing a muscle model for use in the oculomotor system, it is imperative that the model accurately exhibits the static characteristics of rectus eye muscle within the normal range of operation. Thus, any oculomotor muscle model must have length-tension characteristics consisting of straight, parallel lines above 10 g tension. Since oculomotor muscles do not operate below 10 g, it is not unimportant that the linear behavior of the model does not match this nonlinear portion in the length-tension curves observed in the data as was done in the development of the muscle model earlier. As demonstrated in this section, by concentrating on the operational region of the oculomotor muscles, accurate length-tension curves are obtained from the muscle model using just series elastic and length tension elastic elements, even when active-state tension is zero. Thus, there is no need to include a passive elastic element in the muscle model as previously required.

Since the rectus eye muscle is not in equilibrium at primary position (looking straight ahead, 0°) within the oculomotor system, it is necessary to define and account for the equilibrium position of the muscle. Equilibrium denotes the unstretched length of the muscle when the tension is zero, with zero input. It is assumed that the active-state tension is zero on the 45° T length-tension curve. Typically, the equilibrium position for rectus eye muscle is found from within the slack region, where the 45° T length-tension curve intersects the horizontal axis. Note that this intersection point was not shown in the data collected by Collins (see Figure 13.14) but is reported to be approximately 15° (3 mm short of primary position), a value that is typical of those reported in the literature.

Since the muscle does not operate in the slack region during normal eye movements, using an equilibrium point calculated from the operational region of the muscle provides a much more realistic estimate for the muscle. Here, the equilibrium point is defined according to the straight-line approximation to the 45° T length-tension curve above the slack region. The value at the intersection of the straight-line approximation with the horizontal axis gives an equilibrium point of $-19.3°$. By use of the equilibrium point at $-19.3°$, there is no need to include an additional elastic element K_{pe} to account for the passive elasticity associated with unstimulated muscle as others have done.

The tension exerted by the linear muscle model shown in Figure 13.38 is given by

$$T = \frac{K_{se}}{K_{se} + K_{lt}} F - \frac{K_{se} K_{lt}}{K_{se} + K_{lt}} x_1 \tag{13.38}$$

With the slope of the length-tension curve equal to $0.8\ \text{g}/^\circ = 40.86\ \text{N/m}$ in the operating region of the muscle (nonslack region), $K_{se} = 2.5\ \text{g}/^\circ = 125\ \text{N/m}$, and Eq. (13.38) has a slope of

$$\frac{K_{se} K_{lt}}{K_{se} + K_{lt}} \tag{13.39}$$

K_{lt} is evaluated as $1.2\ \text{g}/^\circ = 60.7\ \text{N/m}$.

To estimate the static active-state tension for fixation at the locations detailed in Figure 1 of Collins [7], we use the techniques by Enderle and coworkers [19] by taking Eq. (13.38) to solve for steady-state active-state tensions for each innervation level straight-line approximation, yielding for $\theta > 0°$ (N direction)

$$F = 0.4 + 0.0175\theta\ N \quad \text{for } \theta > 0° (\text{N direction}) \tag{13.40}$$

and

$$F = 0.4 + 0.012\theta\, N \quad \text{for } \theta \leq 0^{\circ}(\text{T direction}) \tag{13.41}$$

where θ is the angle that the eyeball is deviated from the primary position measured in degrees, and $\theta = 5208.7 \times (x_1 - 3.705)$. Note that $5208.7 = \frac{180}{\pi r}$, where r equals the radius of the eyeball with a value of 11 mm.

Figure 13.39 displays a family of static length-tension curves obtained using Eqs. (13.38)–(13.41), which depicts the length-tension experiment. No attempt is made to describe the activity within the slack region, since the rectus eye muscle does not normally operate in that region. The length-tension relationship shown in Figure 13.39 is in excellent agreement with the data shown in Figure 13.14 within the operating region of the muscle.

13.7.3 The Force-Velocity Relationship

The original basis for assuming nonlinear muscle viscosity is that the expected linear relation between external load and maximum velocity was not observed in early experiments by Fenn and Marsh [22]. As they reported, "If the muscle is represented accurately

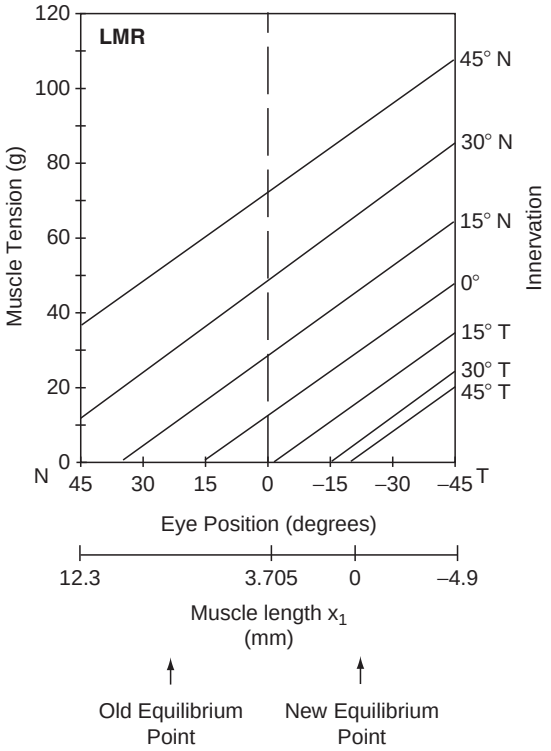


FIGURE 13.39 Length-tension relationship generated using Equations (13.38)–(13.41) derived from the linear muscle model and inputs: $F = 130$ g for 45° N, $F = 94.3$ g for 30° N, $F = 64.9$ g for 15° N, $F = 40.8$ g for 0°, $F = 21.7$ g for 15° T, $F = 5.1$ g for 30° T, and $F = 0$ g for 45° T. These lines were parameterized to match Figure 13.14.

by a viscous elastic system, this force-velocity curve should have been linear, the loss of force being always proportional to the velocity. The slope of the curve would then represent the coefficient of viscosity." Essentially the same experiment was repeated for rectus eye muscle by Close and Luff in 1974 with similar results.

The classic force-velocity experiment was performed to test the viscoelastic model for muscle as described previously in Section 13.5.4. Under these conditions, it was first assumed that the inertial force exerted by the load during isotonic shortening could be ignored. The second assumption was that if mass was not reduced enough by the lever ratio (enough to be ignored), then taking measurements at maximum velocity provided a measurement at a time when acceleration is zero, and, therefore, inertial force equals zero. If these two assumptions are valid, then the experiment would provide data free of the effect of inertial force as the gravity force is varied. Both assumptions are incorrect. The first assumption is wrong, since the inertial force is never minimal (minimal would be zero) and therefore has to be taken into account. The second assumption is wrong, since, given an inertial mass not equal to zero, the maximum velocity depends on the forces that act prior to the time of maximum velocity. The force-velocity relationship is carefully reexamined with the inertial force included in the analysis in this section.

The dynamic characteristics for the linear muscle model are described with a force-velocity curve calculated via the lever system presented in Figure 13.20 and according to the isotonic experiment. For the rigid lever, the displacements x_1 and x_3 are directly proportional to the angle θ_1 and to each other, such that

$$\theta_1 = \frac{x_1}{d_1} = \frac{x_3}{d_3} \tag{13.42}$$

The equation describing the torques acting on the lever is given by

$$Mgd_3 + Md_3^2\ddot{\theta}_1 = d_1K_{se}(x_2 - x_1) + d_1B_2(\dot{x}_2 - \dot{x}_1) \tag{13.43}$$

The equation describing the forces at node 2, inside the muscle, is given by

$$F = K_{lt}x_2 + B_1\dot{x}_2 + B_2(\dot{x}_2 - \dot{x}_1) + K_{se}(x_2 - x_1) \tag{13.44}$$

Equation (13.43) is rewritten by removing θ_1 using Eq. (13.42), hence

$$Mg\frac{d_3}{d_1} + M\left(\frac{d_3}{d_1}\right)^2\ddot{x}_1 = K_{se}(x_2 - x_1) + B_2(\dot{x}_2 - \dot{x}_1) \tag{13.45}$$

Ideally, to calculate the force-velocity curve for the lever system, $x_1(t)$ is found first. Then $\dot{x}_1(t)$ and $\ddot{x}_1(t)$ are found from $x_1(t)$. Finally, the velocity is found from $V_{\max} = \dot{x}_1(T)$, where time T is the time it takes for the muscle to shorten to the stop, according to the experimental conditions of Close and Luff. While this velocity may not be maximum velocity for all data points, the symbol $V_{\max}$ is used to denote the velocities in the force-velocity curve for ease in presentation. Note that this definition of velocity differs from the Fenn and Marsh definition of velocity. Fenn and Marsh denoted maximum velocity as $V_{\max} = \dot{x}_1(T)$, where time T is found when $\ddot{x}_1(T) = 0$.

It should be noted that this is a third-order system and the solution for $x_1(t)$ is not trivial and involves an exponential approximation (for an example of an exponential

approximation solution for V_{max} from a fourth-order model, see the paper by [17], or [14]. It is more expedient, however, to simply simulate a solution for $x_1(t)$ and then find V_{max} as a function of load.

Using a simulation to reproduce the isotonic experiment, elasticities estimated from the length-tension curves as previously described, and data from rectus eye muscle, parameter values for the viscous elements in the muscle model are found as $B_1 = 2\,\text{Nsm}^{-1}$ and $B_2 = 0.5\,\text{Nsm}^{-1}$ as demonstrated by Enderle and coworkers in 1991. The viscous element B_1 is estimated from the time constant from the isotonic time course. The viscous element B_2 is calculated by trial and error so the simulated force-velocity curve matches the experimental force-velocity curve.

Figure 13.40 shows the force-velocity curves using the model described in Eq. (13.45) (with triangles), plotted along with an empirical fit to the data (solid line). It is clear that the force-velocity curve for the linear muscle model is hardly a straight line and that this curve fits the data well.

The muscle lever model described by Eqs. (13.43) and (13.44) is a third-order linear system and is characterized by three poles. Dependent on the values of the parameters, the eigenvalues (or poles) consist of all real poles or a real and a pair of complex conjugate poles. A real pole is the dominant eigenvalue of the system. Through a sensitivity analysis, viscous element B_1 is the parameter that has the greatest effect on the dominant eigenvalue for this system, while viscous element B_2 has very little effect on the dominant eigenvalue. Thus, viscous element B_1 is estimated so the dominant time constant of the lever system model (approximately $\frac{B_1}{K_{lt}}$ when $B_1 > B_2$) matches the time constant from the isotonic experimental data. For rectus eye muscle data, the duration of the isotonic experiment is

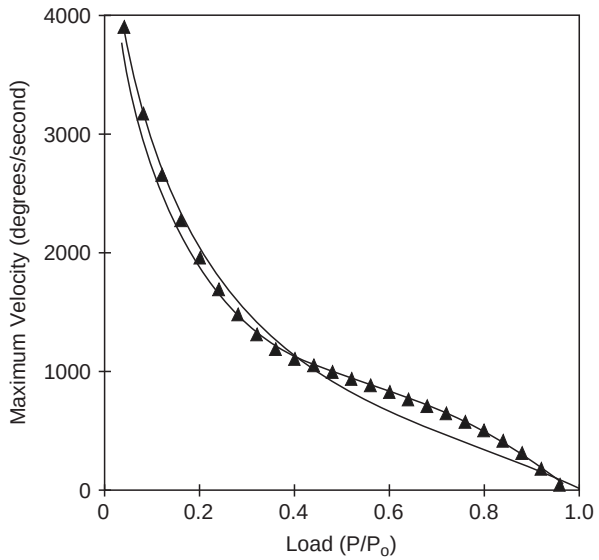


FIGURE 13.40 Force-velocity curve derived from simulation studies with the linear muscle model with an end stop. Shown with triangles indicating simulation calculation points and an empirical fit to the force-velocity data (solid line), as described by Close and Luff [6]. *Adapted from Enderle et al. [19].*

approximately 100 ms. A value for $B_1 = 2\ \text{Nsm}^{-1}$ yields a simulated isotonic response with approximately the same duration. For skeletal muscle data, the duration of the isotonic experiment is approximately 400 ms, and a value for $B_1 = 600\ \text{Nsm}^{-1}$ yields a simulated isotonic response with approximately the same duration. It is known that fast and slow muscle have differently shaped force velocity curves and that the fast muscle force-velocity curve data has less curvature. Interestingly, the changes in the parameter values for B_1 as suggested here give differently shaped force velocity curves consistent with fast (rectus eye) and slow (skeletal) muscle.

The parameter value for viscous element B_2 is selected by trial and error so the shape of the simulation force-velocity curve matches the data. As the value for B_2 is decreased from $0.5\ \text{Nsm}^{-1}$, the shape of the force velocity curves changes to a more linear shaped function. Moreover, if the value of B_2 falls below approximately $0.3\ \text{Nsm}^{-1}$, strong oscillations appear in the simulations of the isotonic experiment, which are not present in the data. Thus, the viscous element B_2 is an essential component in the muscle model. Without it, the shape of the force-velocity curve is linear and the time course of the isotonic experiment does not match the characteristics of the data.

Varying the parameter values of the lever muscle model changes the eigenvalues of the system. For instance, with $M = 0.5$ kg, the system's nominal eigenvalues (as defined with the parameter values previously specified) are a real pole at -30.71 and a pair of complex conjugate poles at $-283.9 \pm j221.2$. If the value of B_2 is increased, three real eigenvalues describe the system. If the value of B_2 is decreased, a real pole and a pair of complex conjugate poles continue to describe the system. Changing the value of B_1 does not change the eigenvalue composition, but it does significantly change the value of the dominant eigenvalue from -292 with $B_1 = 0.1$ to -10 with $B_1 = 6$.

13.7.4 The 1995 Linear Homeomorphic Saccadic Eye Movement Model

The linear model of the oculomotor plant presented in Section 13.6 is based on a nonlinear oculomotor plant model by Hsu and coworkers [33] using a linearization of the force-velocity relationship and elasticity curves. Using the linear model of muscle described earlier in this section, it is possible to avoid the linearization and to derive a truer linear homeomorphic saccadic eye movement model.

The linear muscle model described in this section has the static and dynamic properties of rectus eye muscle, a model without any nonlinear elements. As presented, the model has a nonlinear force-velocity relationship that matches eye muscle data using linear viscous elements, and the length tension characteristics are also in good agreement with eye muscle data within the operating range of the muscle. Some additional advantages of the linear muscle model are that a passive elasticity is not necessary if the equilibrium point $x_e = -19.3°$, rather than 15°, and muscle viscosity is a constant that does not depend on the innervation stimulus level.

Figure 13.41 shows the mechanical components of the updated oculomotor plant for horizontal eye movements, the lateral and medial rectus muscle, and the eyeball. The agonist muscle is modeled as a parallel combination of viscosity B_2 and series elasticity K_{se}, connected to the parallel combination of active-state tension generator F_{ag}, viscosity element B_1, and length tension elastic element K_{lt}. For simplicity, agonist viscosity is set

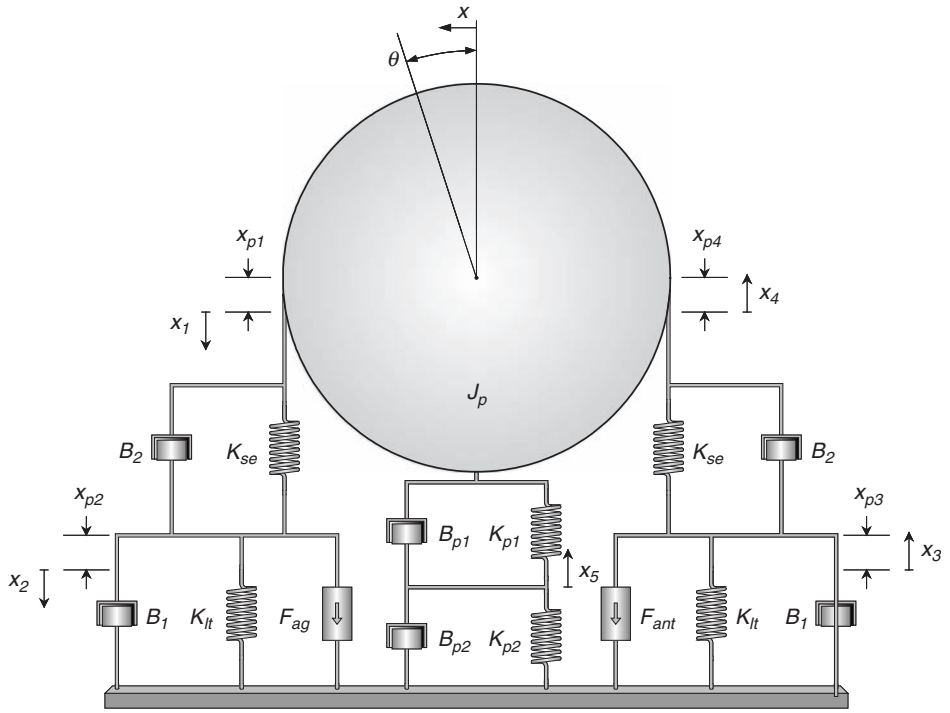


FIGURE 13.41 The mechanical components of the updated oculomotor plant. The muscles are shown to be extended from equilibrium, a position of rest, at the primary position (looking straight ahead), consistent with physiological evidence. The average length of the rectus muscle at the primary position is approximately 40 mm and at the equilibrium position is approximately 37 mm. θ is the angle the eyeball is deviated from the primary position, and variable x is the length of arc traversed. When the eye is at the primary position, both θ and x are equal to zero. Variables x_1 through x_4 are the displacements from equilibrium for the stiffness elements in each muscle, and θ_5 is the rotational displacement for passive orbital tissues. Values x_{p1} through x_{p4} are the displacements from equilibrium for each of the variables x_1 through x_4 at the primary position. The total extension of the muscle from equilibrium at the primary position is x_{p1} plus x_{p2} or x_{p3} plus x_{p4}, which equals approximately 3 mm. It is assumed that the lateral and medial rectus muscles are identical, such that x_{p1} equals x_{p4} and x_{p3} equals x_{p2}. The radius of the eyeball is r.

equal to antagonist viscosity. The antagonist muscle is similarly modeled with a suitable change in active-state tension to F_{ant}. Each of the elements defined in the oculomotor plant is ideal and linear.

The eyeball is modeled as a sphere with moment of inertia J_p, connected to a pair of viscoelastic elements connected in series. The update of the eyeball model is based on observations by Robinson, presented in 1981, and the following discussion. In the model of the oculomotor plant described in Section 3.6, passive elasticity K_{pe} was combined with the passive elastic orbital tissues. In the new linear model muscle presented in this chapter, the elastic element K_{pe} is no longer included in the muscle model. Thus, the passive orbital tissue elasticity needs to be updated due to the elimination of K_{pe} and the observations by Robinson. As reported by Robinson in 1981, "When the human eye, with horizontal recti detached, is displaced and suddenly released, it returns rapidly about 61 percent of the

way with a time constant of about 0.02 sec, and then creeps the rest of the way with a time constant of about 1 sec." As suggested according to this observation, there are at least two viscoelastic elements. Here it is proposed that these two viscoelastic elements replace the single viscoelastic element of the previous oculomotor plant. Connected to the sphere, are $B_{p1}||K_{p1}$ connected in series to $B_{p2}||K_{p2}$. As reported by Robinson, total orbital elasticity is equal to 12.8×10^{-7} g/° (scaled for this model). Thus, with the time constants previously described, the orbital viscoelastic elements are evaluated as $K_{p1} = 1.28 \times 10^{-6}$ g/°, $K_{p2} = 1.98 \times 10^{-6}$ g/°, $B_{p1} = 2.56 \times 10^{-8}$ gs/°, and $B_{p2} = 1.98 \times 10^{-6}$ gs/°. For modeling purposes, θ_5 is the variable associated with the change from equilibrium for these two pairs of viscoelastic elements. Both θ and θ_5 are removed from the analysis for simplicity using the substitution $\theta = 57.296\,\dfrac{x}{r}$ and $\theta_5 = 57.296\,\dfrac{x_5}{r}$.

By summing the forces acting at junctions 2 and 3, and the torques acting on the eyeball and junction 5, a set of four equations is written to describe the oculomotor plant:

$$\begin{aligned} F_{ag} &= K_{lt}x_2 + B_1\dot{x}_2 + K_{se}(x_2 - x_1) + B_2(\dot{x}_2 - \dot{x}_1) \\ B_2(\dot{x}_4 - \dot{x}_3) + K_{se}(x_4 - x_3) &= F_{ant} + K_{lt}x_3 + B_1\dot{x}_3 \\ B_2(\dot{x}_2 + \dot{x}_3 - \dot{x}_1 - \dot{x}_4) + K_{se}(x_2 + x_3 - x_1 - x_4) &= J\ddot{x} + B_3(\dot{x} - \dot{x}_5) + K_1(x - x_5) \\ K_1(x - x_5) + B_3(\dot{x} - \dot{x}_5) &= B_4\dot{x}_5 + K_2x_5 \end{aligned} \tag{13.46}$$

where

$$J = \frac{57.296}{r^2}J_p,\ B_3 = \frac{57.296}{r^2}B_{p1},\ B_4 = \frac{57.296}{r^2}B_{p2},\ K_1 = \frac{57.296}{r^2}K_{p1},\ K_2 = \frac{57.296}{r^2}K_{p2}$$

Using Laplace variable analysis about an operating point yields

$$\begin{aligned} K_{se}K_{12}\left(F_{ag} - F_{ant}\right) + (K_{se}B_{34} + B_2K_{12})\left(\dot{F}_{ag} - \dot{F}_{ant}\right) + B_2B_{34}\left(\ddot{F}_{ag} - \ddot{F}_{ant}\right) \\ = C_4\dddot{x} + C_3\ddot{x} + C_2\ddot{x} + C_1\dot{x} + C_0x \end{aligned} \tag{13.47}$$

where

$$\begin{aligned} B_{12} &= B_1 + B_2,\ B_{34} = B_3 + B_4,\ K_{12} = K_1 + K_2,\ K_{st} = K_{se} + K_{lt} \\ C_4 &= JB_{12}B_{34} \\ C_3 &= B_3B_4B_{12} + 2B_1B_2B_{34} + JB_{34}K_{st} + JB_{12}K_{12} \\ C_2 &= 2B_1B_{34}K_{se} + JK_{st}K_{12} + B_3B_{34}K_{st} + B_3B_{12}K_{12} + K_1B_{12}B_{34} - B_3^2K_{st} - 2K_1B_3B_{12} \\ &\quad + 2B_2K_{lt}B_{34} + 2B_1K_{12}B_2 \\ C_1 &= 2K_{lt}B_{34}K_{se} + 2B_1K_{12}K_{se} + B_3K_{st}K_2 + K_1B_{34}K_{st} + K_1B_{12}K_{12} - K_{st}K_1B_3 \\ &\quad - K_1^2B_{12} + 2B_2K_{lt}K_{12} \\ C_0 &= 2K_{lt}K_{se}K_{12} + K_1K_{st}K_2 \end{aligned}$$

Converting from x to θ gives

$$\begin{aligned} \delta\Big(K_{se}K_{12}\left(F_{ag} - F_{ant}\right) + (K_{se}B_{34} + B_2K_{12})\left(\dot{F}_{ag} - \dot{F}_{ant}\right) + B_2B_{34}\left(\ddot{F}_{ag} - \ddot{F}_{ant}\right)\Big) \\ = \dddot{\theta} + P_3\ddot{\theta} + P_2\ddot{\theta} + P_1\dot{\theta} + P_0\theta \end{aligned} \tag{13.48}$$

where

$$\delta = \frac{57.296}{rJB_{12}B_{34}},\ P_3 = \frac{C_3}{C_4},\ P_2 = \frac{C_2}{C_4},\ P_1 = \frac{C_1}{C_4},\ P_0 = \frac{C_0}{C_4}$$

Based on an analysis of experimental data, suitable parameter estimates for the oculomotor plant are as follows:

$K_{SE} = 125\ \text{Nm}^{-1}$
$K_{LT} = 60.7\ \text{Nm}^{-1}$
$B_1 = 2.0\ \text{Nsm}^{-1}$
$B_2 = 0.5\ \text{Nsm}^{-1}$
$J = 2.2 \times 10^{-3}\ \text{Ns}^2\text{m}^{-1}$
$B_3 = 0.538\ \text{Nsm}^{-1}$
$B_4 = 41.54\ \text{Nsm}^{-1}$
$K_1 = 26.9\ \text{Nm}^{-1}$
$K_2 = 41.54\ \text{Nm}^{-1}$

Based on the updated model of muscle and length tension data presented in the previous section, steady-state active-state tensions are determined as

$$F = \begin{cases} 0.4 + 0.0175\theta & \text{N for } \theta \geq 0^\circ \\ 0.4 + 0.0125\theta & \text{N for } \theta < 0^\circ \end{cases} \tag{13.49}$$

The agonist and antagonist active-state tensions follow from Figure 13.11, which assume no latent period, and are given by the following low-pass filtered waveforms:

$$\dot{F}_{ag} = \frac{N_{ag} - F_{ag}}{\tau_{ag}} \quad \text{and} \quad \dot{F}_{ant} = \frac{N_{ant} - F_{ant}}{\tau_{ant}} \tag{13.50}$$

where N_{ag} and N_{ant} are the neural control inputs (pulse-step waveforms), and

$$\tau_{ag} = \tau_{ac}(u(t) - u(t - t_1)) + \tau_{de}u(t - t_1)$$

$$\tau_{ant} = \tau_{de}(u(t) - u(t - t_1)) + \tau_{ac}u(t - t_1)$$

are the time-varying time constants.

Saccadic eye movements simulated with this model have characteristics that are in good agreement with the data, including position, velocity and acceleration, and the main sequence diagrams. As before, the relationship between agonist pulse magnitude and pulse duration is tightly coupled.

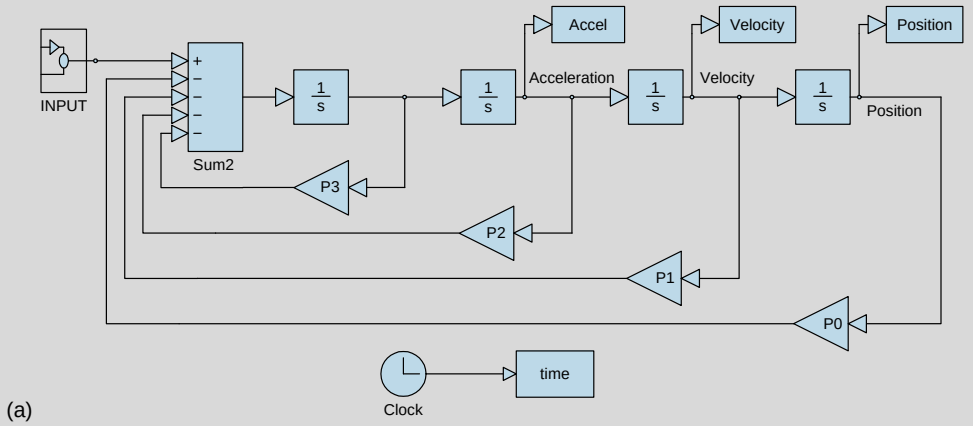
EXAMPLE PROBLEM 13.7

Using the oculomotor plant model described with Eq. (13.48), parameters given in this section, and the steady-state input from Eq. (13.49), simulate a 10° saccade. Plot agonist and antagonist active-state tension, position, velocity, and acceleration versus time. Compare the simulation with the main sequence diagram in Figure 13.6.

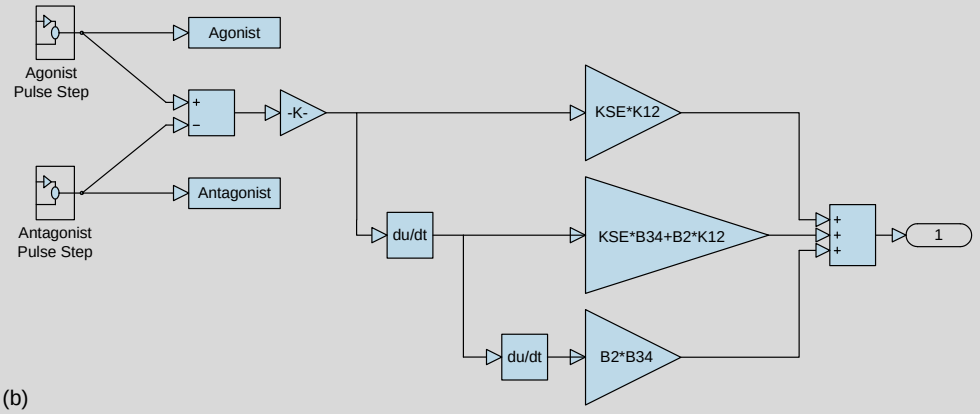
Continued

Solution

The solution to this example involves selecting a set of parameters $(F_p, t_1, \tau_{ac}$, and $\tau_{de})$ that match the characteristics observed in the main sequence diagram shown in Figure 13.6. There is a great deal of flexibility in simulating a 10° saccade. The only constraints for the 10° saccade simulation results are that the duration is approximately 40 to 50 ms and peak velocity is in the 500 to 600°s^{-1} range. For realism, a latent period of 150 ms has been added to the simulation results. A SIMULINK block diagram of Eq. (13.48) is shown in Figure 13.42.



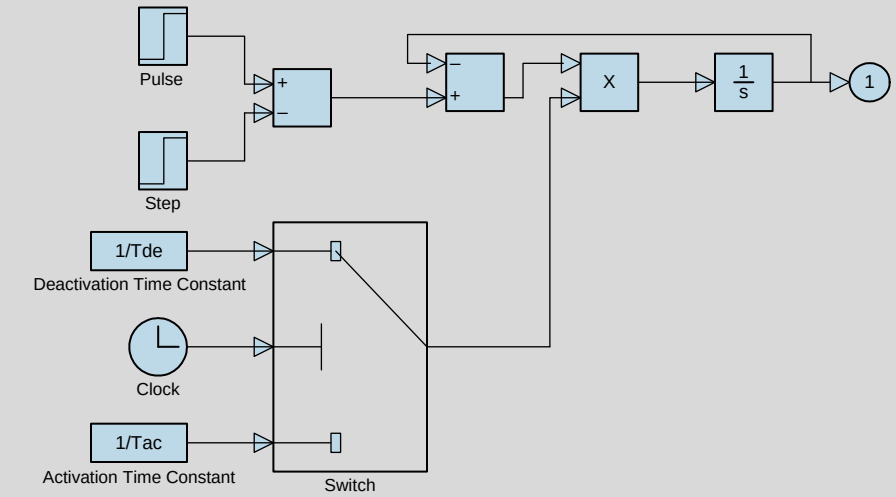
(a)



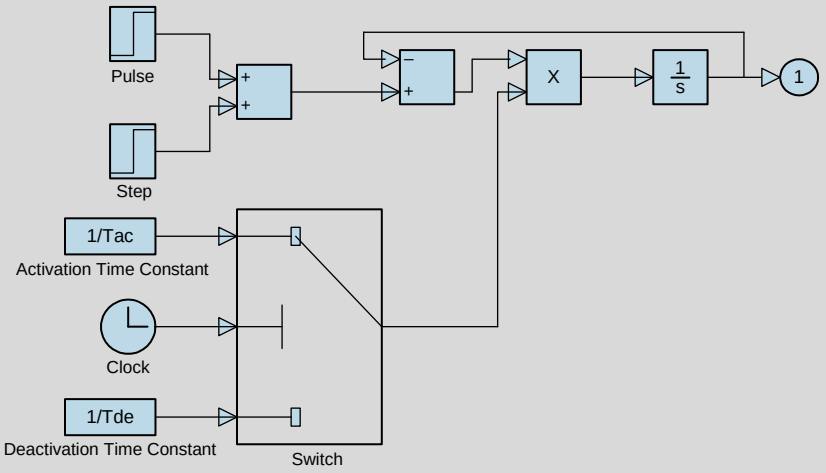
(b)

FIGURE 13.42 SIMULINK block diagrams for Example Problem 13.7. (a) Model defined by Eq. (13.49). (b) The input.

Continued



(c)



(d)

FIGURE 13.42, cont'd (c) Agonist pulse-step. (d) Antagonist pulse-step.

The response of the system is shown in Figure 13.43, with $F_p = 1.3$ N, $t_1 = 0.01$ s, $\tau_{ac} = 0.018$ s, and $\tau_{de} = 0.018$ s. These 10° simulation results have the main sequence characteristics with a peak velocity of $568°s^{-1}$ and a duration of 45 ms.

Continued

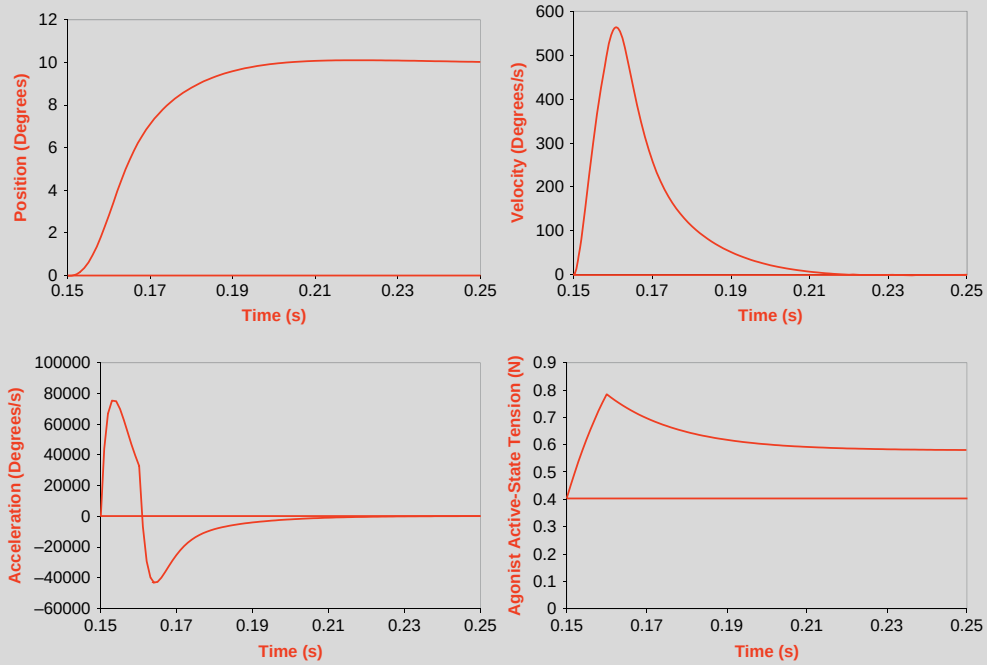


FIGURE 13.43 Simulation for Example Problem 13.7, with $F_p = 1.3\,\text{N}$, $t_1 = 0.10\,\text{s}$, $\tau_{ac} = 0.018\,\text{s}$, and $\tau_{de} = 0.018\,\text{s}$.

Many other parameter sets can also simulate a 10° saccade. For instance, consider reducing τ_{de} to .009 s. Because the antagonist active-state tension activity goes toward zero more quickly than in the last case, a greater total active-state tension ($F_{ag} - F_{ant}$) results. Therefore, to arrive at 10° with the appropriate main sequence characteristics, F_p needs to be reduced to 1.0 N if τ_{ac} remains at 0.018 s and t_1 equals 0.0115 s. This 10° simulation is shown in Figure 13.44.

To simulate larger saccades with main sequence characteristics, the time constants for the agonist and antagonist active-state tensions can be kept at the same values as the 10° saccades or made functions of saccade amplitude (see [1] for several examples of amplitude-dependent time constants). Main sequence simulations for 15° and 20° saccades are obtained with $F_p = 1.3\,\text{N}$ and the time constants are both fixed at 0.018 s (the first case) by changing t_1 to 0.0155 and 0.0223 s, respectively. For example, the 20° simulation results are shown in Figure 13.45, with a peak velocity of $682°\text{s}^{-1}$ and a duration of 60 ms.

In general, as F_p increases, t_1 decreases to maintain the same saccade amplitude. Additionally, peak velocity increases as F_p increases. For the saccade amplitude to remain a constant as either or both of the time constants increase, F_p should also increase.

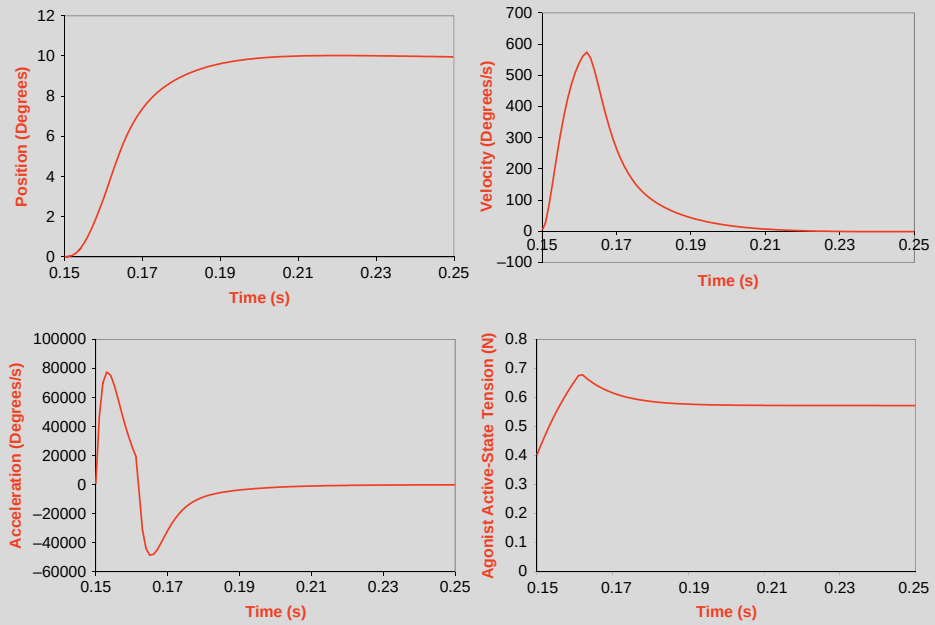


FIGURE 13.44 Simulation for Example Problem 13.7, with $F_p = 1.0\,\text{N}$, $t_1 = 0.010\,\text{s}$, $\tau_{ac} = 0.018\,\text{s}$, and $\tau_{de} = 0.009\,\text{s}$.

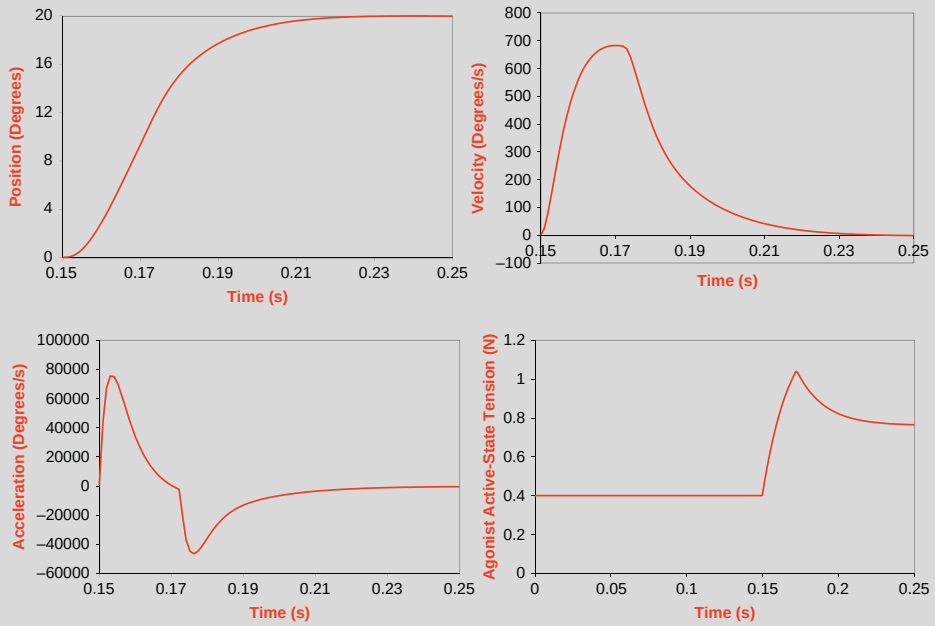


FIGURE 13.45 Simulation for Example Problem 13.7, with $F_p = 1.3\,\text{N}$, $t_1 = 0.0223\,\text{s}$, $\tau_{ac} = 0.018\,\text{s}$, and $\tau_{de} = 0.018\,\text{s}$.

13.8 THE 2009 LINEAR HOMEOMORPHIC SACCADIC EYE MOVEMENT MODEL

In this section, we further explore the fast eye movement system that has postsaccade behavior, including normal saccades and those with a dynamic or a glissadic overshoot based on a model by Zhou, Chen, and Enderle [50] and Enderle and Zhou [18]. Postsaccade phenomena, such as a dynamic or a glissadic overshoot, are usually observed during human saccades [47]. In dynamic overshoot, the eyes move beyond the target, and then, with a quick saccade-like return with no time delay, the eyes move back to the target. Glissadic overshoot is similar to dynamic overshoot but with a return that is slower. To analyze postsaccade behavior, the neural input to the muscles is now described by a pulse-slide-step of neural activity, supported by physiological evidence [26]. The slide is a slow exponential transition from the pulse to the step.

The oculomotor plant is shown in Figure 13.46. It should be noted that the passive elasticity of the eyeball in Figure 13.46 is changed from the model in Figure 13.41 that included two Voigt passive elements connected in series to a single Voigt element. The Voigt element with time constant 0.02 s is used in the model presented here. The other Voigt element, with a time constant of 1 s, is neglected, since it has an insignificant effect on the accuracy because we are modeling a single saccade and not a series of saccades. Further, eliminating

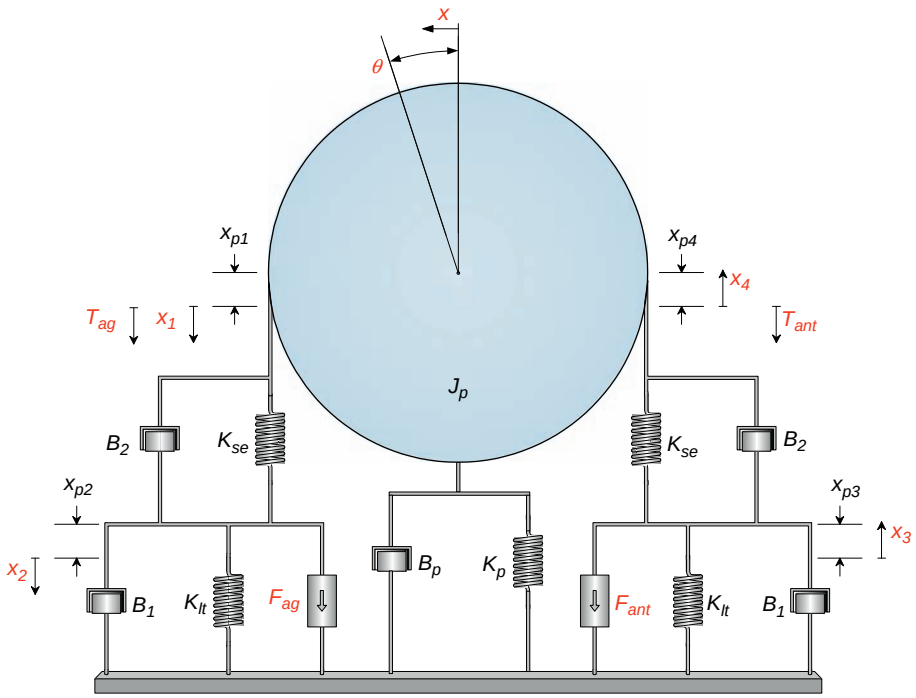


FIGURE 13.46 Oculomotor plant used for analyzing saccades with postsaccade behavior.

this Voigt element reduces the order of the model from fourth to third order and simplifies the system identification. The net torque generated by the muscles during a saccade rotates the eyeball to a new orientation and, after the saccade is completed, compensates the passive restraining torques generated by orbital tissues.

By summing the forces at junctions 2 and 3 (the equilibrium positions for x_2 and x_3) and the torques acting on the eyeball, using Laplace variable analysis about the operating point, the linear homeomorphic model, as shown in Figure 13.46, is derived as

$$\delta\left(B_2\left(\dot{F}_{ag}-\dot{F}_{ant}\right)+K_{se}\left(F_{ag}-F_{ant}\right)\right)=\ddot{\theta}+P_2\ddot{\theta}+P_1\dot{\theta}+P_0\theta \tag{13.51}$$

where

$$\delta=\frac{5208.7}{JB_{12}}$$

$$P_2=\frac{JK_{st}+B_{12}B+2B_1B_2}{JB_{12}}$$

$$P_1=\frac{2B_1K_{se}+2B_2K_{lt}+B_{12}K+K_{st}B}{JB_{12}}$$

$$P_0=\frac{K_{st}K+2K_{lt}K_{se}}{JB_{12}}$$

Full details of the derivation are provided in Enderle and Zhou [18].

13.8.1 Neural Input

Previously, we modeled the neural input to the saccade system as a pulse-step waveform. This input has been used in many studies because of its simplicity and ease of use [4, 17, 20]. To create a more realistic input based on physiological evidence, a pulse-slide-step input is used as shown in Figure 13.47 (based on [26]). The slide is an exponential transition from the pulse to the step. This model is consistent with the data published in the literature (for example, see Figure 4 in [42], and Figure 2 in [46]). The diagram in Figure 13.47 (top) closely approximates the data shown in Figure 13.47 (bottom) for the agonist input.

At steady state, the eye is held steady by the agonist and antagonist inputs F_{g0} and F_{t0}. We typically define the time when the target moves as $t=0$. This is a common assumption, since many simulation studies ignore the latent period and focus on the actual movement.

The overall agonist pulse occurs in the interval $0-T_2$, with a more complex behavior than the pulse described earlier. We view the overall pulse process as the intention of the system, which is limited by its physical capabilities. The start of the pulse occurs with an exponential rise from the initial firing rate, F_{g0}, to peak magnitude, F_{p1}, with a time constant τ_{gn1}. At T_1, the input decays to F_{p2}, with a time constant τ_{gn2}. The slide occurs at T_2, with a time constant τ_{gn3}, to F_{gs}, the force necessary to hold the eye at its destination. The input F_{gs} is applied during the step portion of the input.

At $t=0$, the antagonist neural input is completely inhibited and exponentially decays to zero from F_{t0} with time constant τ_{tn1}. At time T_3, the antagonist input exponentially increases with time constant τ_{tn2}. The antagonist neural input shown in Figure 13.47 (middle)

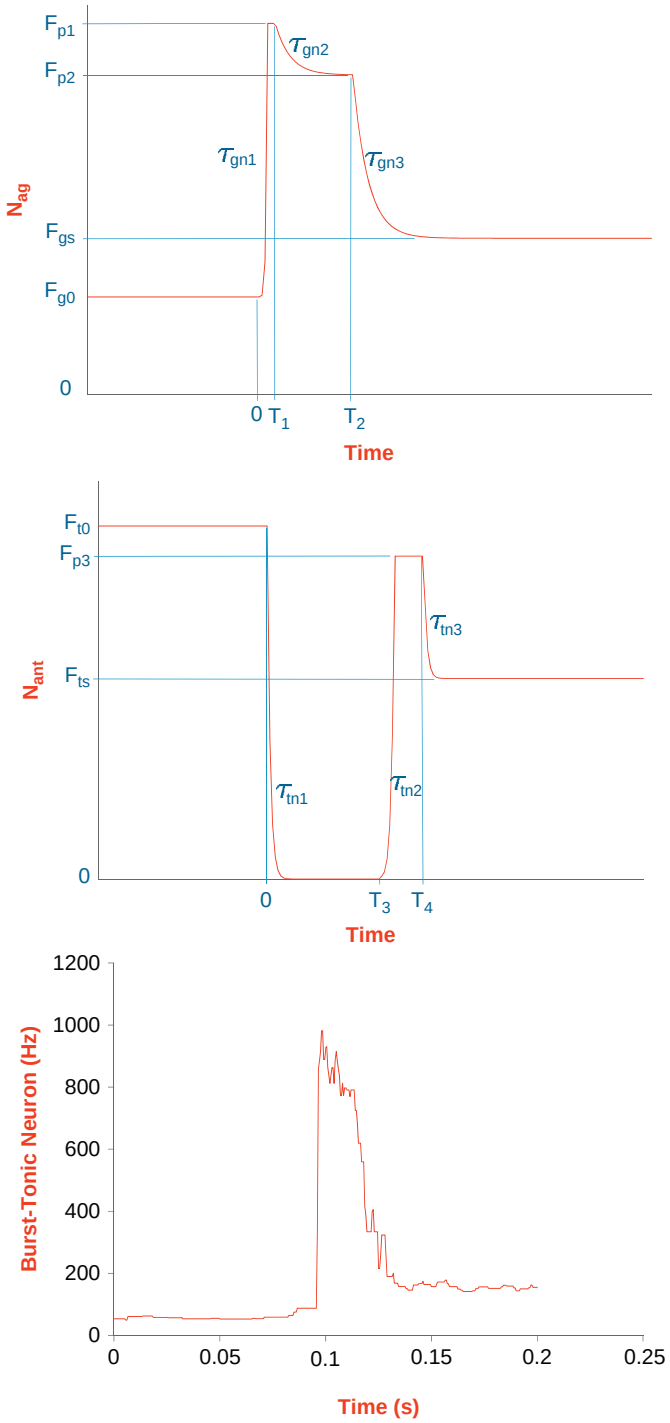


FIGURE 13.47 Neural input to the horizontal saccade system. (Top) Agonist input. (Middle) Antagonist input. (Bottom) Discharge rate of a single burst-tonic neuron during a saccade (agonist input). Details of the experiment and training for the bottom diagram are reported elsewhere. *From Sparks et al. [44]. Data provided personally by Dr. David Sparks.*

includes a PIRB pulse with duration of $T_4 - T_3$. At T_4, the antagonist input exponentially decays to F_{ts}, with a time constant τ_{tn3}. If no PIRB occurs in the antagonist input, the input exponentially rises to F_{ts} with time constant τ_{tn2}.

The agonist pulse includes an interval (T_1) that is constant for saccades of all sizes as supported by physiological evidence [11, 50]. We choose to model the change in the firing rate with an exponential function because this seems to match the data fairly well.

After complete inhibition, the antagonist neural input has a brief excitatory pulse starting at T_3 with duration of approximately 10 ms. Enderle proposed that this burst is generated by PIRB, a property that contributes to the postsaccade phenomena such as dynamic and glissadic overshoot (2002). The active-state tensions are again defined as low-pass filtered neural signals:

$$\dot{F}_{ag} = \frac{N_{ag} - F_{ag}}{\tau_{ag}} \tag{13.52}$$

$$\dot{F}_{ant} = \frac{N_{ant} - F_{ant}}{\tau_{ant}} \tag{13.53}$$

where

$$\tau_{ag} = \tau_{gac}(u(t - T_1) - u(t - T_2)) + \tau_{gde}u(t - T_2) \tag{13.54}$$

$$\tau_{ant} = \tau_{tde}(u(t) - u(t - T_3)) + \tau_{tac}(u(t - T_3) - u(t - T_4)) + \tau_{tde}u(t - T_4) \tag{13.55}$$

The activation and deactivation time constants represent the different dynamic characteristics of muscle under increasing and decreasing stimulation.

13.8.2 Parameter Estimation and System Identification

The model presented here involves a total of 25 parameters describing the oculomotor plant, neural inputs, and active-state tensions that are estimated by system identification. Initial estimates of the model parameters are important, since they affect the convergence of the estimation routine. In this model, the initial estimates are derived from previously published experimental observations, with a more detailed discussion of the parameter estimates for human and monkey given in Zhou, Chen, and Enderle [50] and Enderle and Zhou [18]. Oculomotor parameters are given in Table 13.1. The transfer function for the oculomotor plant is

$$H(s) = \frac{\theta}{\Delta F} = \frac{\delta B_2\left(s + \frac{K_{se}}{B_2}\right)}{s^3 + P_2 s^2 + P_1 s + P_0} \tag{13.56}$$

where $\Delta F = F_{ag} - F_{ant}$. Using the parameter values in Table 13.1, we have the transfer function for humans as

$$H(s) = \frac{1.9406 \times 10^5 (s + 250)}{s^3 + 596s^2 + 1.208 \times 10^5 s + 1.3569 \times 10^6} \tag{13.57}$$

TABLE 13.1 Comparison of Parameters for Monkeys and Humans

Parameter	Human	Rhesus Monkey
Radius of eyeball	11 mm (11.8 mm in model)	10 mm
K_{se}	125 N/m	125 N/m
K_{lt}	60.7 N/m	77.66 N/m
B_1	5.6 Ns/m	4 Ns/m
B_2	0.5 Ns/m	0.4 Ns/m
F	$F = \begin{cases} 0.4 + 0.0175\lvert\theta\rvert, & \text{for } \theta > 0 \\ 0.4 - 0.0125\lvert\theta\rvert, & \text{for } \theta \leq 0 \end{cases}$	$F = \begin{cases} 0.55 + 0.0175\lvert\theta\rvert, & \text{for } \theta > 0 \\ 0.55 - 0.0125\lvert\theta\rvert, & \text{for } \theta \leq 0 \end{cases}$
K	16.34 N/m	10.21 N/m
B	0.327 Ns/m	0.204 Ns/m
J	2.2×10^{-3} Ns^2/m	1.76×10^{-3} Ns^2/m

and for monkeys

$$H(s) = \frac{2.6904 \times 10^5 (s + 312.5)}{s^3 + 575.2s^2 + 1.4829 \times 10^5 s + 2.7743 \times 10^6} \tag{13.58}$$

There are three poles and the one zero in the transfer function. Using the parameter values in Table 13.1 for humans, the poles are

$-292.22 + j168.63$

$-292.22 - j168.63$

-11.92

and the zero is

250

For monkeys, the poles are

$-277.48 + j245.09$

$-277.48 - j245.09$

-20.24

and the zero is

312.5

For humans, the time constant for the real pole is 3.4 ms, and the time constant for the complex pole is −83.9 ms. Similarly, for monkeys we have 3.6 and 49.4 ms.

EXAMPLE PROBLEM 13.8

Consider the oculomotor system shown in Figure 13.46 and Eq. (13.51). Given the following initial conditions and parameter, create a SIMULINK program and plot the neural inputs, active-state tensions, position, velocity, and acceleration.

$\theta(0) = 0^\circ$, $\dot{\theta}(0) = 0^\circ s^{-1}$, $\ddot{\theta}(0) = 0^\circ s^{-2}$, $T_1 = 0.0044$ s, $T_2 = 0.0259$ s, $T_3 = 0.0293$ s,
$T_4 = 0.0462$ s, $F_{p1} = 1.06$ N, $F_{p2} = 0.9331$ N, $F_{p3} = 0.3790$ N, $F_{g0} = 0.4$ N, $F_{gs} = 0.5546$ N,
$F_{t0} = 0.4$ N, $F_{ts} = 0.2895$ N, $\tau_{gn1} = 0.000287$ s, $\tau_{gn2} = 0.0034$ s, $\tau_{gn3} = 0.0042$ s,
$\tau_{gac} = 0.0112$ s, $\tau_{tn1} = 0.000939$ s, $\tau_{tn2} = 0.0012$ s, $\tau_{tn3} = 0.001$ s, $\tau_{tac} = 0.0093$ s,
$\tau_{tde} = 0.0048$ s, $K_{se} = 124.9582$ Nm, $K_{lt} = 60.6874$ Nm, $K = 16.3597$ Nm,
$B_1 = 5.7223$ Nms^{-1}, $B_2 = 0.5016$ Nms^{-1}, $B = 0.327$ Nms^{-1}, $J = 0.0022$ Nms^{-1},
and radius $= 0.0118$ m.

Solution

We first compute the following intermediate results in MATLAB:

```
b12=b1+b2=6.2239
kst=kse+klt=185.6456
c3=b12*j=0.0137
c2=b12*bp+kst*j+2*b1*b2=8.1855
c1=b12*kp+kst*bp+2*(b2*klt+b1*kse)= 1.6535e+003
c0=kst*kp+2*kse*klt=1.8204e+004
delta=57.296/(r*c3)= 3.5397e+005
p2=c2/c3=596.7159
p1=c1/c3=1.2054e+005
p0=c0/c3=1.3271e+006
```

The SIMULINK model is shown in Figure 13.48. The diagram in (a) is implemented using Eq. (13.51). The input to the oculomotor plant is shown in diagram (b), with the agonist and antagonist active-state tensions shown in (c) and (d) based on Eqs. (13.52) and (13.53), respectively. The agonist and antagonist neural input is shown in Figures 13.48e and f.

Figure 13.49 shows plots of position, velocity, acceleration, agonist neural input and active-state tension, and antagonist neural input and active-state tension. From the antagonist neural input and active-state tension plot, it is clear that the eye movement has postsaccade behavior. Peak return velocity is –20 $^\circ s^{-1}$ in Figure 13.49b, which makes it a glissade.

Continued

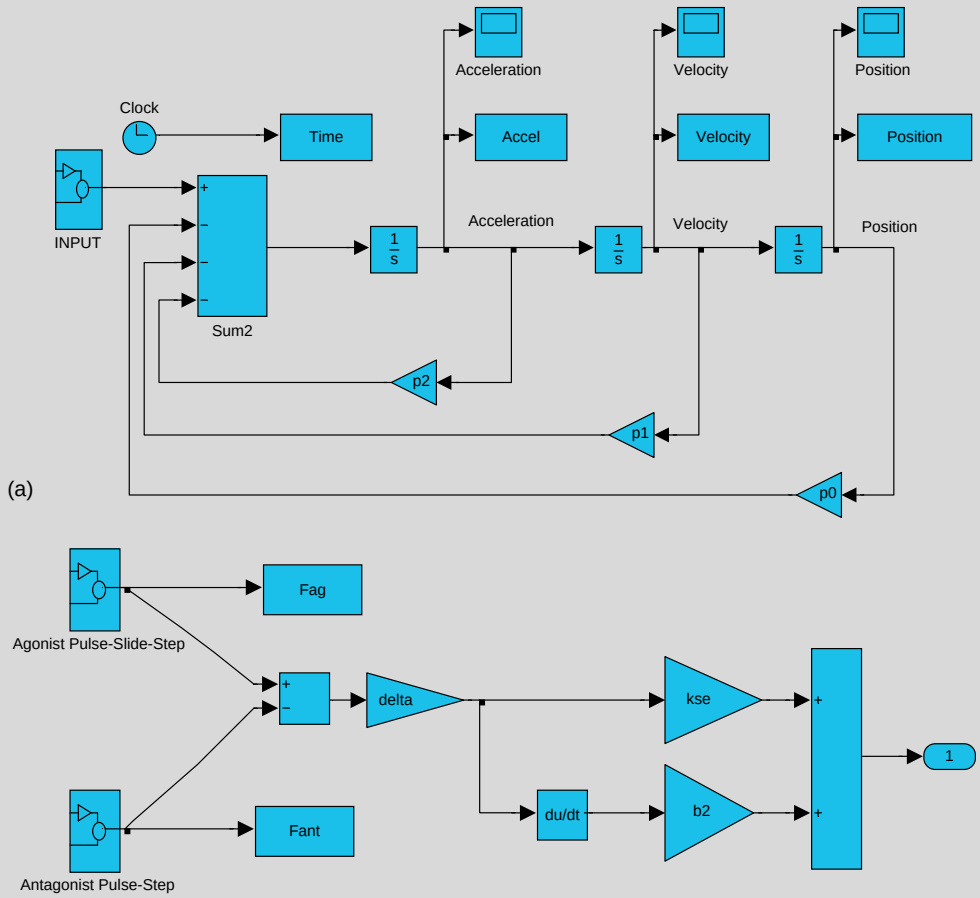


FIGURE 13.48 SIMULINK program for Example 13.8. (a) Main program. (b) Input to plant.

Continued

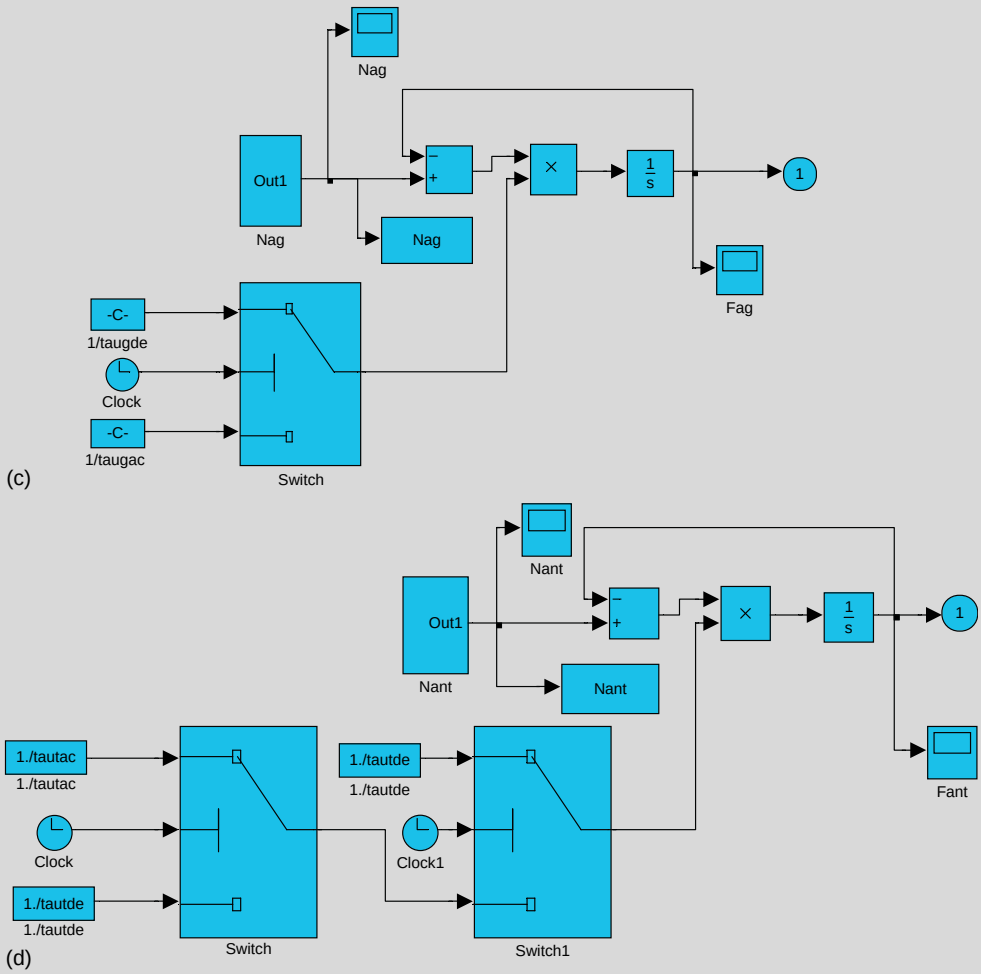


FIGURE 13.48, cont'd (c) Agonist active-state tension. (d) Antagonist active-state tension.

Continued

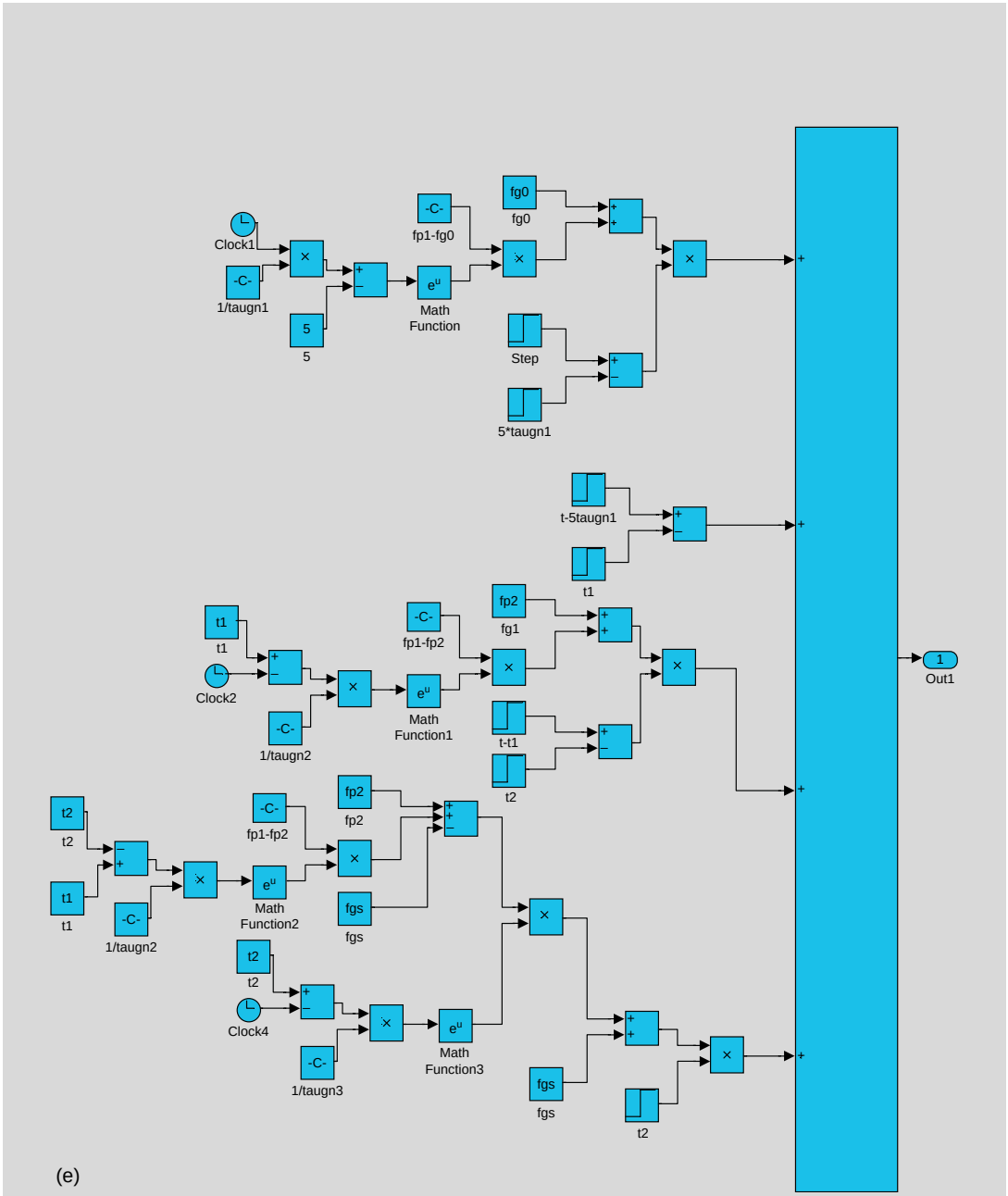


FIGURE 13.48, cont'd (e) Agonist neural input.

Continued

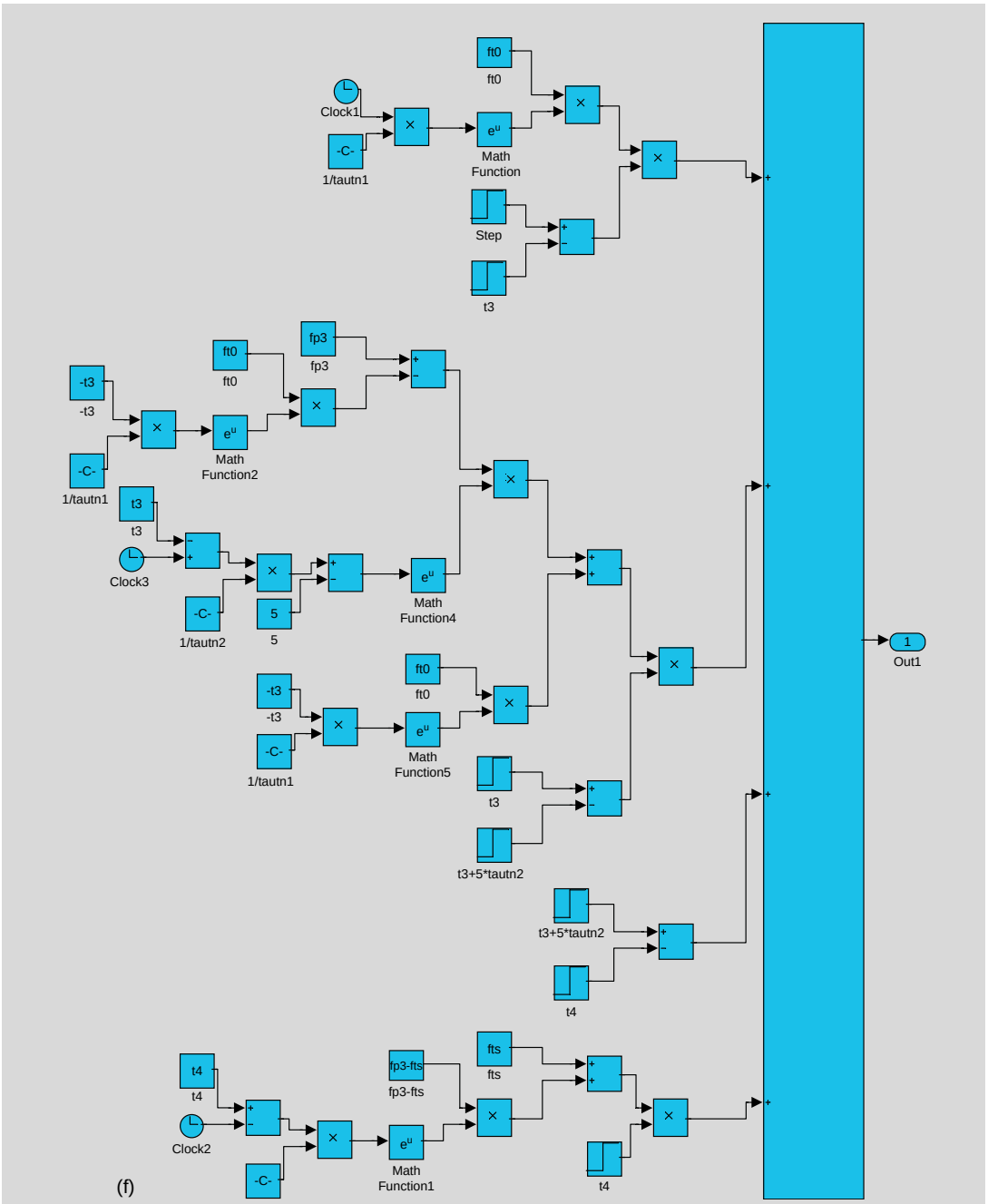
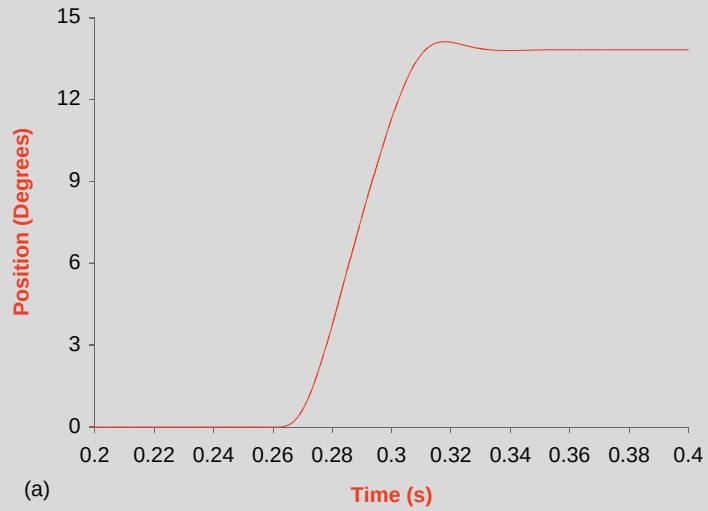


FIGURE 13.48, cont'd (f) Antagonist neural input.



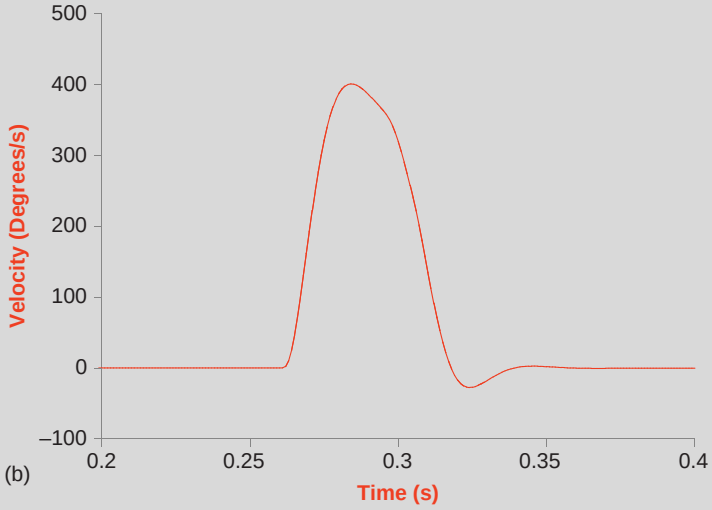
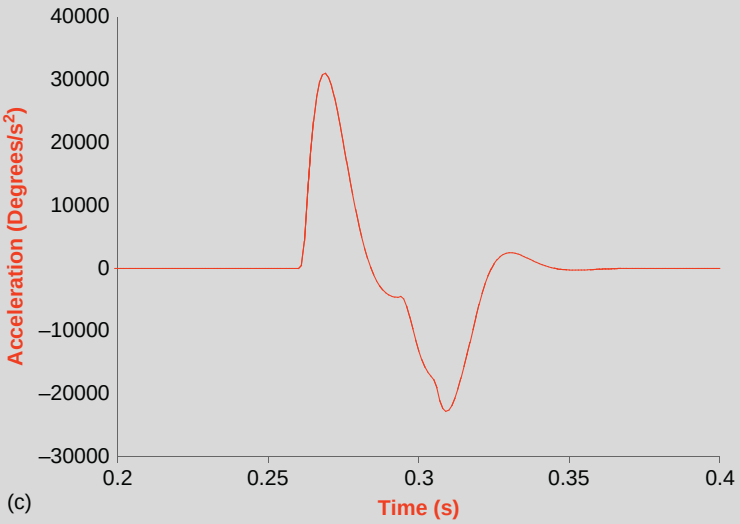
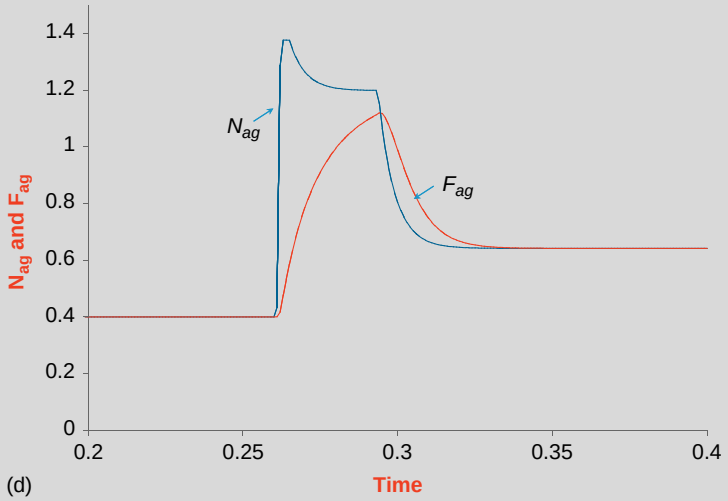


FIGURE 13.49 Plots of position, velocity, and acceleration (a–c) for Example Problem 13.8. Plots of agonist neural input and active-state tension, and antagonist neural input and active-state tension for Example Problem 13.8, (d) and (e).

Continued



(c)



(d)

FIGURE 13.49, cont'd

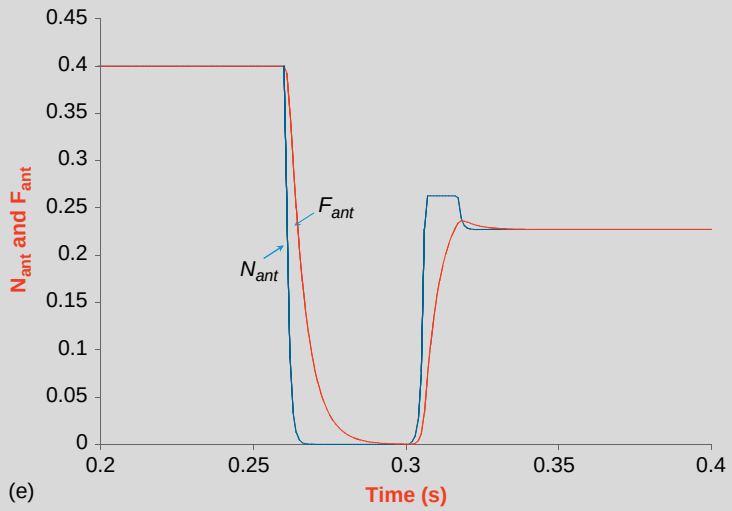


FIGURE 13.49, cont'd

13.8.3 Monkey Data and Results

Data[1] were collected from a rhesus monkey that executed a total of 27 saccades in our data set for 4°, 8°, 16°, and 20° target movements. Neuron data were recorded from the long lead burst neuron (5 saccades), excitatory burst neuron (17 saccades), and the agonist burst-tonic neuron (5 saccades) (the neuron types are described in Section 13.9). The firing of the burst-tonic neuron is similar to the motoneuron that drives the agonist muscle during a saccade. Figure 13.50 shows the estimation results for three saccades (4°, 8°, and 15°).

Figure 13.51 shows the estimated neural inputs and active-state tensions that generate the saccades shown in Figure 13.50. Also shown are the firing rates recorded from a single burst-tonic cell in a rhesus monkey for these saccades, scaled to match the height of N_{ag}. The shapes of the model's neural inputs approximate the burst-tonic data during the pulse and slide very closely. The estimated agonist neural input N_{ag} clearly has a similar shape to the firing rate data. It should be noted that the firing activity in the data comes from a single burst-tonic neuron. The neural input to the oculomotor plant is actually due to the firing of more than 1,000 motoneurons.

13.8.4 Human Data and Results

Data[2] were collected from three human subjects executing 127 saccades, many with dynamic overshoots or glissades. Figure 13.52 shows representative model estimates of saccades generated with a dynamic overshoot, a glissadic overshoot, and normal characteristics.

[1]Details of the experiment and training are reported in Sparks et al. [44]. Data provided personally by Dr. David Sparks.

[2]Details of the experiment are reported in Enderle and Wolfe [17].

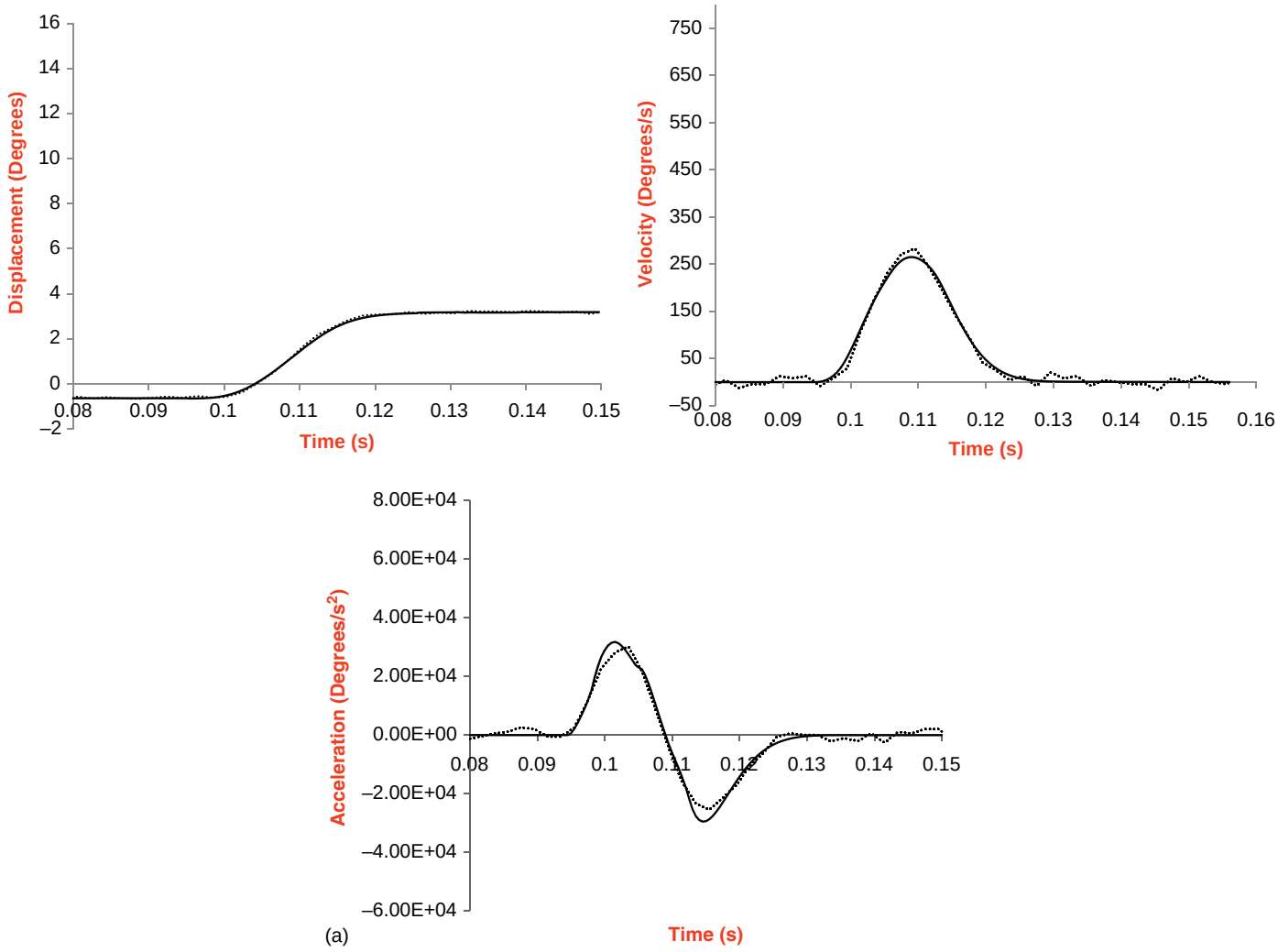


FIGURE 13.50 Eye position, velocity, and acceleration for three different saccades in a rhesus monkey (a: 4°, b: 8°, and c: 15°). The solid lines are the model predictions, and the doted lines are the experimental data during the saccadic eye movement.

Continued

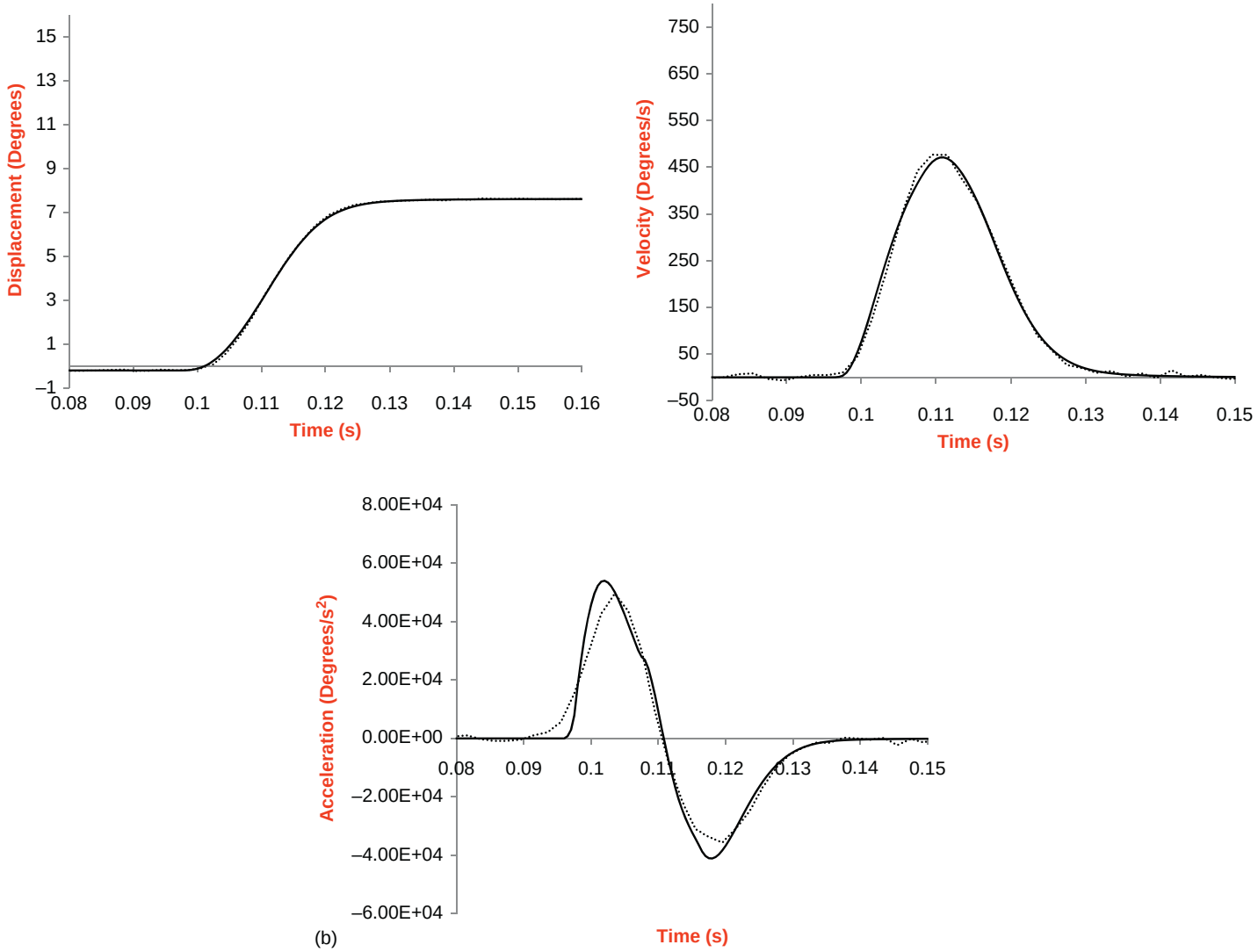
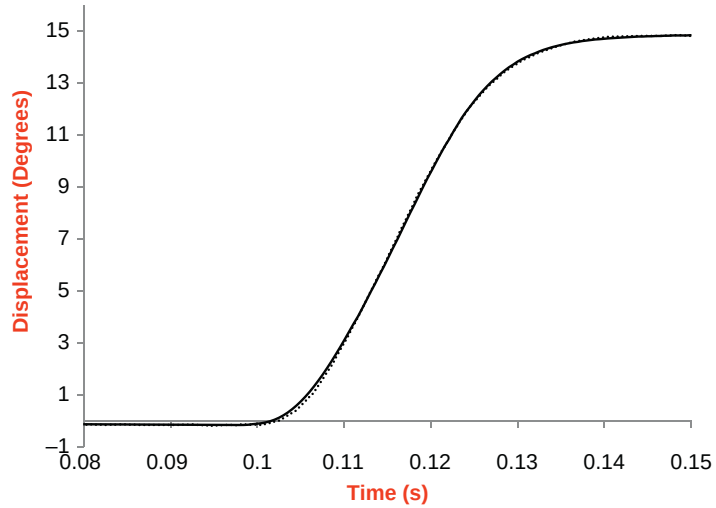
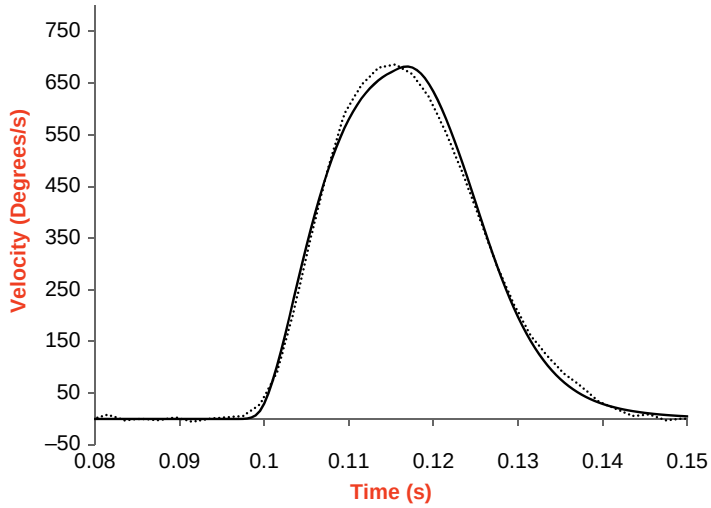
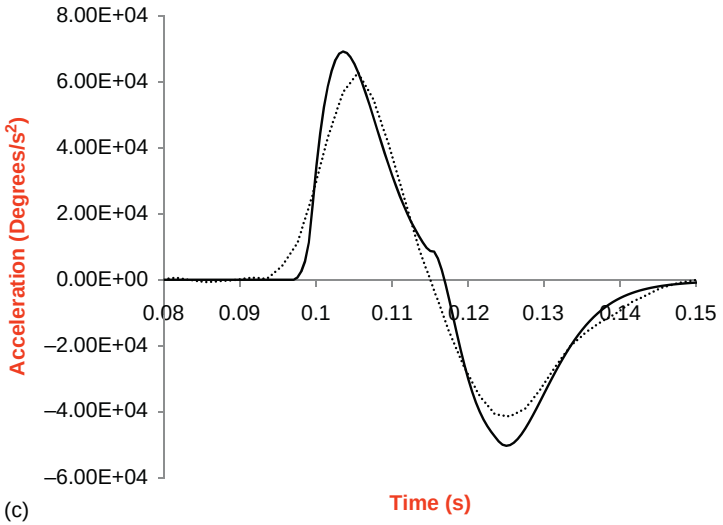


FIGURE 13.50, cont'd

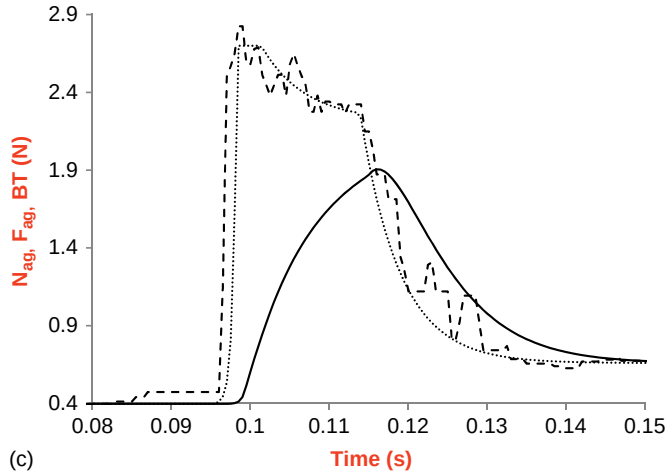






(c)

FIGURE 13.50, cont'd



(c)

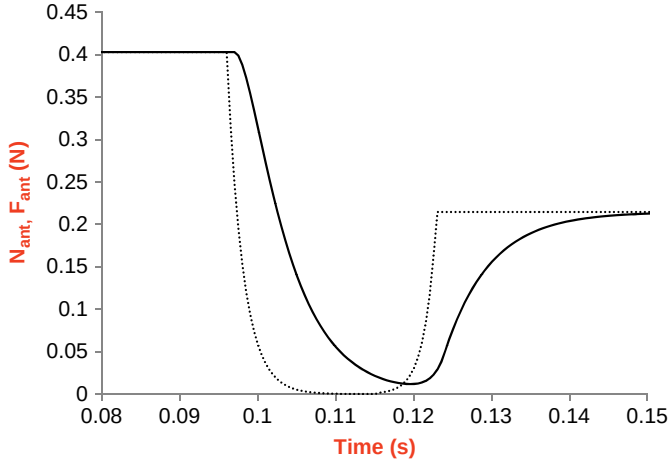
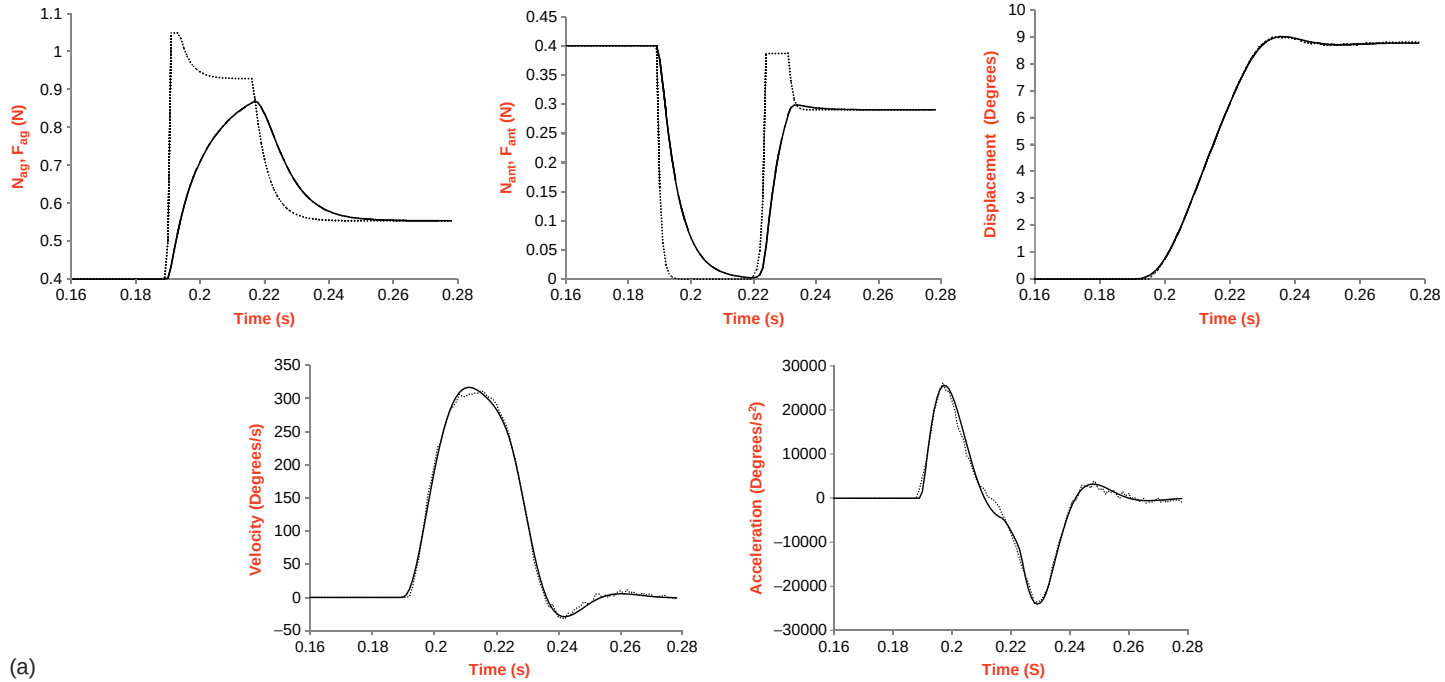


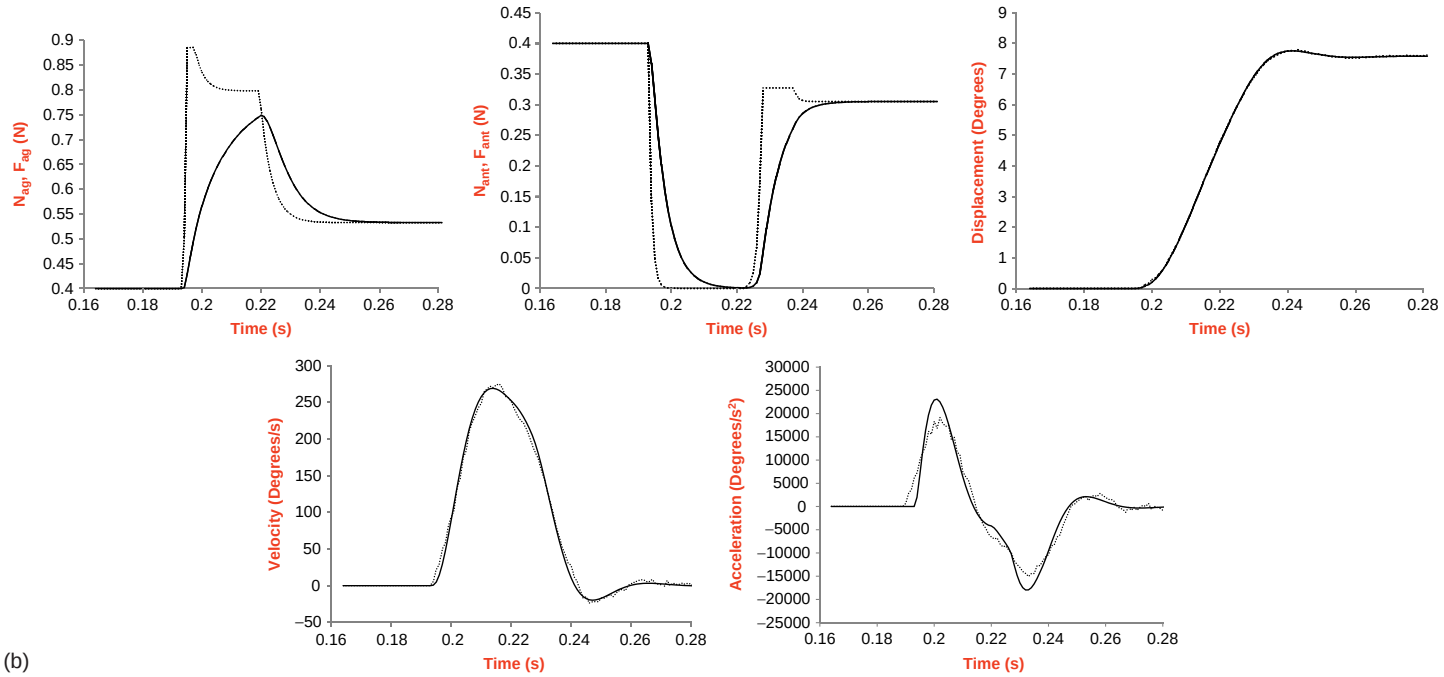
FIGURE 13.51, cont'd



(a)

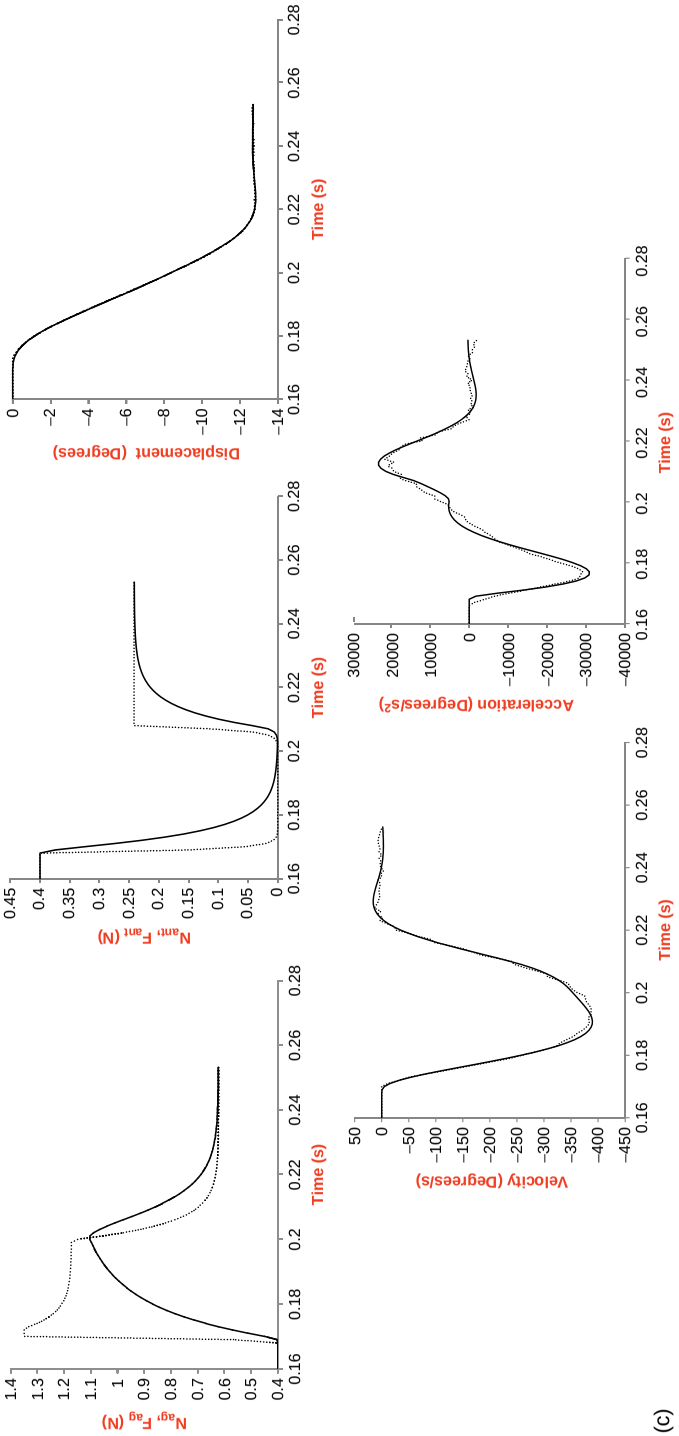
FIGURE 13.52 (a) Dynamic overshoot saccade of 8°, (b) glissadic overshoot saccade of 8°, and (c) normal –12° saccade. The first two lines of graphs are the active-state tension (solid line) and neural input (dotted line) calculated from the parameter estimation. Also shown are the model predictions using the parameter estimates from the system identification technique for displacement, velocity, and acceleration and the data (dotted line is the data and solid line is the model predictions).

Continued



(b)

FIGURE 13.52, cont'd



(c)

FIGURE 13.52, cont'd

The model predictions for all saccades match displacement data and estimates of velocity very well, including saccades with a dynamic or a glissadic overshoot, with accuracy similar to those in Figure 13.52.

The 8° saccade shown in Figure 13.52a of data and model predictions has dynamic overshoot. Note that the saccade with dynamic overshoot is caused by a PIRB firing in the antagonist neural input at approximately 220 ms. The PIRB induces prominent reverse peak velocity as shown.

Figure 13.52b shows model predictions and data for an 8° saccade with glissadic overshoot. The glissade is caused by the PIRB in the antagonist neural input at approximately 223 ms. Notice the peak firing for a saccade with glissadic overshoot is smaller than one with dynamic overshoot. The PIRB induces reverse peak velocity that is smaller than the one with dynamic overshoot. In glissadic overshoot, the eye has an overshoot that returns to steady state more gradually. As a result, the glissade has a smaller peak velocity.

A –12° normal saccade is shown in Figure 13.52c. Normal saccades usually do not have a PIRB, although this is not absolute, since the timing of the PIRB might offset the impact of the burst.

The main sequence diagram is shown in Figure 13.53. Peak-velocity estimates from the model are in close agreement with the data estimates of peak velocity and follow an exponential shape as a function of saccade magnitude. Duration has a linear relationship with saccade magnitude for saccades above 7°. For saccades between 3 to 7°, duration is approximately constant. It should be noted that saccade duration is difficult to determine, especially for small saccades, and may be a source of differences with other published data. The latent period is relatively independent of saccade magnitude.

The estimated agonist pulse magnitudes and durations are shown in Figure 13.54 for all 127 saccades. The agonist pulse magnitude does not significantly increase as a function of saccade magnitude for saccades larger than 7°, consistent with the time-optimal controller proposed by Enderle [11], Zhou and coworkers [50], and Enderle and Zhou [18]. For saccades under 7°, agonist pulse magnitude shows a linear increase in pulse magnitude versus saccade magnitude, again in agreement with our theory for the saccade controller. A great variability is observed in the pulse magnitude estimates for saccades of the same magnitude, which is also observed by Hu and coworkers [34] in their analysis of the firing rates in the monkey EBN. The agonist pulse duration increases as a function of saccade magnitude for saccades larger than 7°. For saccades between 3 to 7°, the duration of the agonist pulse is relatively constant as a function of saccade magnitude. Note that for all saccades, the pulse magnitude is tightly coordinated with the pulse duration.

13.8.5 Postinhibitory Rebound Burst and Postsaccade Phenomena

Inhibition of antagonist burst neurons[3] is postulated to cause an unplanned postinhibitory rebound burst (PIRB) toward the end of a saccade that causes dynamic overshoots or glissades [11]. While some studies do not observe the rebound firing in the abducens neurons in monkeys [24, 38, 45], PIRB are observed in the abducens motoneurons at the

[3]The neurons describe in this section are fully described in Section 14.9, including a neural network. Only a brief description of the neurons involved in PIRB is given here.

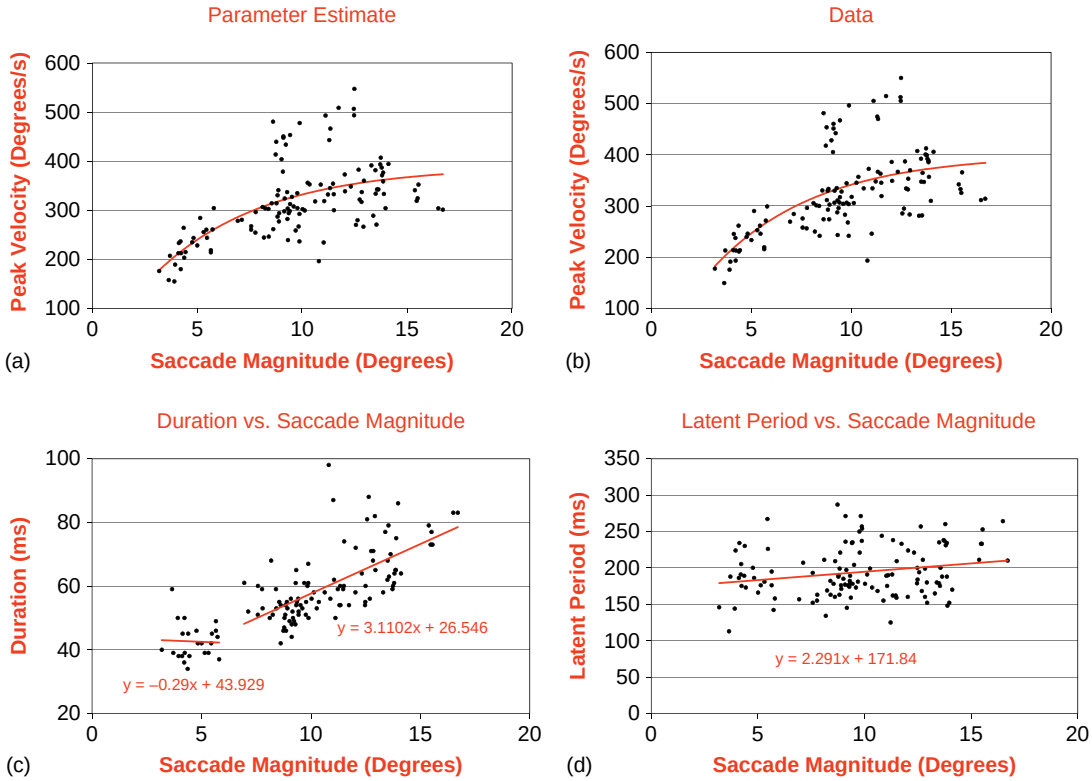


FIGURE 13.53 Main sequence diagram for all 127 saccades from three human subjects. (a) Peak velocity versus saccade magnitude from the model estimates, with regression fit $\dot{\theta}_{pv} = 390\left(1 - e^{-0.2\theta_{ss}}\right)$. (b) Peak velocity versus saccade magnitude from the data, with regression fit $\dot{\theta}_{pv} = 401\left(1 - e^{-0.2\theta_{ss}}\right)$. (c) Duration versus saccade magnitude based on the data. (d) Latent period versus saccade magnitude based on the data. Note that the parameter estimation program did not update the duration or the latent period, so a single graph for each is drawn.

end of off-saccades in monkeys in other studies (for examples, see [42] and [46]). It was noted earlier that saccades with dynamic overshoots or glissades do not occur with the same frequency in the monkey as in humans, and that they are absent from our monkey data.

The theory presented here is that, at least in humans, the antagonist PIRB causes a reverse peak velocity during dynamic overshoots or glissades in humans. The model predictions accurately match the velocity data for the entire saccade, including saccades with dynamic or glissadic overshoot. We were unable to generate saccades with postsaccade behavior based on just the timing of the antagonist step, but we needed the PIRB to generate saccades with dynamic or glissadic overshoot.

Figures 13.55 and 13.56 summarize the characteristics of the 127 saccades collected from the three human subjects. The number of saccades with a glissade is larger than the number of normal saccades or those with a dynamic overshoot. Additionally, the incidence of dynamic overshoot decreases as saccade size increases. As shown, saccades with a dynamic overshoot typically have larger rebound burst magnitude than those with a glissade or with

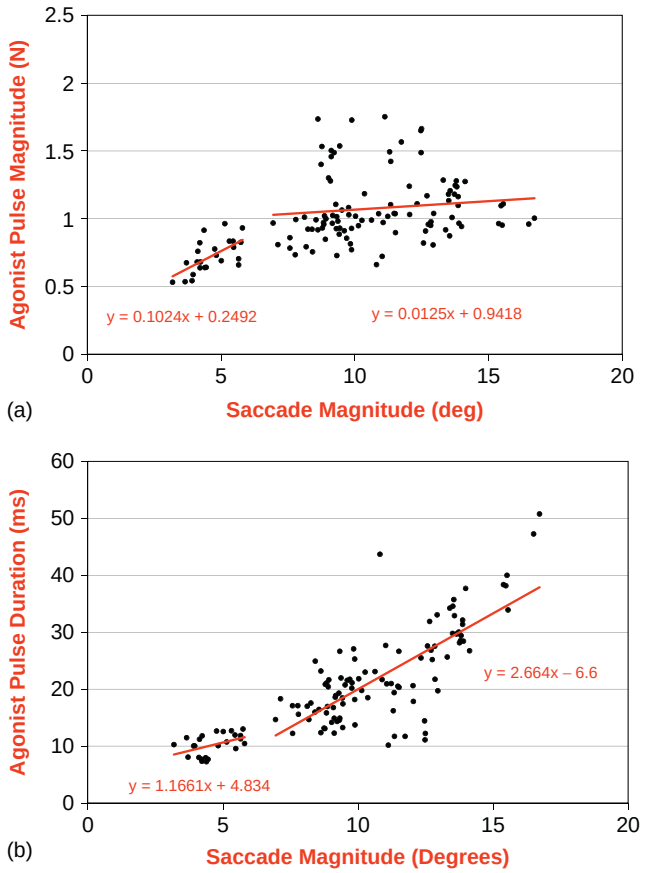


FIGURE 13.54 Agonist pulse magnitude (a) and duration (b) as functions of saccade magnitude.

normal characteristics. The antagonist onset delay varies from 3 ms to approximately 25 ms. With a larger rebound burst, the onset delay is typically shorter for each type of saccade.

An inherent coordination error exists between the return to tonic firing levels in the abducens and oculomotor motoneurons during the completion of a saccade. During an abducting saccade, ipsilateral abducens motoneurons fire without inhibition and oculomotor motoneurons are inhibited during the pulse phase. Because the IBN inhibits antagonist motoneurons, resumption of tonic firing and PIRB activity in the motoneurons does not begin until shortly after the ipsilateral IBNs cease firing. This same delay exists in the abducens motoneurons for adducting saccades.

There are significantly more internuclear neurons between the contralateral EBN and the TN and the ipsilateral oculomotor motoneurons (antagonist neurons during an abducting saccade) than the ipsilateral EBN and TN and ipsilateral abducens motoneurons (antagonist neurons during an adducting saccade). Due to the greater number of internuclear neurons operating during an abducting saccade, a longer time delay exists before the resumption of activity in the oculomotor motoneurons after the pulse phase for abducting than adducting saccades.

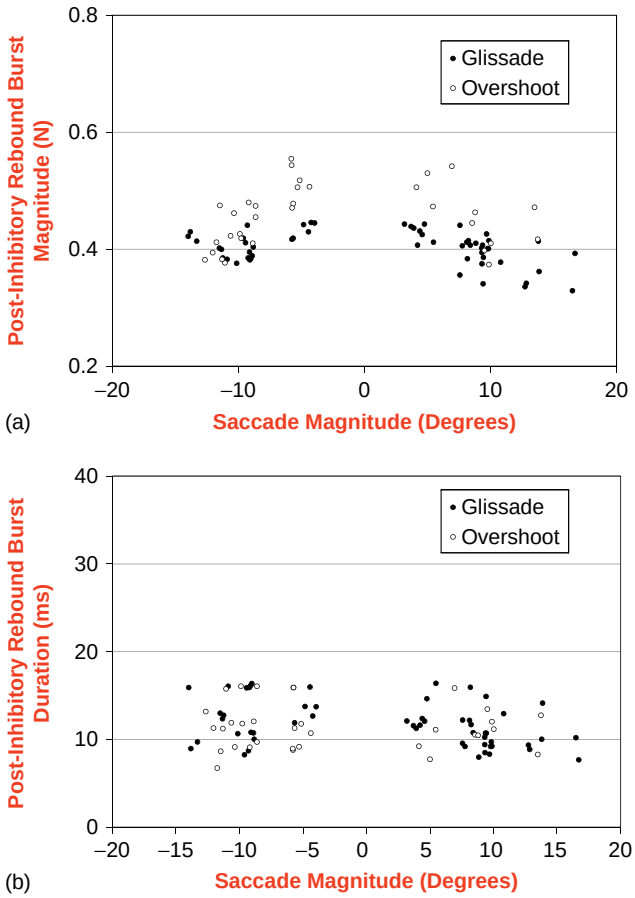


FIGURE 13.55 Postinhibitory rebound burst magnitude (a) and duration (b) as functions of saccade magnitude.

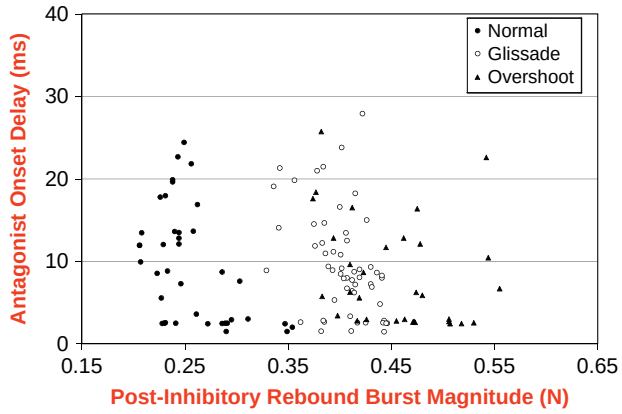


FIGURE 13.56 Postsaccade phenomena involving normal, glissade, and dynamic overshoot saccades.

Since the time delay before the resumption of activity in the oculomotor motoneurons after the pulse phase of a saccade is greater for abducting saccades than with adducting saccades, the incidence of saccades with dynamic overshoot should be greater for abducting saccades than adducting saccades. This is precisely what is observed in saccadic eye movement recordings; most saccades with dynamic overshoot occur in the abducting direction. Additionally, because the contralateral TN's firing rate decreases as ipsilateral saccade amplitude increases, the rate of dynamic overshoot decreases, since fewer saccades have sufficiently high PIRB magnitudes. This is also what is observed in saccadic eye movement recordings.

It is possible for a normal saccade to have a small PIRB as long as the onset delay is small. As the onset delay increases, the PIRB must decrease or a saccade with dynamic or glissadic overshoot occurs.

13.8.6 Time-Optimal Controller

The general principle for a time-optimal controller for the horizontal saccade system is that the eyes reach their destination in minimum time that involves over 1,000 neurons. Each neuron contributes to the neural input to the oculomotor plant. Enderle and Wolfe [16] described the time-optimal control of saccadic eye movements with a single switch-time using a linear homeomorphic oculomotor plant for the lateral and medical rectus muscles. Here, we reexamine the 1987 study using the updated oculomotor plant and a time-optimal controller constrained by a more realistic pulse-slide-step motoneuron stimulation of the agonist muscle with a pause and step in the motoneuron stimulation of the antagonist muscle, and physiological constraints.

The time-optimal controller proposed here has a firing rate in individual neurons that is maximal during the agonist pulse and independent of eye orientation, while the antagonist muscle is inhibited. We refer to maximal firing in the neuron as the intent of the system, which, because of biophysical properties of the neuron membrane, slowly decays over time, as described in Enderle [11]. The type of time-optimal controller described here is more complex than the one in 1987 due to physiological considerations. The time-optimal controller operates in two modes: one for small saccades and one for large saccades.

The duration of small saccades has been reported as approximately constant [23, 50, Enderle and Zhou, 2010b], and also as a function of saccade amplitude (e.g., [4]). Estimating the saccade start and end time is quite difficult because it is contaminated by noise. Enderle and coworkers have used a Kaiser filter to reduce the impact of noise, which others may not have implemented, and possibly introduced a difference in results. Moreover, synchrony of firing will have a greater impact on the start time for small saccades than larger saccades, since the beginning of the saccades is much more drawn out, making detection more difficult. In our analysis, a regression fit for the data is carried out in two intervals: one between 3 and 7°, and one for those greater than 7°. Our results indicate an approximately constant duration for small saccades and a duration that increases with saccade size for large saccades. Other investigators have used a single interval for the regression fit to a straight line or a nonlinear function. It is possible that using the technique used here will result in a similar conclusion to ours. Since we did not analyze saccades less than 3°, judgment on saccade duration in this interval is delayed and supports future investigation.

We propose that there is a minimum time period that EBNs can be switched on and off and that this is a physical constraint of the system. As shown in Figure 13.53c, small saccades have approximately the same duration of 44 ms, and they do not significantly change as a function of saccade magnitude. Also note that there is randomness in the response, where saccades with large pulse magnitudes are matched with shorter durations, and vice versa. As the saccade size increases for small saccades, we propose that additional neurons are added to the agonist neural input up to 7°, where above this, all neurons are engaged.

In our model, we sum the input of all active motoneurons into the firing of a single neuron. Thus, as the magnitude of the saccades increases, the firing rate of the single neuron in our model increases up to 7°, after which it is maximal, since all neurons are firing. Keep in mind, however, that the firing rate of a real neuron is maximal and does not change as a function of saccade magnitude, as is easily seen in Figure 4 in Robinson [42] and Figure 2 in Van Gisbergen and coworkers [46]. The overall neural input for the agonist pulse is given by

$$N_{ag} = \begin{cases} N(\theta_T) N_{ag_i} & \theta < 7^\circ \\ N_{ag_{\max}} & \theta \geq 7^\circ \end{cases} \tag{13.59}$$

where $N(\theta_T)$ is the number of neurons firing for a saccade of θ_T degrees, N_{ag_i} is the contribution from an individual neuron, and $N_{ag_{\max}}$ is the combined input from all neurons. For small saccades, the commencement of firing of the individual neurons, or synchrony of firing, has a great impact on the overall neural input, since the period of firing during the pulse is small (10 ms for the estimate in Figure 13.54b). Randomness in the start time among the active neurons means that the beginning of the saccade is more drawn out than if they all started together. For smaller saccades, this may result in an incorrect start time, which then effects the duration. Any lack of synchrony can cause the overall agonist input to be smaller; this is a much larger factor for a small saccade than a large saccade, since the pulse duration is much larger. It is very likely that during a saccade, neurons do not all commence firing at the same instant. This is seen in Figure 13.54b, where there is a small slope to the regression fit.

Above 7°, the magnitude of the saccade is dependent on the duration of the agonist pulse with all neurons firing maximally. The agonist pulse magnitude as shown in Figure 13.54a is approximately a constant according to the regression fit. The duration of the agonist pulse increases as a function of saccade magnitude, as shown in Figure 13.54b.

The saccade controller described here is a time-optimal controller that differs from the one describe by Enderle and Wolfe [16] because of the physiology of the system. Active neurons during the pulse phase of the saccade all fire maximally. For saccades greater than 7°, this is the same time-optimal controller described earlier by Enderle and Wolfe [16]. For saccades from 3 to 7°, the system is constrained by a minimum duration of the agonist pulse; saccade magnitude is dependent on the number of active neurons, all firing maximally, consistent with physiological evidence. In terms of control, it is far easier to operate the system for small saccades based on the number of active neurons firing maximally, rather than adjusting the firing rate for all neurons as a function of saccade magnitude as proposed by others. Thus, the system described here is still time-optimal based on physiological constraints.

Generally, saccades recorded for any size magnitude are extremely variable, with wide variations in the latent period, time to peak velocity, peak velocity, and duration. Furthermore, this variability is well coordinated by the neural controller. Saccades with lower peak velocity are matched with longer saccade durations, and saccades with higher peak velocity are matched with shorter saccade durations. Thus, saccades driven to the same destination usually have different trajectories.

13.9 SACCADE NEURAL PATHWAYS

Clinical evidence, lesion, and stimulation studies all point toward the participation of vitally important neural sites in the control of saccades, including the cerebellum, superior colliculus (SC), thalamus, cortex, and other nuclei in the brain stem, and that saccades are driven by two parallel neural networks [10, 11, 15, 18, 50]. From each eye, the axons of retinal ganglion cells exit and join other neurons to form the optic nerve. The optic nerves from each eye then join at the optic chiasm, where fibers from the nasal half of each retina cross to the opposite side. Axons in the optic tract synapse in the lateral geniculate nucleus (a thalamic relay) and continue to the visual cortex. This portion of the saccade neural network is concerned with the recognition of visual stimuli. Axons in the optic tract also synapse in the SC. This second portion of the saccade neural network is concerned with the location of visual targets and is primarily responsible for goal-directed saccades.

Saccadic neural activity of the SC and cerebellum, in particular, have been identified as the saccade initiator and terminator, respectively, for a goal-directed saccade. The impact of the frontal eye field and the thalamus, while very important, have less important roles in the generation of goal-directed saccades to visual stimuli. The frontal eye fields are primarily concerned with voluntary saccades, and the thalamus appears to be involved with corrective saccades. Figure 13.57 shows the important sites for the generation of a conjugate goal-directed horizontal saccade in both eyes. Each of the sites and connections detailed in Figure 13.57 is fully supported by physiological evidence. Some of these neural sites will be briefly described herein, with abbreviations provided in Table 13.2.

The Paramedian Pontine Reticular Formation (PPRF) has neurons that burst at frequencies up to 1,000 Hz during saccades and are silent during periods of fixation, and neurons that fire tonically during periods of fixation. Neurons that fire at steady rates during fixation are called tonic neurons (TN) and are responsible for holding the eye steady. The TN firing rate depends on the position of the eye (presumably through a local integrator type network). The TNs are thought to provide the step component to the motoneuron. There are two types of burst neurons in the PPRF: the long-lead burst neuron (LLBN) and a medium-lead burst neuron (MLBN). During periods of fixation, these neurons are silent. The LLBN burst at least 12 ms before a saccade and the MLBN burst less than 12 ms (typically 6–8 ms) before the saccade. The MLBNs are connected monosynaptically with the Abducens Nucleus.

The two types of neurons in the MLBN are the excitatory burst neurons (EBN) and the inhibitory burst neurons (IBN). The EBN and IBN labels describe the synaptic activity upon the motoneurons; the EBNs excite and are responsible for the burst firing, and the IBNs inhibit and are responsible for the pause. A mirror image of these neurons exists on both sides of the midline. The IBNs inhibit the EBNs on the contralateral side.

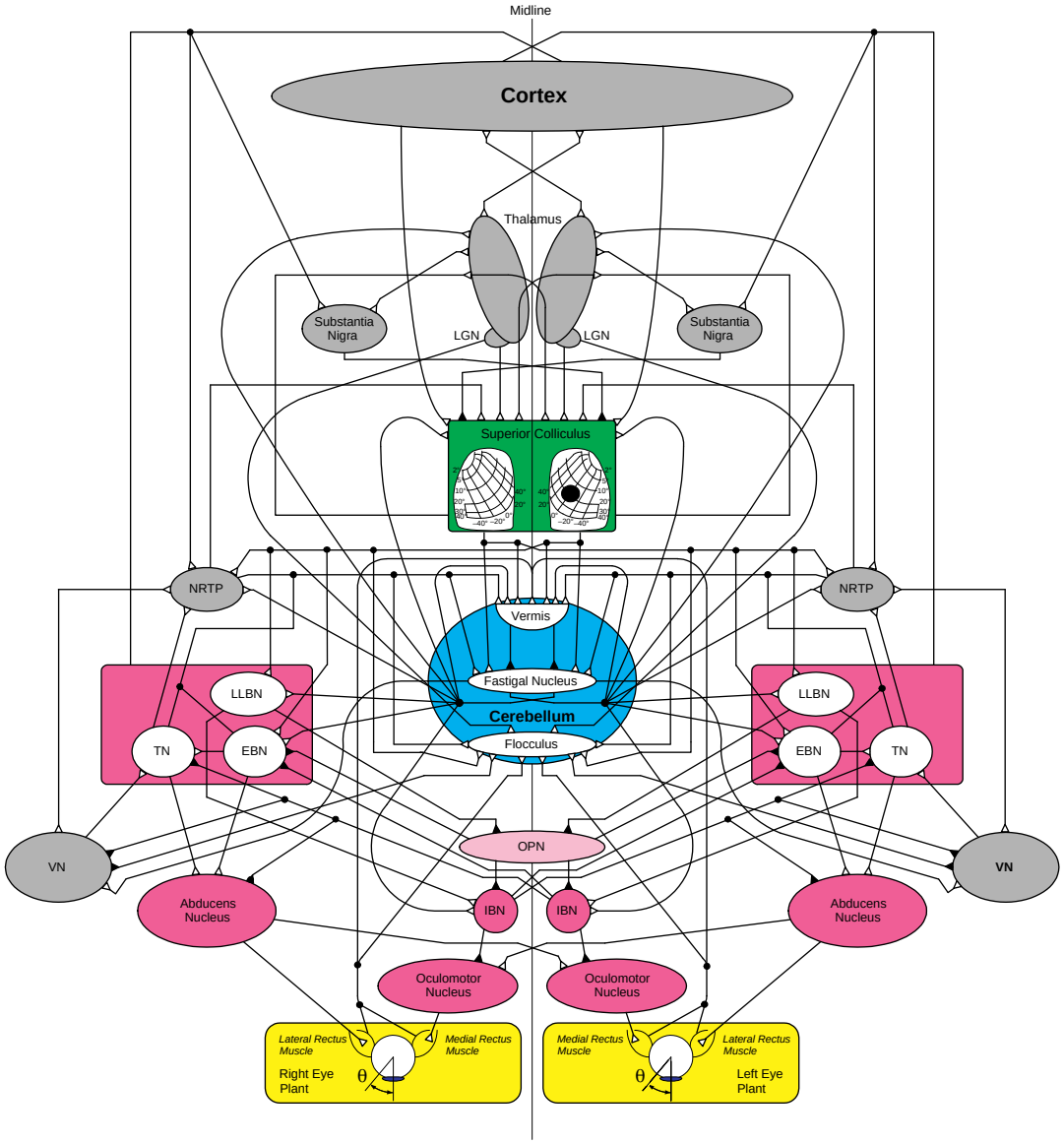


FIGURE 13.57 The important sites for the generation of a conjugate horizontal saccade in both eyes. Excitatory inputs are shown with  , and inhibitory inputs are shown with  . Consistent with current knowledge, the left and right structures of the neural circuit model are maintained. Since interest is in goal-directed visual saccades, the cortex has not been partitioned into the frontal eye field and posterior eye field (striate, prestriate, and inferior parietal cortices).

TABLE 13.2 List of Abbreviations

Neural Site	Abbreviation
Excitatory Burst Neuron	EBN
Fastigial Nucleus	FN
Inhibitory Burst Neuron	IBN
Long Lead Burst Neuron	LLBN
Medium Lead Burst Neuron	MLBN
Nucleus Reticularis Tegmenti Pontis	NRTP
Paramedian Pontine Reticular Formation	PPRF
Omnipause Neuron	OPN
Substantia Nigra	SN
Superior Colliculus	SC
Tonic Neuron	TN
Vestibular Nucleus	VN

Also within the brain stem is another type of saccade neuron called the omnipause neuron (OPN). The OPN fires tonically at approximately 200 Hz during periods of fixation and is silent during saccades. The OPN stops firing approximately 10–12 ms before a saccade and resumes tonic firing approximately 10 ms before the end of the saccade. The OPNs are known to inhibit the MLBNs and are inhibited by the LLBNs. The OPN activity is responsible for the precise timing between groups of neurons that causes a saccade.

The SC contains two major functional divisions: a superficial division and an intermediate or deep. Inputs to the superficial division are almost exclusively visual and originate from the retina and the visual cortex. The deep layers provide a site of convergence for sensory signals from several modalities and a source of efferent commands for initiating saccades. The SC is the initiator of the saccade and is thought to translate visual information into motor commands.

The deep layers of the SC initiate a saccade based on the distance between the current position of the eye and the desired target. The neural activity in the SC is organized into movement fields that are associated with the direction and saccade amplitude and does not involve the initial position of the eyeball whatsoever. The movement field is shown in Figure 13.58 for a 20° and 2° saccade. Neurons that are active during a particular saccade are shown as the dark circle. Notice that the movement field for the 2° saccade is smaller than that of the 20° saccade.

Active neurons in the deep layers of the SC generate a high-frequency burst of activity beginning 18–20 ms before a saccade and ending sometime toward the end of the saccade; the exact timing for the end of the burst firing is quite random and can occur slightly before or slightly after the saccade ends. Neurons discharging for small saccades have smaller movement fields, and those for larger saccades have larger movement fields. All of the movement fields are connected to the same set of LLBNs.

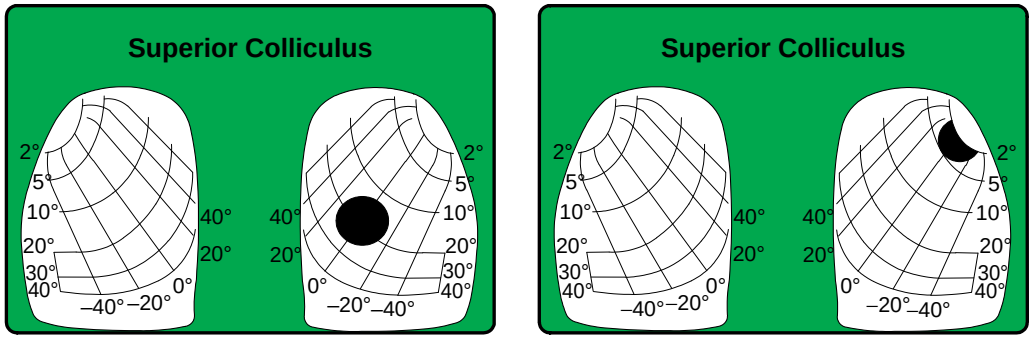


FIGURE 13.58 A detailed view of the retinotopic mapping over of the superior colliculus for a (a) 20° movement and a (b) 2° movement. Notice the locus of points for the 2° movement is smaller than that for the 20° movement, which implies that fewer neurons are firing for the smaller movement. The movement fields within the superior colliculus also reflect that the number of neurons firing for saccades less than 7° is smaller than those firing for saccades greater than 7°. For saccades above 7°, the movement field is approximately constant.

The cerebellum is responsible for the coordination of movement and is composed of a cortex of gray matter, internal white matter, and three pairs of deep nuclei: fastigial nucleus (FN), the interposed and globose nucleus, and dentate nucleus. The deep cerebellar nuclei and the vestibular nuclei transmit the entire output of the cerebellum. Output of the cerebellar cortex is carried through Purkinje cells. Purkinje cells send their axons to the deep cerebellar nuclei and have an inhibitory effect on these nuclei. The cerebellum is involved with both eye and head movements, and both tonic and phasic activities are reported in the cerebellum. The cerebellum is not directly responsible for the initiation or execution of a saccade but contributes to saccade precision. Sites within the cerebellum important for the control of eye movements include the oculomotor vermis, FN, and the flocculus. Consistent with the operation of the cerebellum for other movement activities, the cerebellum is postulated here to act as the coordinator for a saccade and as a precise gating mechanism.

The cerebellum is included in the saccade generator as a time-optimal gating element, using three active sites during a saccade: the vermis, FN, and flocculus. The vermis is concerned with the absolute starting position of a saccade in the movement field and corrects control signals for initial eye position. Using proprioceptors in the oculomotor muscles and an internal eye position reference, the vermis is aware of the current position of the eye. The vermis is also aware of the signals (dynamic motor error) used to generate the saccade via the connection with the Nucleus Reticularis Tegmenti Pontis (NRTP) and the SC.

With regard to the oculomotor system, the cerebellum has inputs from SC, lateral geniculate nucleus (LGN), oculomotor muscle proprioceptors, and striate cortex via NRTP. The cerebellum sends inputs to the NRTP, LLBN, EBN, VN, thalamus, and SC. The oculomotor vermis and fastigial nuclei are important in the control of saccade amplitude, and the flocculus, perihypoglossal nuclei of the rostral medulla, and possibly the pontine and mesencephalic reticular formation are thought to form the integrator within the cerebellum. One important function of the flocculus may be to increase the time constant of the neural integrator for saccades starting at locations different from primary position.

The FN receives input from the SC, as well as other sites. The output of the FN is excitatory and projects ipsilaterally and contralaterally, as shown in Figure 13.58. During fixation, the FN fires tonically at low rates. Twenty ms prior to a saccade, the contralateral FN bursts, and the ipsilateral FN pauses and then discharges with a burst. The pause in ipsilateral firing is due to Purkinje cell input to the FN. The sequential organization of Purkinje cells along beams of parallel fibers suggests that the cerebellar cortex might function as a delay, producing a set of timed pulses that could be used to program the duration of the saccade. If one considers nonprimary position saccades, different temporal and spatial schemes, via cerebellar control, are necessary to produce the same-size saccade. It is postulated here that the cerebellum acts as a gating device that precisely terminates a saccade based on the initial position of the eye in the orbit.

To execute a saccade, a sequence of complex activities takes place within the brain, beginning from the detection of an error on the retina to the actual movement of the eyes. A saccade is directly caused by a burst discharge (pulse) from motoneurons stimulating the agonist muscle and a pause in firing from motoneurons stimulating the antagonist muscle. During periods of fixation, the motoneurons fire at a rate necessary to keep the eye stable (step). The pulse discharge in the motoneurons is caused by the EBN, and the step discharge is caused by the TN in the PPRF.

Consider the saccade network in Figure 13.57 that is programmed to move the eyes 20°. Qualitatively, a saccade occurs according to the following sequence of events:

1. The deep layers of the SC initiate a saccade based on the distance between the current position of the eye and the desired target. The neural activity in the SC is organized into movement fields that are associated with the direction and saccade amplitude, and it does not involve the initial position of the eyeball whatsoever. Neurons active in the SC during this particular saccade are shown as the dark circle, representing the desired 20° eye movement. Active neurons in the deep layers of the SC generate an irregular high-frequency burst of activity that changes over time, beginning 18–20 ms before a saccade and ending sometime toward the end of the saccade; the exact timing for the end of the SC firing is quite random and can occur either before or after the saccade ends.
2. The ipsilateral LLBN and EBN are stimulated by the contralateral SC burst cells. The LLBN then inhibits the tonic firing of the OPN. The contralateral FN also stimulates the ipsilateral LLBN and EBN.
3. When the OPN cease firing, the MLBN (EBN and IBN) is released from inhibition. Some report that the ipsilateral EBN is probably not stimulated by the SC [25, 39]. This conflict doesn't impact our model as we propose the stimulation of the EBN by other sites does not reflect the firing rate of the EBN, but that the EBN fire autonomously given weak stimulation.
4. The ipsilateral IBN is stimulated by the ipsilateral LLBN and the contralateral FN of the cerebellum. When released from inhibition, the ipsilateral EBN responds with a PIRB for a brief period of time. The EBN when stimulated by the contralateral FN (and perhaps the SC) enables a special membrane property that causes a high-frequency burst that decays slowly until inhibited by the contralateral IBN. The IBN may also have the same type of special membrane properties.
5. The burst firing in the ipsilateral IBN inhibits the contralateral EBN and Abducens Nucleus, and the ipsilateral Oculomotor Nucleus.

6. The burst firing in the ipsilateral EBN causes the burst in the ipsilateral Abducens Nucleus, which then stimulates the ipsilateral lateral rectus muscle and the contralateral Oculomotor Nucleus. With the stimulation of the ipsilateral lateral rectus muscle by the ipsilateral Abducens Nucleus and the inhibition of the ipsilateral rectus muscle via the Oculomotor nucleus, a saccade occurs in the right eye. Simultaneously, the contralateral medial rectus muscle is stimulated by the Contralateral Oculomotor Nucleus, and with the inhibition of the contralateral lateral rectus muscle via the Abducens Nucleus, a saccade occurs in the left eye. Thus, the eyes move conjugately under the control of a single drive center.
7. At the termination time, the cerebellar vermis, operating through the Purkinje cells, inhibits the contralateral FN and stimulates the ipsilateral FN. Some of the stimulation of the ipsilateral LLBN and IBN is lost because of the inhibition of the contralateral FN. The ipsilateral FN stimulates the contralateral LLBN, EBN, and IBN. Further simulation of the contralateral IBN occurs from the contralateral LLBN. The contralateral EBN then stimulates the contralateral Abducens Nucleus. The contralateral IBN then inhibits the ipsilateral EBN, TN, Abducens Nucleus, and contralateral Oculomotor Nucleus. With this inhibition, the stimulus to the agonist muscles ceases. In most saccades, the SC continues to fire even though the saccade has ended.
8. The ipsilateral FN stimulation of the contralateral EBN allows for modest bursting in the contralateral EBN (while still being inhibited by the ipsilateral IBN whose activity has been reduced). This then stimulates the contralateral Abducens Nucleus and ipsilateral Oculomotor Nucleus. With the stimulation from the contralateral EBN through the contralateral Abducens Nucleus and ipsilateral Oculomotor Nucleus, the antagonist muscles fire, causing the antagonist muscles to contract. Once the SC ceases firing, the stimulus to the LLBN stops, allowing the resumption of OPN firing that inhibits the ipsilateral and contralateral MLBN and the saccade ends.

13.10 SYSTEM IDENTIFICATION

In traditional applications of electrical, mechanical, and chemical engineering, the main application of modeling is as a design tool to allow the efficient study of the effects of parametric variation on system performance as a means of cost containment. In modeling physiological systems, the goal is not to design a system but to identify the parameters and structure of the system. Ideally, the input and the output of the physiological system are known, and some information about the internal dynamics of the system is available (Figure 13.59). In many cases, either the input or output is not measurable or observable but is estimated from a remote signal, and no information about the system is known. System identification is the process of creating a model of a system and estimating the

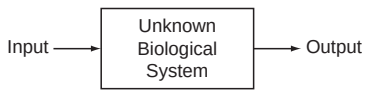


FIGURE 13.59 Block diagram of a typical physiological system without feedback.

parameters of the model. This section introduces the concept of system identification in both the frequency and time domain.

A variety of signals is available to the biomedical engineer, as described earlier. Those produced by the body include action potentials, EEGs, EKGs, EMGs, EOG, and pressure transducer output. Additional signals are available through ultrasound, x-ray tomography, MRI, and radiation. From these signals a model is built and parameters estimated according to the modeling plan in Figure 13.1. Before work on system identification begins, understanding the characteristics of the input and output signals is important as described in Chapter 10—that is, knowing the voltage range, frequency range, whether the signal is deterministic or stochastic, and if coding (i.e., neural mapping) is involved. Most biologically generated signals are low frequency and involve some coding. For example, EEGs have an upper frequency of 30 Hz and eye movements have an upper frequency under 100 Hz. The saccadic system uses neural coding that transforms burst duration into saccade amplitude, as described earlier in this chapter. After obtaining the input and output signals, these signal must be processed. A fundamental block is the amplifier, which is characterized by its gain and frequency, as described in Chapter 8. Note that the typical amplifier is designed as a low-pass filter (LPF), since noise amplification is not desired. Interestingly, most amplifiers have storage elements (i.e., capacitors and inductors), so the experimenter must wait until the transient response of the amplifier has been completed before any useful information can be extracted. An important point to remember is that the faster the cutoff of the filter, the longer the transient response of the amplifier.

In undergraduate classes, a system (the transfer function or system description) and input are usually provided and a response or output is requested. While this seems difficult, it is actually much easier than trying to determine the parameters of a physiological system when all that is known are the input (and perhaps not the direct input as described in the saccadic eye movement system) and noisy output characteristics of the model. In the ideal case, the desired result here is the transfer function, which can be determined from

$$H(s) = \frac{V_o(s)}{V_i(s)} \tag{13.60}$$

13.10.1 Classical System Identification

The simplest and most direct method of system identification is sinusoidal analysis. A source of sinusoidal excitation is needed that usually consists of a sine wave generator, a measurement transducer, and a recorder to gather frequency response data. Many measurement transducers are readily available for changing physical variables into voltages, as described in Chapter 9. Devices that produce the sinusoidal excitation are much more difficult to obtain and are usually designed by the experimenter. Recording the frequency response data can easily be obtained from an oscilloscope. Figure 13.60 illustrates the essential elements of sinusoidal analysis.

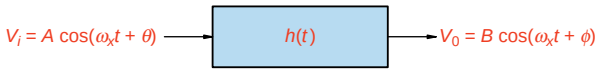


FIGURE 13.60 Impulse response block diagram.

The experiment to identify model parameters using sinusoidal analysis is simple to carry out. The input is varied at discrete frequencies over the entire spectrum of interest, and the output magnitude and phase are recorded for each input. To illustrate this technique, consider the following analysis. From Figure 13.60, it is clear that the transfer function is given by

$$H(j\omega) = \frac{V_o(j\omega)}{V_i(j\omega)} \tag{13.61}$$

The Fourier transform of the input is

$$V_i(j\omega) = F\{A\cos(\omega_x t + \theta)\} = A\int e^{-j\omega\lambda}\cos(\omega_x\lambda + \theta)d\lambda = A\int e^{-j\omega\left(-\frac{\theta}{\omega_x}+\tau\right)}\cos(\omega_x\tau)d\tau \tag{13.62}$$

by substituting $\lambda = \tau - \frac{\theta}{\omega_x}$. Factoring out the terms not involving τ, gives

$$V_i(j\omega) = Ae^{\frac{j\omega\theta}{\omega_x}}\int e^{-j\omega\tau}\cos\omega_x\tau\, d\tau = Ae^{\frac{j\omega\theta}{\omega_x}}[\pi\delta(\omega-\omega_x) + \pi\delta(\omega+\omega_x)] \tag{13.63}$$

Similarly,

$$V_o(j\omega) = Be^{\frac{j\omega\phi}{\omega_x}}[\pi\delta(\omega-\omega_x) + \pi\delta(\omega+\omega_x)] \tag{13.64}$$

According to Eq. (13.61)

$$H(j\omega) = \frac{V_o(j\omega)}{V_i(j\omega)} = \frac{B}{A}\frac{e^{\frac{j\omega\phi}{\omega_x}}}{e^{\frac{j\omega\theta}{\omega_x}}} = \frac{B}{A}e^{\frac{j\omega(\phi-\theta)}{\omega_x}} \tag{13.65}$$

At steady state with $\omega = \omega_x$, Eq. (13.65) reduces to

$$H(j\omega_x) = \frac{B}{A}e^{j(\phi-\theta)} \tag{13.66}$$

Each of these quantities in Eq. (13.66) is known (i.e., B, A, ϕ, and θ), so the magnitude and phase angle of the transfer function are also known. Thus, ω_x can be varied over the frequency range of interest to determine the transfer function.

In general, a transfer function, $G(s)|_{s=j\omega} = G(j\omega)$, is composed of the following terms:

1. Constant term K
2. M poles or zeros at the origin of the form $(j\omega)^M$
3. P poles of the form $\prod_{p=1}^{P}(1+j\omega\tau_p)$ or Z zeros of the form $\prod_{z=1}^{Z}(1+j\omega\tau_z)$. Naturally, the poles or zeros are located at $-\frac{1}{\tau}$.
4. R complex poles of the form $\prod_{r=1}^{R}\left(1+\left(\frac{2\zeta_r}{\omega_{n_r}}\right)j\omega+\left(\frac{j\omega}{\omega_{n_r}}\right)^2\right)$ or S zeros of the form $\prod_{s=1}^{S}\left(1+\left(\frac{2\zeta_s}{\omega_{n_s}}\right)j\omega+\left(\frac{j\omega}{\omega_{n_s}}\right)^2\right)$
5. Pure time delay $e^{-j\omega T_d}$

where M, P, Z, R, S, and T_d are all positive integers. Incorporating the preceding terms, the transfer function is written as

$$G(j\omega) = \frac{K \times e^{-j\omega T_d} \times \left(\prod_{z=1}^{Z}(1 + j\omega\tau_z)\right) \times \left(\prod_{s=1}^{S}\left(1 + \left(\frac{2\zeta_s}{\omega_{n_s}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_s}}\right)^2\right)\right)}{(j\omega)^M \times \left(\prod_{p=1}^{P}(1 + j\omega\tau_p)\right) \times \left(\prod_{r=1}^{R}\left(1 + \left(\frac{2\zeta_r}{\omega_{n_r}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_r}}\right)^2\right)\right)} \tag{13.67}$$

This equation is used as a template when describing the data with a model. To determine the value of the unknown parameters in the model, the logarithm and asymptotic approximations to the transfer function are used. In general, the logarithmic gain, in dB, of the transfer function template is

$$\begin{aligned} 20\log|G(j\omega)| = {} & 20\log K + 20\sum_{z=1}^{Z}\log|1 + j\omega\tau_z| + 20\sum_{s=1}^{S}\log\left|1 + \left(\frac{2\zeta_s}{\omega_{n_s}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_s}}\right)^2\right| \\ & - 20\log\left|(j\omega)^M\right| - 20\sum_{p=1}^{P}\log|1 + j\omega\tau_p| - 20\sum_{r=1}^{R}\log\left|1 + \left(\frac{2\zeta_r}{\omega_{n_r}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_r}}\right)^2\right| \end{aligned} \tag{13.68}$$

and the phase, in degrees, is

$$\begin{aligned} \phi(\omega) = {} & -\omega T_d + \sum_{z=1}^{Z}\tan^{-1}(\omega\tau_z) + \sum_{s=1}^{S}\tan^{-1}\left(\frac{2\zeta_s\omega_{n_s}\omega}{\omega_{n_s}^2 - \omega^2}\right) - M \times (90^\circ) \\ & - \sum_{p=1}^{P}\tan^{-1}(\omega\tau_p) - \sum_{r=1}^{R}\tan^{-1}\left(\frac{2\zeta_r\omega_{n_r}\omega}{\omega_{n_r}^2 - \omega^2}\right) \end{aligned} \tag{13.69}$$

where the phase angle of the constant is 0° and the magnitude of the time delay is 1. Evident from these expressions is that each term can be considered separately and added together to obtain the complete Bode diagram. The asymptotic approximations to the logarithmic gain for the poles and zeros are given by the following:

- **Poles at the origin**
  Gain: $-20\log|(j\omega)| = -20\log\omega$. The logarithmic gain at $\omega = 1$ is 0 (i.e., the line passes through 0 dB at $\omega = 1$ radian/s).
  Phase: $\phi = -90^\circ$. If there is more than one pole, the slope of the gain changes by M $\times$ (-20) and the phase by M $\times$ (-90°).

- **Pole on the real axis**

$$\textit{Gain: } -20\log|1 + j\omega\tau_p| = \begin{cases} 0 & \text{for } \omega < \dfrac{1}{\tau_p} \\ -20\log(\omega\tau_p) & \text{for } \omega \geq \dfrac{1}{\tau_p} \end{cases}$$

Phase: an asymptotic approximation to $-\tan^{-1}(\omega\tau_p)$ is drawn with a straight line from 0° at one decade below $\omega = \dfrac{1}{\tau_p}$ to −90° at 1 decade above $\omega = \dfrac{1}{\tau_p}$.

The pole is located at $-\dfrac{1}{\tau_p}$.

- **Zero on the real axis**

$$\textit{Gain: } 20\log|1 + j\omega\tau_z| = \begin{cases} 0 & \text{for } \omega < \dfrac{1}{\tau_z} \\ 20\log(\omega\tau_z) & \text{for } \omega \geq \dfrac{1}{\tau_z} \end{cases}$$

Phase: an asymptotic approximation to $\tan^{-1}(\omega\tau_z)$ is drawn with a straight line from 0° at one decade below $\omega = \dfrac{1}{\tau_z}$ to +90° at 1 decade above $\omega = \dfrac{1}{\tau_z}$.

The zero is located at $-\dfrac{1}{\tau_z}$.

- **Complex poles**

$$\textit{Gain: } -20\log\left|1 + \left(\frac{2\zeta_r}{\omega_{n_r}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_r}}\right)^2\right| = \begin{cases} 0 & \text{for } \omega < \omega_{\mathrm{n_r}} \\ -40\log\left(\dfrac{\omega}{\omega_{n_r}}\right) & \text{for } \omega \geq \omega_{\mathrm{n_r}} \end{cases}$$

A graph of the actual magnitude-frequency is shown in Figure 13.61, with $\omega_n = 1.0$ and ζ ranging from 0.05 to 1.0. Notice that as ζ decreases from 1.0, the magnitude peaks at correspondingly larger values. As ζ approaches zero, the magnitude approaches infinity at $\omega = \omega_{n_r}$. For values of $\zeta > 0.707$ there is no resonance.

Phase: depending on the value of ζ_r, the shape of the curve is quite variable but in general is 0° at one decade below $\omega = \omega_{n_r}$ and −180° at 1 decade above $\omega = \omega_{n_r}$.

A graph of the actual phase-frequency is shown in Figure 13.61 with $\omega_n = 1.0$ and ζ ranging from 0.05 to 1.0. Notice that as ζ decreases from 1.0, the phase changes more quickly from 0° to 180° over a smaller frequency interval.

The poles are located at $-\zeta_r\omega_{n_r} \pm j\omega_{n_r}\sqrt{1-\zeta_r^2}$.

- **Complex zeros**

$$\textit{Gain: } 20\log\left|1 + \left(\frac{2\zeta_s}{\omega_{n_s}}\right)j\omega + \left(\frac{j\omega}{\omega_{n_s}}\right)^2\right| = \begin{cases} 0 & \text{for } \omega < \omega_{\mathrm{n_s}} \\ 40\log\left(\dfrac{\omega}{\omega_{n_s}}\right) & \text{for } \omega \geq \omega_{\mathrm{n_s}} \end{cases}$$

Phase: depending on the value of ζ_s, the shape of the curve is quite variable but in general is 0° at one decade below $\omega = \omega_{n_s}$ and +180° at 1 decade above $\omega = \omega_{n_s}$.

Both the magnitude and phase follow the two previous graphs and discussion with regard to the complex poles with the exception that the slope is +40 dB/decade rather than –40 dB/decade.

The zeros are located at $-\zeta_s\omega_{n_s} \pm j\omega_{n_s}\sqrt{1-\zeta_s^2}$.

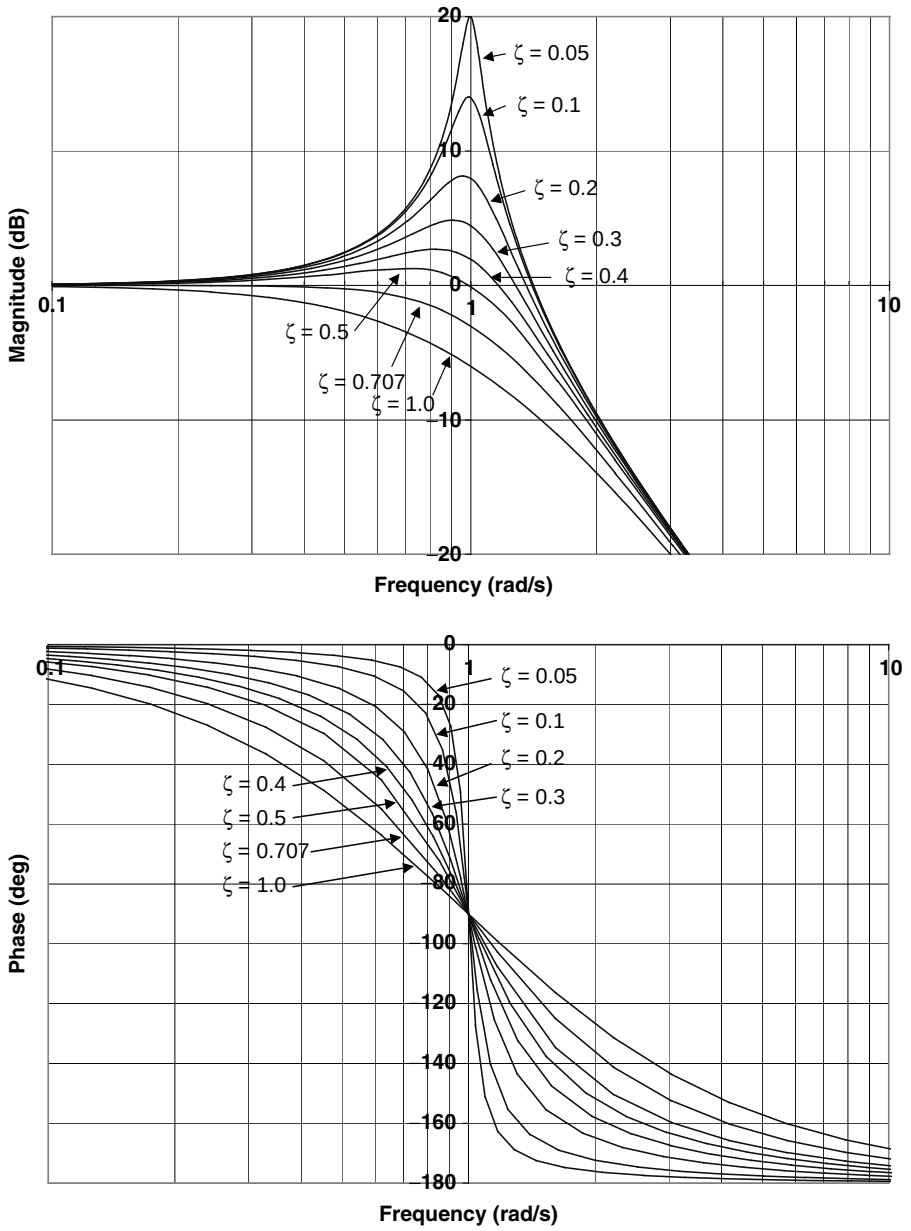


FIGURE 13.61 Bode plot of complex poles with $\omega_n = 1.0$ rad/s.

- **Time delay**
  Gain: 1 for all ω
  Phase: $-\omega T_d$
- **Constant *K***
  Gain: 20 log K
  Phase: 0

The frequency at which the slope changes in a Bode magnitude-frequency plot is called a break or corner frequency. The first step in estimating the parameters of a model involves identifying the break frequencies in the magnitude-frequency and/or phase-frequency responses. This simply involves identifying points at which the magnitude changes slope in the Bode plot. Poles or zeros at the origin have a constant slope of –20 or +20 dB/decade, respectively, from $-\infty$ to ∞. Real poles or zeros have a change in slope at the break frequency of –20 or +20 dB/decade, respectively. The value of the pole or zero is the break frequency. Estimating complex poles or zeros is much more difficult. The first step is to locate the break frequency ω_n, the point at which the slope changes by 40 dB/decade. To estimate ζ, use the actual magnitude-frequency (size of the peak) and phase-frequency (quickness of changing 180°) curves in Figure 13.61 to closely match the data.

The error between the actual logarithmic gain and straight-line asymptotes at the break frequency is 3dB for a pole on the real axis (the exact curve equals the asymptote − 3dB). The error drops to 0.3dB one decade below and above the break frequency. The error between the real zero and the asymptote is similar except the exact curve equals the asymptote + 3dB. At the break frequency for the complex poles or zeros, the error between the actual logarithmic gain and straight-line asymptotes depends on ζ and can be quite large as observed from Figure 13.61.

EXAMPLE PROBLEM 13.9

Sinusoids of varying frequencies were applied to an open-loop system, and the results in Table 13.3 were measured. Construct a Bode diagram and estimate the transfer function.

TABLE 13.3 Data for Example Problem 14.9

Frequency (radians/s)	0.01	0.02	0.05	0.11	0.24	0.53	1.17	2.6	5.7	12.7	28.1	62	137	304	453	672	1000
20 log|G| (dB)	58	51	44	37	30	23	15	6	−7	−20	−34	−48	−61	−75	−82	−89	−96
Phase (degrees)	−90	−91	−91	−93	−97	−105	−120	−142	−161	−171	−176	−178	−179	−180	−180	−180	−180

Solution

Bode plots of gain and phase versus frequency for the data given are shown in Figure 13.62. From the phase-frequency graph, it is clear that this system has two more poles than zeros because the phase angle approaches −180° as ω→∞. Also note that there is no peaking observed in the gain-frequency graph or sharp changes in the phase-frequency graph, so there does not appear to be any lightly damped ($\zeta < 0.5$) complex poles. However, this does not imply that there are no heavily damped complex poles at this time.

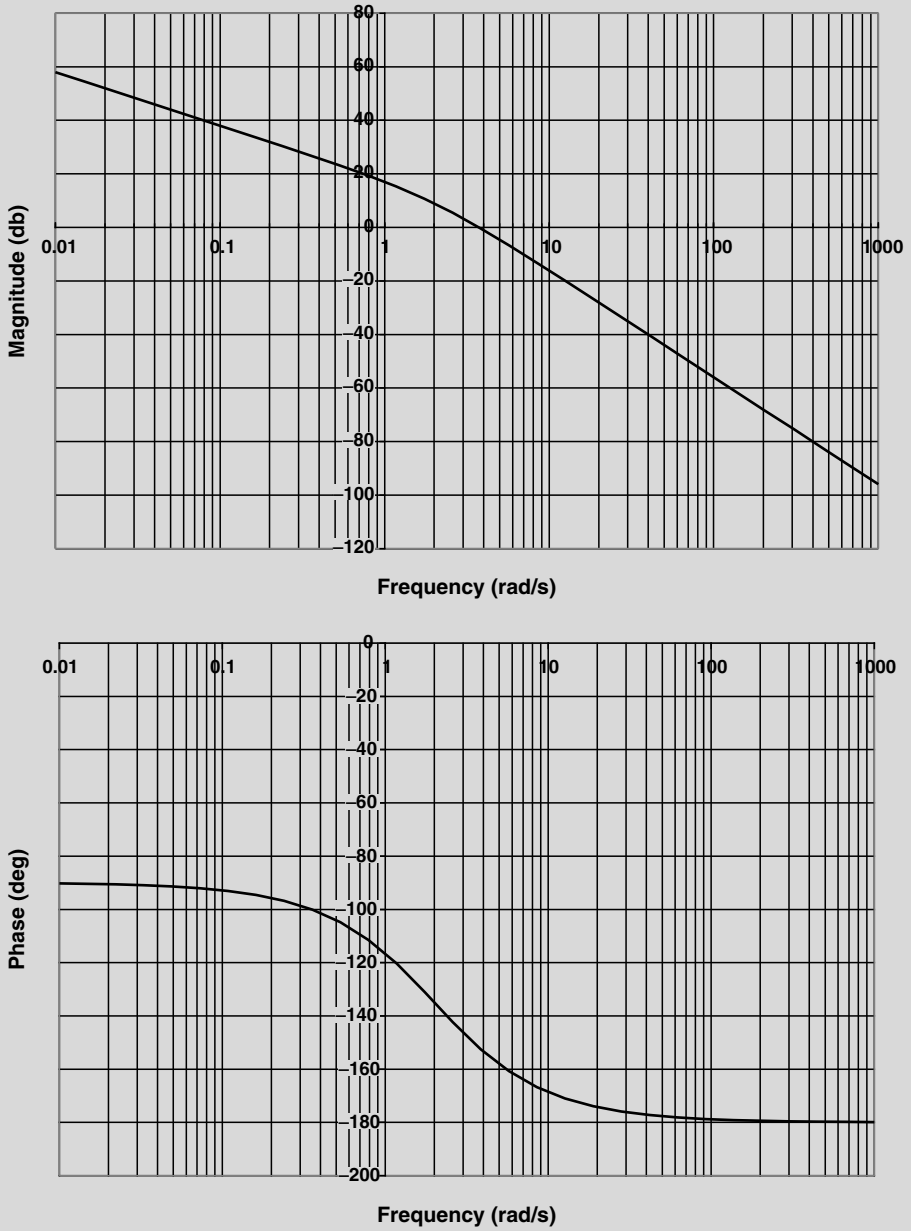


FIGURE 13.62 Data for Example Problem 13.9.

For frequencies in the range of 0.01 to 1 radians/s, the slope of the gain-frequency graph is −20 dB per decade. Since it is logical to assume that the slope remains at −20 dB for frequencies less than 0.01 radians/s, a transfer function with a pole at the origin provides such a response. If

Continued

possible, it is important to verify that the magnitude-frequency response stays at −20 dB per decade for frequencies less than 0.01 radians/s to the limits of the recording instrumentation.

For frequencies in the range of 2 to 1,000 radians/s, the slope of the gain-frequency graph is −40 dB per decade. Since a pole has already been identified in the previous frequency interval, it is reasonable to conclude that there is another pole in this interval. Other possibilities exist such as an additional pole and a zero that are closely spaced, or a complex pole and a zero. However, in the interests of simplicity and because these possibilities are not evident in the graphs, the existence of a pole is all that is required to describe the data.

Summarizing at this time, the model contains a pole at the origin and another pole somewhere in the region above 1 radians/s. It is also safe to conclude that the model does not contain heavily damped complex poles since the slope of the gain-frequency graph is accounted for completely.

To estimate the unknown pole, straight lines are drawn tangent to the magnitude-frequency curve as shown in Figure 13.63. The intersection of the two lines gives the break frequency for the pole at approximately 2 radians/s. Notice that the actual curve is approximately 3dB below the asymptotes as discussed previously and observed in the following figure.

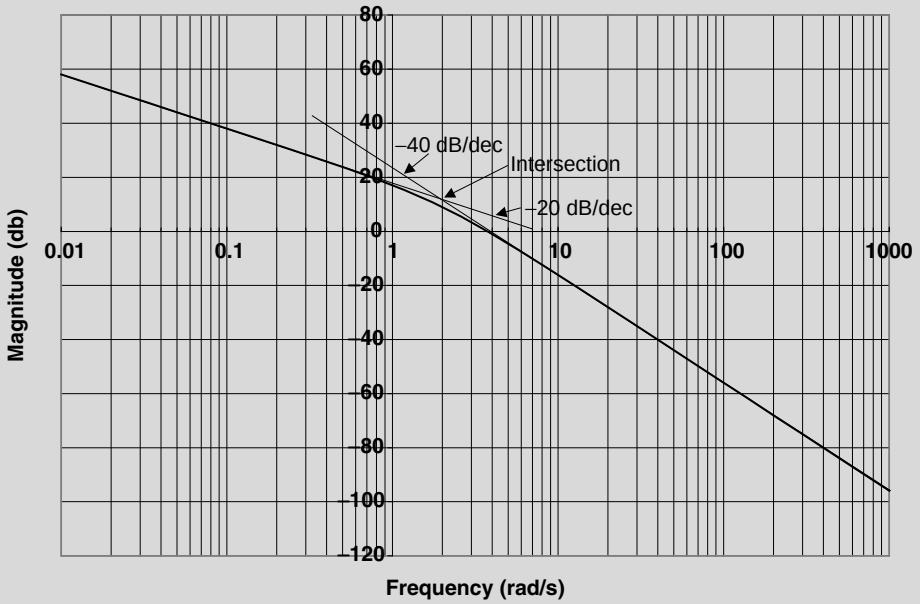


FIGURE 13.63 Straight-line approximations for Example Problem 13.9.

The model developed thus far is

$$G(j\omega) = \frac{1}{jw\left(\frac{j\omega}{2} + 1\right)}$$

To determine the constant K, the magnitude at $\omega = 1$ radian/s is investigated. The pole at zero contributes a value of 0 toward the logarithmic gain. The pole at 2 contributes a value of –1 dB ($-20\log\left|\frac{j\omega}{2}+1\right| = -20\log\left(\sqrt{\left(\frac{1}{2}\right)^2+1^2}\right) = -1$ dB) toward the logarithmic gain. The reason that the contribution of the pole at 2 was computed was that the point 1 radian/s was within the range of ± a decade of the break frequency. At $\omega = 1$ radian/s, the nonzero terms are

$$20\log|G(\omega)| = 20\log K - 20\log\left|\frac{j\omega}{2}+1\right|$$

From the gain-frequency graph, $20\log|G(\omega)| = 17$ dB. Therefore,

$$17 = 20\log K - 1$$

or $K = 8$. The model now consists of

$$G(j\omega) = \frac{8}{jw\left(\frac{j\omega}{2}+1\right)}$$

The last term to investigate is whether there is a time delay in the system. At the break frequency $\omega = 2.0$ radians/s, the phase angle from the current model should be

$$\phi(\omega) = -90 - \tan^{-1}\left(\frac{\omega}{2}\right) = -90 - 45 = -135^\circ$$

This value is approximately equal to the data, and thus there does not appear to be a time delay in the system.

Example Problem 13.9 illustrated a process of thinking in determining the structure and parameters of a model. Carrying out an analysis in this fashion on complex systems is extremely difficult, if not impossible. Software packages that automatically carry out estimation of poles, zeros, a time delay, and a gain of a transfer function from data are available, such as the System Identification toolbox in MATLAB. There are also other programs that provide more flexibility than MATLAB in analyzing complex systems, such as the FORTRAN program written by Seidel [43].

There are a variety of other inputs that one can use to stimulate the system to elicit a response. These include such transient signals as a pulse, step, and ramp, and noise signals such as white noise and pseudo random binary sequences. The reason these other techniques might be of interest is that not all systems are excited via sinusoidal input. One such system is the fast eye movement system. Here we typically use a step input to analyze the system (see [17]).

In analyzing the output data obtained from step input excitation to determine the transfer function, we use a frequency response method using the Fourier Transform and the fast Fourier Transform. The frequency response of the input is known. The frequency response of the output is calculated via a numerical algorithm called the fast Fourier Transform (FFT). To calculate the FFT of the output, the data must first be digitized using an A/D converter and stored in disk memory. Care must be taken to antialias (low-pass) filter

any frequency content above the highest frequency of the signal, and sample at a rate of at least 2.5 times the highest frequency. The transfer function is then estimated.

13.10.2 Identification of a Linear First-Order System

The technique for solving for the parameters of a first-order system is presented here. Assume that the system of interest is a first-order system that is excited with a step input. The response to the input is

$$y(t) = \left(y_{ss} + Ke^{\frac{-t}{\tau}}\right)u(t) \tag{13.70}$$

where $K = -(y_{ss} - y(0))$ and is shown in Figure 13.64. Note that at $t = \tau$, the response is 63 percent of the way from the initial to the steady-state value. Similarly, at $t = 4\tau$, the response is 98 percent of the way from the initial to the steady-state value.

Suppose step input data are collected from an unknown first-order system shown in Figure 13.64. To fully describe the system, the parameters of Eq. (13.70) need to be estimated. One way to estimate the system time constant is from the initial slope of the response and a smoothed steady-state value (via averaging). That is, the time constant τ is found from

$$\dot{y}(t) = \frac{1}{\tau}(y_{ss} - y(0))e^{\frac{-t}{\tau}} \quad \rightarrow \quad \tau = \frac{y_{ss} - y(0)}{\dot{y}(t)}e^{\frac{-t}{\tau}}$$

At $t = 0$, $\tau = \dfrac{y_{ss} - y(0)}{\dot{y}(0)}$, where $\dot{y}(0)$ is the initial slope of the response. The equation for estimating τ is nothing more than the equation of a straight line. This technique is illustrated in Figure 13.65.

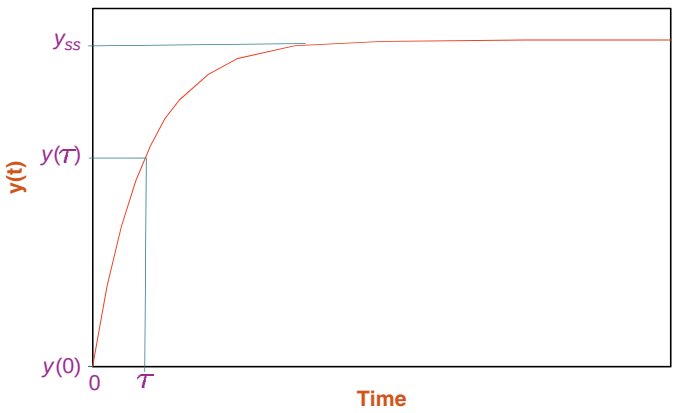


FIGURE 13.64 First-order system response to a step input.

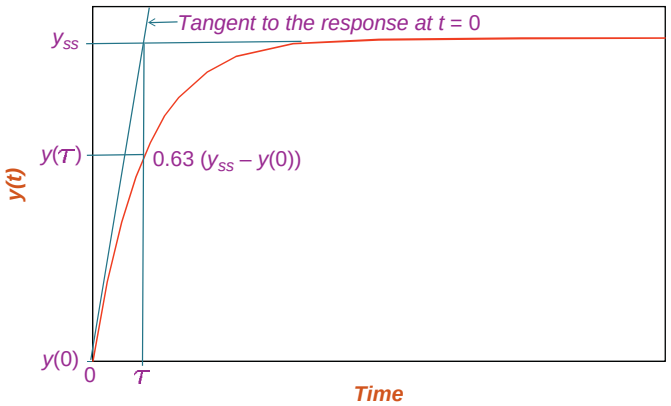


FIGURE 13.65 Estimating the time constant from the initial slope of the response.

EXAMPLE PROBLEM 13.10

The data in Table 13.4 were collected for the step response for an unknown first-order system. Find the parameters that describe the model.

TABLE 13.4 Data for Example Problem 13.10

t	00.0	00.05	00.1	00.15	00.2	0.25	00.3	00.4	00.5	00.7	1.0	1.5	2.0
$y(t)$	00.0	00.56	00.98	1.30	1.54	1.73	1.86	2.04	2.15	2.24	2.27	2.28	2.28

Solution

The model under consideration is described by Eq. (13.70), $y(t) = y_{ss} + Ke^{\frac{-t}{\tau}}u(t)$, with unknown parameters y_{ss}, K, and τ. The data are plotted in Figure 13.66, along with the tangent. Clearly $y_{ss} = 2.28$ and $K = -2.88$.

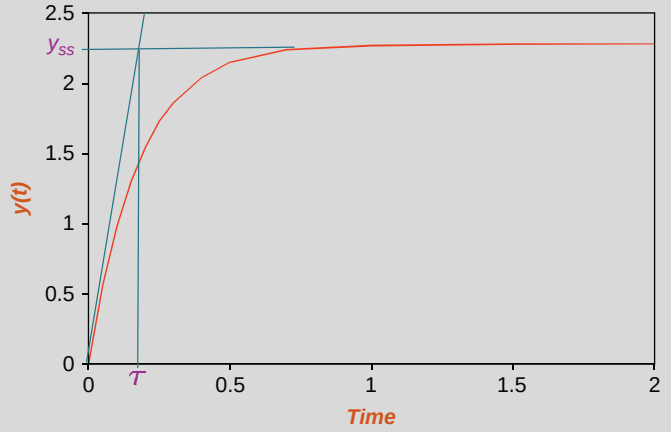


FIGURE 13.66 Illustration for Example Problem 13.10.

Continued

From the graph, $\tau = 0.17$ s. Thus, we have

$$y(t) = 2.88\left(1 - e^{\frac{-t}{0.17}}\right)u(t).$$

13.10.3 Identification of a Linear Second-Order System

Consider estimating the parameters of a second-order system, such as $M\ddot{y} + B\dot{y} + Ky = f(t)$. It is often convenient to rewrite the original differential equation in the standard form for ease in analysis as

$$\ddot{y} + 2\zeta\omega_n\dot{y} + \omega_n^2 y = f(t) \tag{13.71}$$

where ω_n is the undamped natural frequency and ζ is the damping ratio. The roots of the characteristic equation are

$$s_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1} = -\zeta\omega_n \pm j\omega_n\sqrt{1 - \zeta^2} = -\zeta\omega_n \pm j\omega_d \tag{13.72}$$

A system with $0 < \zeta < 1$ is called underdamped, with $\zeta = 1$ critically damped and with $\zeta > 1$ overdamped.

Step Response

The complete response of the system in Eq. (13.72) to a step input with arbitrary initial conditions is given in Table 13.5, where y_{ss} is the steady-state value of $y(t)$, and A_1, A_2, B_1, ϕ, C_1, and C_2 are the constants that describe the system evaluated from the initial conditions. For the underdamped system, one can estimate ζ and ω_n to a step input with magnitude γ from data as follows. Consider the step response in Figure 13.67. A suitable model for the system is a second-order underdamped model (i.e., $0 < \zeta < 1$), with solution

$$y\,(t) = C\left[1 + \frac{e^{-\zeta\omega_n t}}{\sqrt{1 - \zeta^2}}\cos\left(\omega_d t + \phi\right)\right] \tag{13.73}$$

where C is the steady-state response, y_{ss}, and $\omega_d = \omega_n\sqrt{1 - \zeta^2}$ and $\phi = \pi + \tan^{-1}\dfrac{-\zeta}{\sqrt{1 - \zeta^2}}$.

The following terms illustrated in Figure 13.67 are typically used to describe, quantitatively, the response to a step input:

TABLE 13.5 Step Response for a Second-Order System

Damping	Natural Response Equation
Overdamped	$y(t) = y_{ss} + A_1 e^{s_1 t} + A_2 e^{s_2 t}$
Underdamped	$y(t) = y_{ss} + B_1 e^{-\zeta\omega_n t}\cos(\omega_d t + \phi)$
Critically damped	$y(t) = y_{ss} + (C_1 + C_2 t)e^{-\zeta\omega_n t}$

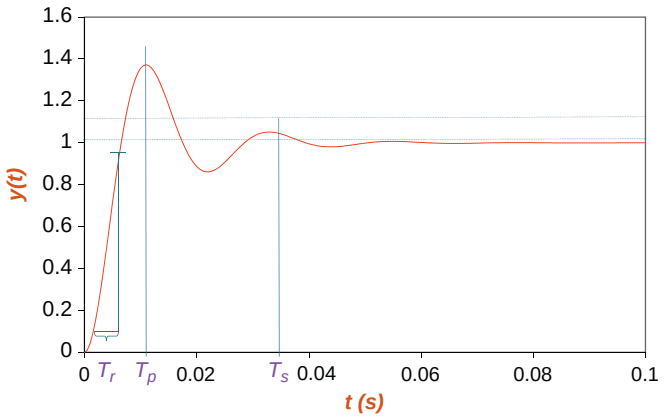


FIGURE 13.67 Sample second-order response with graphical estimates of rise time, time to peak overshoot, and settling time.

- Rise time, T_r. The time for the response to rise from 10 to 90 percent of steady state.
- Settling time, T_s. The time for the response to settle within ±5 percent of the steady-state value.
- Peak overshoot time, T_p. The time for the response to reach the first peak overshoot.

Using values graphically determined from the data for the preceding quantities, as shown in Figure 13.67, provides estimates for the parameters of the model. The time to peak overshoot, T_p, is found by first calculating

$$\frac{\partial y}{\partial t} = \frac{\gamma\, e^{-\zeta\omega_n t}}{K\sqrt{1-\zeta^2}}\left[-\zeta\omega_n \cos\left(\omega_d t + \phi\right) - \omega_d \sin\left(\omega_d t + \phi\right)\right] \tag{13.74}$$

using the chain rule on Eq. (13.74) and then determining T_p from $\frac{\partial y}{\partial t}\big|_{t=T_p} = 0$, yielding

$$T_p = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}} \tag{13.75}$$

The response at T_p is

$$y(T_P) = C\left(1 + e^{-\zeta\frac{\pi}{\sqrt{1-\zeta^2}}}\right) \tag{13.76}$$

and is found as follows

$$y\left(T_p\right) = C\left[1 + \frac{e^{-\zeta\omega_n\frac{\pi}{\omega_n\sqrt{1-\zeta^2}}}}{\sqrt{1-\zeta^2}}\cos\left(\omega_d\frac{\pi}{\omega_n\sqrt{1-\zeta^2}} + \phi\right)\right]$$

$$= C\left[1 + \frac{e^{\frac{-\zeta\omega_n \pi}{\omega_n\sqrt{1-\zeta^2}}}}{\sqrt{1-\zeta^2}}\cos(\pi+\phi)\right]$$

$$= C\left[1 + \frac{e^{\frac{-\zeta\omega_n \pi}{\omega_n\sqrt{1-\zeta^2}}}}{\sqrt{1-\zeta^2}}\sqrt{1-\zeta^2}\right]$$

$$= C\left(1 + e^{\frac{-\zeta\pi}{\sqrt{1-\zeta^2}}}\right)$$

With the performance estimates calculated from the data and Eqs. (13.75) and (13.76), it is possible to estimate ζ and ω_n. First, find ζ by using the Eq. (13.76). Then, using the solution for ζ, substitute this value into Eq. (13.75) to find ω_n. The phase angle ϕ in Eq. (13.73) is determined from the estimate for ζ—that is,

$$\phi = \pi + \tan^{-1}\frac{-\zeta}{\sqrt{1-\zeta^2}}$$

EXAMPLE PROBLEM 13.11

Find ζ and ω_n for the data in Figure 13.67.

Solution

From the data in Figure 13.67, $C = 1.0$, $T_p = 0.011$, and $y(T_p) = 1.37$. Therefore,

$$y(T_P) = C\left(1 + e^{\frac{-\zeta\pi}{\sqrt{1-\zeta^2}}}\right) \quad \rightarrow \quad \zeta = \sqrt{\frac{\frac{(\ln(y(T_p)-1))^2}{\pi^2}}{1+\frac{(\ln(y(T_p)-1))^2}{\pi^2}}} = 0.3$$

$$T_p = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}} \quad \rightarrow \quad \omega_n = \frac{\pi}{T_p\sqrt{1-\zeta^2}} = 300 \text{ radians/s}$$

and

$$\phi = \pi + \tan^{-1}\frac{-\zeta}{\sqrt{1-\zeta^2}} = 2.8369 \text{ radians.}$$

Impulse Response

The complete response of the system, $a_2\ddot{y} + a_1\dot{y} + a_0 y = \delta(t)$, to an impulse with zero initial conditions is given in Table 13.6. Let us consider estimating ζ and ω_n for the overdamped case. Here, the solution is

$$y(t) = \frac{1}{2\omega_n\sqrt{\zeta^2 - 1}}\left(e^{s_1 t} - e^{s_2 t}\right) \tag{13.77}$$

where $s_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$. As before, we use the time to peak overshoot, T_p, to find the parameter values. The time to peak overshoot is found by first calculating

$$\frac{\partial y}{\partial t} = \frac{1}{2\omega_n\sqrt{\zeta^2 - 1}}\left(s_1 e^{s_1 t} - s_2 e^{s_2 t}\right) \tag{13.78}$$

and then determining T_p from $\frac{\partial y}{\partial t}\big|_{t=T_p} = 0$. This gives

$$T_p = \frac{\ln\left(\frac{s_2}{s_1}\right)}{s_1 - s_2} \tag{13.79}$$

In this situation, the time to peak overshoot is easy to estimate from the data.

If one of the time constants is considerably larger than the other, then the larger or dominant time constant can be determined from the data by locating when the response is less than 1 percent of the initial value. For example, consider the impulse response $y(t) = e^{-t} - e^{-4t}$ as shown in Figure 13.68. The response of the smaller time constant ($-e^{-4t}$) goes to zero by approximately 1 s, well before the response of the dominant time constant ($-e^{-t}$). So the impulse response is essentially equal to the response of the dominant time constant ($-e^{-t}$) for $t > 1$ s. To determine the dominant time constant, locate the time at which the response for the dominant time is at 1 percent of the value at $t = 0$ s; that time is 5τ. To estimate the initial value of the response for the dominant time constant at $t = 0$, project the exponential back, as shown in Figure 13.68, to give a value of 1, then take 1 percent of that value to find 5τ. The dominant time constant is then easily determined. Thus, given T_p and one of the time constants, the smaller time constant can then be determined from Eq. 13.79 and using MATLAB's "solve" command, from which ζ and ω_n are calculated (i.e., $\omega_n^2 = s_1 s_2$ and $\zeta = \dfrac{s_1 + s_2}{2\omega_n}$).

TABLE 13.6 Impulse Response for a Second-Order System

Damping	Natural Response Equation
Overdamped	$y(t) = \dfrac{1}{2\omega_n\sqrt{\zeta^2 - 1}}\left(e^{s_1 t} - e^{s_2 t}\right)$
Underdamped	$y(t) = \dfrac{\omega_n e^{-\omega_n \zeta t}}{\sqrt{1 - \zeta^2}}\sin(\omega_d t)$
Critically damped	$y(t) = te^{-\omega_n t}$

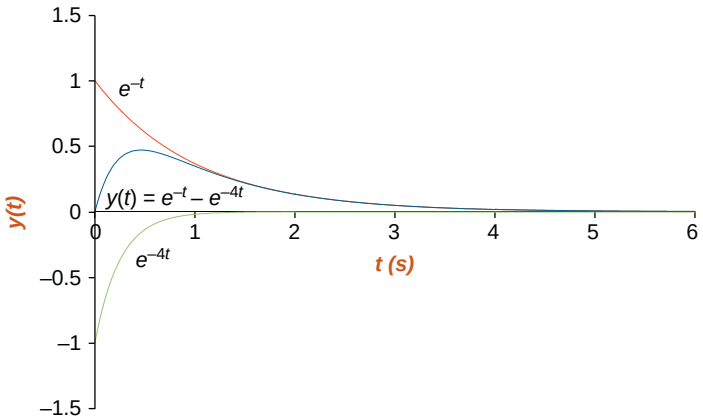


FIGURE 13.68 Impulse response $y(t) = e^{-t} - e^{-4t}$, with functions e^{-t} and e^{-4t} also shown.

EXAMPLE PROBLEM 13.12

Find ζ and ω_n for the impulse response shown in Figure 13.69.

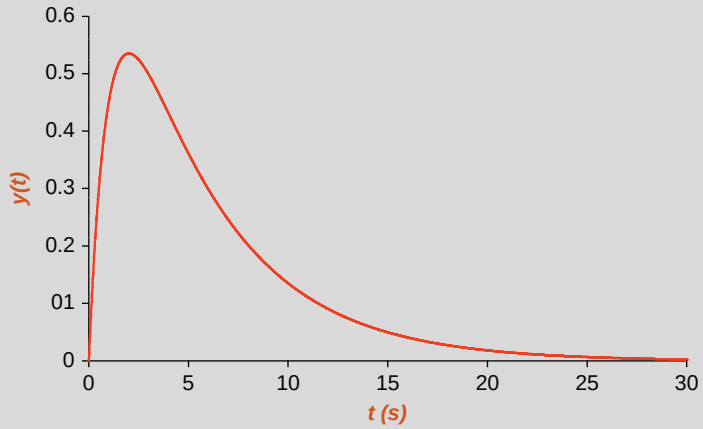


FIGURE 13.69 Illustration for Example Problem 13.12.

Solution

The data (solid line) is redrawn in Figure 13.70 with the exponential of the dominant time constant extrapolated back to zero (dotted line). This gives an initial value of 1 for the exponential. The response falls to 1 percent of 1 at approximately 25 s, giving $5\tau_2 \approx 25$, or $\tau_2 = 5$ s for the dominant time constant and $s_2 = -0.2$.

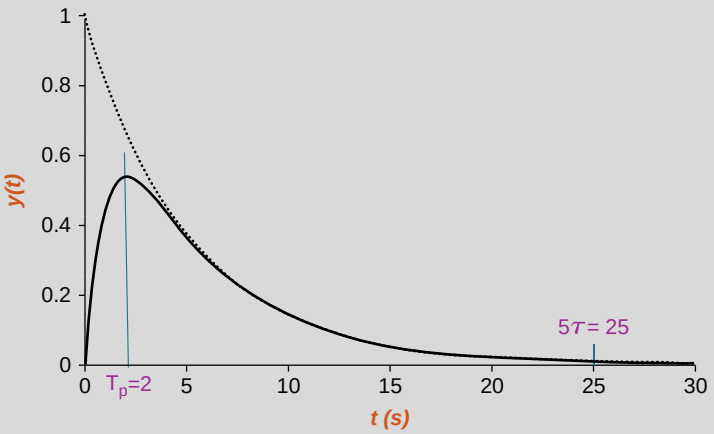


FIGURE 13.70 Graphical analysis for Example Problem 13.12.

From Figure 13.70, the time to peak overshoot occurs at approximately 2 s. Using Eq. (13.79), we have

$$s_1 T_p + \ln(s_1) = s_2 T_p + \ln(s_2)$$

and MATLAB's "solve" command gives s_1.

```
>> solve('s1*2+log(s1)=s2*2+log(s2)','s2=-.2')
ans =
s1: [2x1 sym]
s2: [2x1 sym]
>> s1
s1 =
-1.00939406263234159227988829136295
-0.2
```

Notice that there are two roots for s_1, since both are solutions to the equation. Since $s_2 = -.2$, then $s_1 = -1$. Finally, we have

$$\omega_n^2 = s_1 s_2 = 0.2 \text{ and } \zeta = \frac{s_1 + s_2}{2\omega_n} = 3.$$

13.11 EXERCISES

1. Consider the system shown in Figure 13.71 defined with $M_1 = 1$ kg, $M_2 = 2$ kg, $B_1 = .5$ N-s/m, $B_2 = 1$ N-s/m, $K_1 = 3$ N/m, and $K_2 = 2$ N/m . Let $f(t)$ be the applied force, and x_1 and x_2 be the displacements from rest. (a) Find the transfer function. (b) Solve for $x_1(t)$ if $f(t) = 3u(t)$ N and zero initial conditions. (c) Simulate the solution with SIMULINK if $f(t) = 3u(t)$ N and zero initial conditions. (c) Use MATLAB to draw the Bode diagram.

Continued

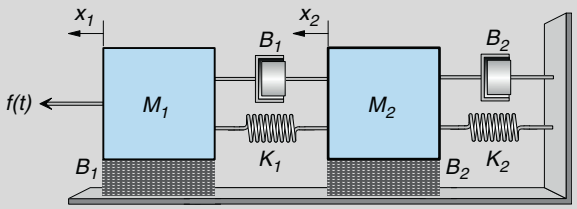


FIGURE 13.71 Illustration for Exercise 1.

2. Consider the system shown in Figure 13.72 defined with $M_1 = 2$ kg, $B_1 = 1$ N-s/m, $B_2 = 2$ N-s/m, $K_1 = 5$ N/m, and $K_2 = 2$ N/m. Let $f(t)$ be the applied force, and x_1 and x_2 be the displacements from rest. (a) Find the transfer function. (b) Solve for $x_2(t)$ if $f(t) = u(t)$ N and zero initial conditions. (c) Simulate the solution with SIMULINK if $f(t) = u(t)$ N and zero initial conditions. (c) Use MATLAB to draw the Bode diagram.

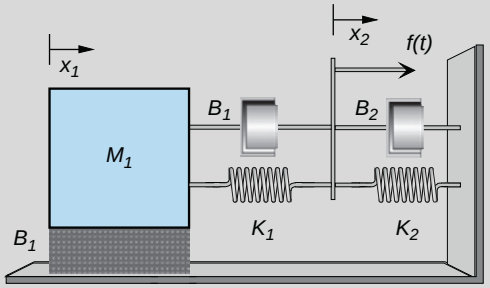


FIGURE 13.72 Illustration for Exercise 2.

3. Consider the system illustrated in Figure 13.73 defined with $M_1 = 3$ kg, $M_2 = 1$ kg, $B_1 = 1$ N-s/m, $B_2 = 2$ N-s/m, $K_1 = 1$ N/m, and $K_2 = 1$ N/m. Let $f(t)$ be the applied force, and x_1 and x_2 be

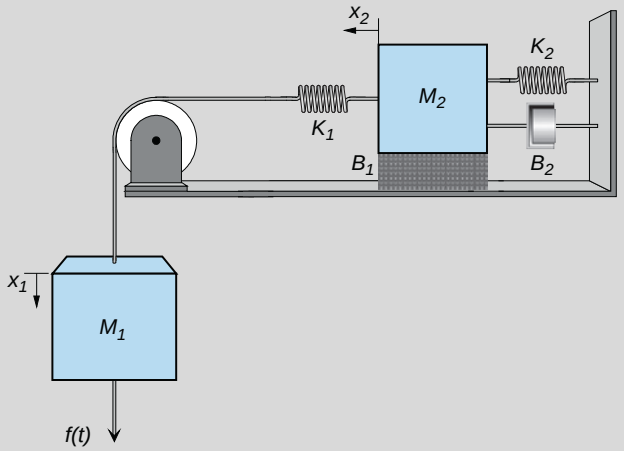


FIGURE 13.73 Illustration for Exercise 3.

the displacements from rest. The pulley is assumed to have no inertia or friction. (a) Write the differential equations that describe this system. (b) Solve for $x_2(t)$ if $f(t) = 10u(t)$ N and zero initial conditions. (c) If the input is a step with magnitude 10 N and zero initial conditions, simulate the solution with SIMULINK.

4. Consider the system in Figure 13.74, where there are two viscous elements B_2, K_1 is a translational element, and K_2 is a rotational element. Let $\tau(t)$ be the applied torque, x_1 the displacement of M from rest, and θ the angular displacement of the element J from rest (i.e., when the springs are neither stretched nor compressed). The pulley has no inertia or friction, and the cable does not stretch. (a) Write the differential equations that describe the system. (b) Write the state variable equations that describe the system. (c) If the input is a step with magnitude 3 N and zero initial conditions, simulate the solution with SIMULINK.

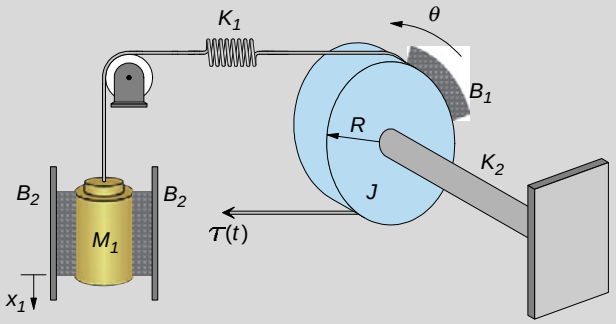


FIGURE 13.74 Illustration for Exercise 4.

5. Consider the system in Figure 13.75, where there are two viscous elements B_2, K_1 is a translational element, and K_2 is a rotational element. Let $f(t)$ be the applied force, x_1 the displacement of M_1 and x_2 the displacement of M_2 from rest, and θ the angular displacement of the element J from rest (i.e., when the springs are neither stretched nor compressed). The pulley has no inertia or friction, and the cable does not stretch. (a) Write the differential equations that describe the system. (b) Write the state variable equations that describe the system.

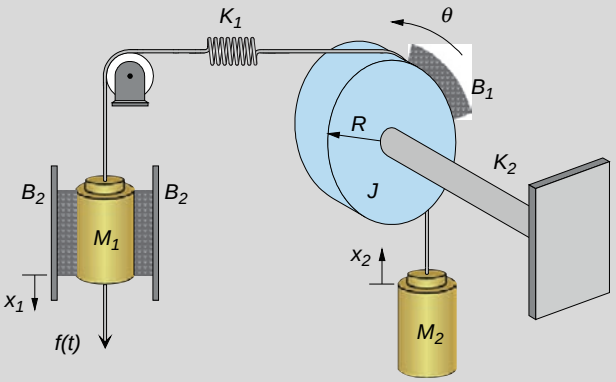


FIGURE 13.75 Illustration for Exercise 5.

Continued

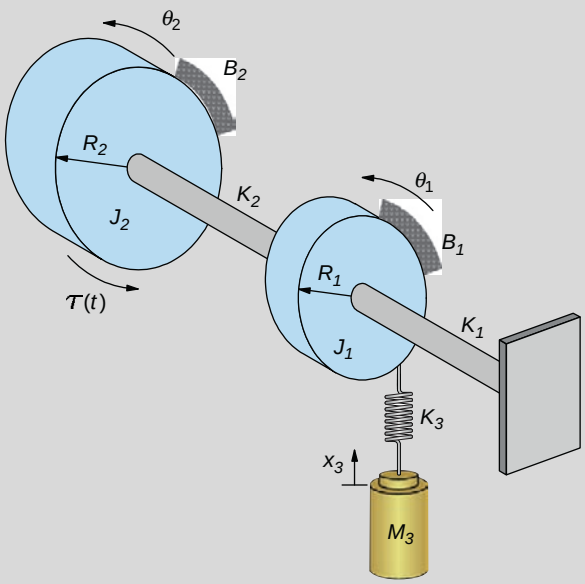


FIGURE 13.76 Illustration for Exercise 6.

6. Consider the system in the following diagram (Figure 13.76) defined with $M_3 = 1$ kg, $K_3 = 4$ N/m (translational), $J_1 = 0.5$ kg-m^2, $B_1 = 1$ N-s-m, $B_2 = 2$ N-s-m, $K_1 = 2$ N-m, $J_2 = 2$ kg-m^2, $K_2 = 1$ N-m, $R_1 = 0.2$ m, and $R_2 = 1.0$ m. Let $\tau(t)$ be the applied torque, x_3 be the displacement of M_3 from rest, and θ_1 and θ_2 be the angular displacement of the elements J_1 and J_2 from rest (i.e., when the springs are neither stretched nor compressed). (a) Write the differential equations that describe this system. (b) If the input is a step with magnitude 10 N, simulate the solution with SIMULINK.
7. What is the main sequence diagram? How do results from the Westheimer model in Section 13.3 compare with the main sequence diagram?
8. Simulate a 20° saccade with the Westheimer model in Section 13.3 with $\zeta = 0.7$ and $\omega_n = 120$ radians/s using SIMULINK. Assume that $K = 1$ N-m. Repeat the simulation for a 5, 10, and 15° saccade. Compare these results with the main sequence diagram in Figure 13.6.
9. Suppose the input to the Westheimer model is a pulse-step waveform as described in Section 13.3, and $\zeta = 0.7$, $\omega_n = 120$ radians/s, and $K = 1$ N-m. (a) Estimate the size of the step necessary to keep the eyeball at 20°. (b) Using SIMULINK, find the pulse magnitude that matches the main sequence diagram in Figure 13.6 necessary to drive the eyeball to 20°. (c) Repeat part (b) for saccades of 5, 10, and 15°. (d) Compare these results with those of the Westheimer model.
10. A model of the saccadic eye movement system is characterized by the following equation.

$$\tau = 1.74 \times 10^{-3}\,\ddot{\theta} + 0.295\dot{\theta} + 25\theta$$

where τ is the applied torque. Suppose $\tau = 200u(t)$ and the initial conditions are zero. Use Laplace transforms to solve for θ(t). Sketch θ(t).

11. Given the Westheimer model described in Section 13.3 with $\zeta = \frac{1}{\sqrt{2}}$, ω_n =100 radians/s, and $K = 1$, solve for the general response with a pulse-step input as described in Figure 13.11. Examine the change in the response as the pulse magnitude is increased and the duration of the pulse, t_1, is decreased while the steady-state size of the saccade remains constant.

12. With the Westheimer model described in Section 13.3, separately estimate ζ and ω_n for 5°, 10°, 15°, and 20° saccades using information in the main sequence diagram in Figure 13.11. Assume that peak overshoot, $\theta(T_p)$, is 1 degree greater than the saccade size. Simulate the four saccades. Develop a relationship between ζ and ω_n as a function of saccade size that matches the main sequence diagram. With these relationships, plot T_p and peak velocity as a function of saccade size. Compare these results to the original Westheimer main sequence results and those in Figure 13.11.

13. Consider an unexcited muscle model as shown in Figure 13.27 with $K_{lt} = 32\ \text{Nm}^{-1}$, $K_{se} = 125\ \text{Nm}^{-1}$, and $B = 3.4\ \text{Nsm}^{-1}$ ($F = 0$ for the case of an unexcited muscle). (a) Find the transfer function $H(j\omega) = \frac{X_1}{T}$. (b) Use MATLAB to draw the Bode diagram.

14. Consider an unexcited muscle model in Figure 13.38 with $K_{lt} = 60.7\ \text{Nm}^{-1}$, $K_{se} = 125\ \text{Nm}^{-1}$, $B_1 = 2\ \text{Nsm}^{-1}$, and $B_2 = 0.5\ \text{Nsm}^{-1}$ ($F = 0$ for the case of an unexcited muscle). (a) Find the transfer function $H(j\omega) = \frac{X_1}{T}$. (b) Use MATLAB to draw the Bode diagram.

15. Consider an unexcited muscle model as shown in Figure 13.27 with $K_{lt} = 32\ \text{Nm}^{-1}$, $K_{se} = 125\ \text{Nm}^{-1}$, and $B = 3.4\ \text{Nsm}^{-1}$ ($F = 0$ for the case of an unexcited muscle). If the muscle is linearly stretched from 3 mm to 6 mm over a time interval of 0.003 s, that is, $x_1(t) = tu(t) - (t - 0.003)u(t - 0.003)+0.003$, then find the tension T.

16. Consider an unexcited muscle model in Figure 13.38 with $K_{lt} = 60.7\ \text{Nm}^{-1}$, $K_{se} = 125\ \text{Nm}^{-1}$, $B_1 = 2\ \text{Nsm}^{-1}$, and $B_2 = 0.5\ \text{Nsm}^{-1}$ ($F = 0$ for the case of an unexcited muscle). If the muscle is linearly stretched from 3 mm to 6 mm over a time interval of 0.003 s—that is, $x_1(t) = tu(t) - (t - 0.003)u(t - 0.003)+0.003$, then find the tension T.

17. From the horizontal eye movement model in Figure 13.30, derive Eq. (13.35).

18. From the horizontal eye movement model in Figure 13.41, derive Eq. (13.48).

19. From the horizontal eye movement model in Figure 13.46, derive Eq. (13.51).

20. Using the linear homeomorphic saccadic eye movement model from Section 13.6, simulate the following saccades using SIMULINK: (a) 5°, (b) 10°, (c) 15°, (d) 20°.

21. Using the linear homeomorphic saccadic eye movement model from Section 13.7, simulate the following saccades using SIMULINK: (a) 5°, (b) 10°, (c) 15°, (d) 20°.

22. Using the linear homeomorphic saccadic eye movement model from Section 13.8, simulate the following saccades using SIMULINK: (a) 5°, (b) 10°, (c) 15°, (d) 20°.

23. Consider the linear homeomorphic saccadic eye movement model given in Eq. (13.35). (a) Find the transfer function. (b) Use MATLAB to draw the Bode diagram.

24. Consider the linear homeomorphic saccadic eye movement model given in Eq. (13.48). (a) Find the transfer function. (b) Use MATLAB to draw the Bode diagram.

25. Consider the linear homeomorphic saccadic eye movement model given in Eq. (13.51). (a) Find the transfer function. (b) Use MATLAB to draw the Bode diagram.

26. Verify the length-tension curves in Figure 13.39.

Continued

27. Verify the force-velocity curve in Figure 13.40 for the muscle model in Figure 13.38. (Hint: Use SIMULINK to calculate peak velocity for each value of M.)

28. Find the force-velocity curve for the experiment in which a lever is not used for the muscle model in Figure 13.77. Compare the results with those in Figure 13.40.

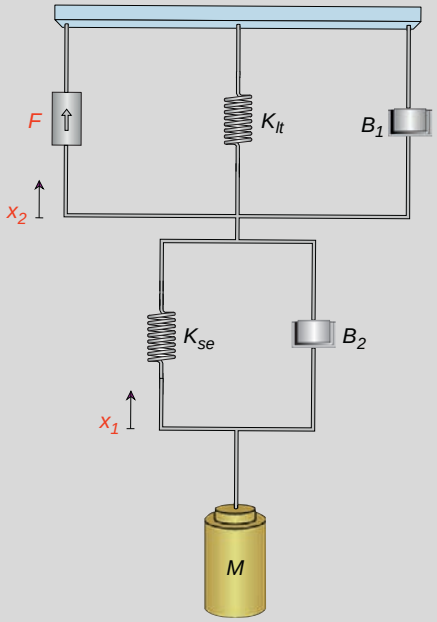


FIGURE 13.77 Illustration for Exercise 28.

29. Consider the model in Figure 13.78A of the passive orbital tissues driven by torque, $\tau(t)$, with $K_p = 0.5$ g/°, $B_p = 0.06$ gs/°, and $J_p = 4.3 \times 10^{-5}$ g ^{2}/°. All elements are rotational. (a) Find the transfer function $\dfrac{\theta(s)}{\tau(s)}$. (b) Use MATLAB to draw the Bode diagram.

30. Consider the model in Figure 13.78B of the passive orbital tissues driven by torque τ with $J_p = 4.308 \times 10^{-5}$ gs^2/°, $K_{p1} = 0.5267$ g/°, $K_{p2} = 0.8133$ g/°, $B_{p1} = 0.010534$ gs/°, and $B_{p2} = 0.8133$ gs/°. All elements are rotational. (a) Find the transfer function $H(s) = \dfrac{\theta(s)}{\tau(s)}$. (b) Use MATLAB to draw the Bode diagram.

31. Consider the model in Figure 13.79 of the eye movement system. The elements are all rotational and $f_K(\theta) = K_1\theta^2$ (a nonlinear rotational spring). (a) Write the nonlinear differential equation that describes this system. (b) Write a linearized differential equation using a Taylor Series first-order approximation about an operating point.

32. Suppose the passive elasticity of unexcited muscle is given by the following nonlinear translational force-displacement relationship $f_K(\theta) = x|x|$, where x is the displacement from equilibrium position. Determine a linear approximation for this nonlinear element in the vicinity of the equilibrium point.

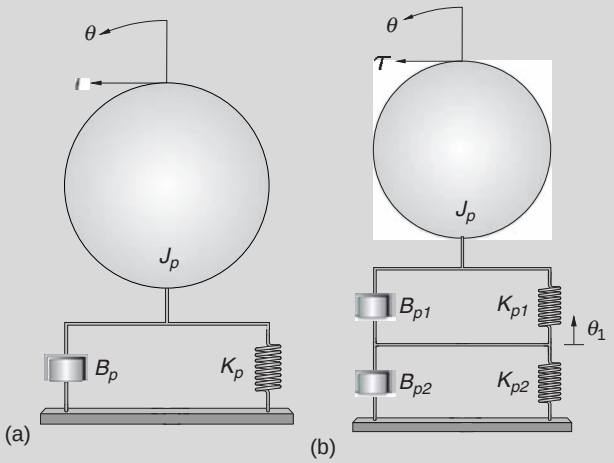


FIGURE 13.78 Illustrations for Exercises 29 and 30.

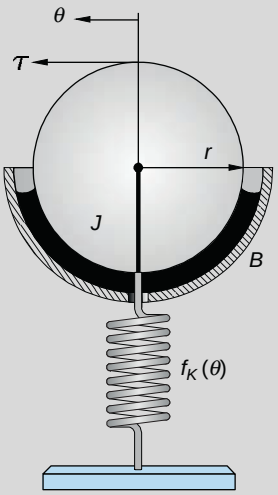


FIGURE 13.79 Illustration for Exercise 31.

33. Sinusoids of varying frequencies were applied to an open-loop system and the following results were measured. Construct a Bode diagram and estimate the transfer function.

Frequency (radians/s)	0.6	1.6	2.6	3.6	5.5	6.1	7.3	9.8	12.7	32.9	62.1	100
Magnitude Ratio, |G|	2.01	2.03	2.09	2.17	2.37	2.43	2.49	2.16	1.39	0.18	0.05	0.018
Phase (degrees)	−3.3	−8.9	−15.0	−21.8	−38.1	−44.8	−60.0	−93.6	−123.2	−164.2	−170.0	−175.0

34. Sinusoids of varying frequencies were applied to an open-loop system and the following results were measured. Construct a Bode diagram and estimate the transfer function.

Continued

Frequency (radians/s)	.001	.356	1.17	2.59	8.53	12.7	18.9	41.8	62.1	137	304	1000
20 log|G| (dB)	6.02	6.02	5.96	5.74	3.65	1.85	−0.571	−6.64	−9.95	−16.8	−23.6	−34.0
Phase (degrees)	−.086	−3.06	−10	−22	−64.9	−88.14	−116.0	−196.0	−259.0	−479.0	−959.0	−2950.0

35. Sinusoids of varying frequencies were applied to an open-loop system and the following results were measured. Construct a Bode diagram and estimate the transfer function.

Frequency (radians/s)	0.11	0.24	0.53	1.17	2.6	5.7	12.7	28.1	62	137	304	453	672	1000
Magnitude Ratio, |G|	2.0	2.0	2.0	2.0	1.93	1.74	1.24	0.67	0.32	0.15	.07	0.044	0.03	0.02
Phase (degrees)	−0.62	−1.37	−3.03	−6.7	−14.5	−29.8	−51.8	−70.4	−80.9	−85.8	−88.1	−88.7	−89.7	−89.4

36. Sinusoids of varying frequencies were applied to an open-loop system and the following results were measured. (Data from [43].) Construct a Bode diagram for the data.

Frequency (radians/s)	1	3	7	10	15	20	25	30	35	40	50	60	70	80	90	100	110	120	130	140
Magnitude Ratio, |G|	1	.95	.77	.7	.67	.63	.6	.53	.48	.44	.35	.31	.33	.35	.32	.32	.3	.29	.27	.26
Phase (radians)	−.035	−.227	−.419	−.541	−.611	−.768	−.995	−1.08	−1.24	−1.31	−1.52	−1.92	−1.61	−1.83	−2.08	−2.23	−2.53	−2.72	−2.9	−3.0

Estimate the transfer function if it consists of (a) two poles; (b) a pole and a complex pole pair; (c) two poles, a zero, and a complex pole pair; (d) three poles, a zero, and a complex pole pair. (Hint: It may be useful to solve this program using the MATLAB System Identification toolbox or Seidel's program.)

37. The following data were collected for the step response for an unknown first-order system. Find the parameters that describe the model.

T	0.0	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05	0.055	0.06	0.1
$v(t)$	0.00	3.41	5.65	7.13	8.11	8.75	9.18	9.46	9.64	9.76	9.84	9.90	9.93	10.0

38. Suppose a second-order underdamped system response to a step is given by Eq. (13.73) and has $C = 10$, $T_p = 0.050$, and $y(T_p) = 10.1$. Find ζ and ω_n. A stylized 10° saccade is shown in the following figure. Estimate ζ and ω_n for the Westheimer model. Calculate the time to peak velocity and peak velocity.

39. Consider the data in Figure 13.80. Estimate ζ, ω_n, and ϕ if the system has the solution of the form of Eq. (13.73).

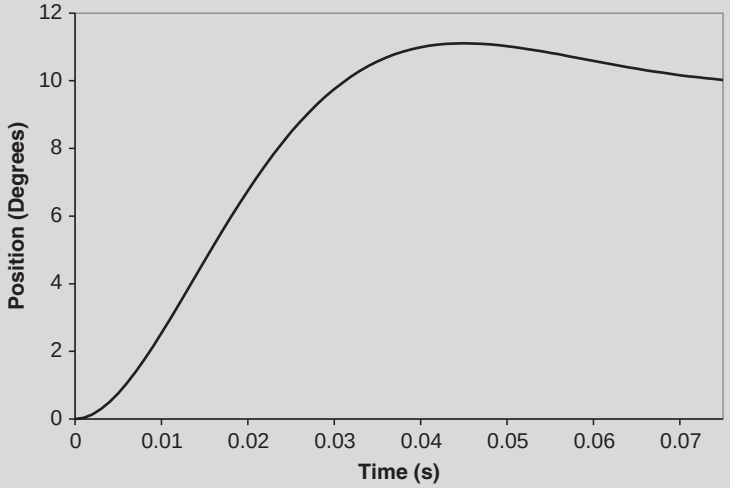


FIGURE 13.80 Illustration for Exercise 39.

References

[1] A.T. Bahill, Bioengineering: Biomedical, Medical and Clinical Engineering, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[2] A.T. Bahill, L. Stark, The high-frequency burst of motoneuronal activity lasts about half the duration of saccadic eye movements, J. Neurophys. 33 (1975b) 382–392.

[3] A.T. Bahill, M.R. Clark, L. Stark, The main sequence, a tool for studying human eye movements, Math. Biosci. 24 (1975a) 194–204.

[4] A.T. Bahill, J.R. Latimer, B.T. Troost, Linear homeomorphic model for human movement, IEEE Trans. Biomed Eng. BME-27 (1980) 631–639.

[5] R.H.S. Carpenter, Movements of the Eyes, second ed., Pion Ltd., London, 1988.

[6] M.R. Close, A.R. Luff, Dynamic properties of inferior rectus muscle of the rat, J. Physiol., London 236 (1974) 258.

[7] C.C. Collins, The human oculomotor control system, in: G. Lennerstrand, P. Bach-y-Rita (Eds.), Basic Mechanisms of Ocular Motility and Their Clinical Implications, Pergamon Press, Inc., Oxford, 1975, pp. 145–180.

[8] C.C. Collins, D. O'Meara, A.B. Scott, Muscle tension during unrestrained human eye movements, J. Physiol. 245 (1975) 351–369.

[9] J.D. Enderle, Observations on Pilot Neurosensory Control Performance During Saccadic Eye Movements, Aviat. Space Environ. Med. 59 (1988) 309–313.

[10] J.D. Enderle, A Physiological Neural Network for Saccadic Eye Movement Control. Air Force Material Command, in: Armstrong Laboratory AL/AO-TR-1994-0023, 1994, p. 48.

[11] J.D. Enderle, Neural Control of Saccades, in: J. Hyönä, D. Munoz, W. Heide, R. Radach (Eds.), The Brain's Eyes: Neurobiological and Clinical Aspects to Oculomotor Research, Progress in Brain Research, vol. 140, Elsevier, Amsterdam, 2002, pp. 21–50.

[12] J.D. Enderle, Eye Movements, in: M. Akay (Ed.), Wiley Encyclopedia of Biomedical Engineering, John Wiley & Sons, Inc., Hoboken, 2006a.

[13] J.D. Enderle, The Fast Eye Movement Control System, in: J. Bronzino (Ed.), The Biomedical Engineering Handbook Biomedical Engineering Fundamentals, third ed., CRC Press, Inc., Boca Raton, FL, 2006b, Chapter 16, pages 16-1 to 16-21.

[14] J.D. Enderle, Models of Horizontal Eye Movements. Part 1: Early Models of Saccades and Smooth Pursuit, Morgan & Claypool Publishers, San Rafael, CA, 2010, p. 149.

[15] J.D. Enderle, E.J. Engelken, Simulation of Oculomotor Post-Inhibitory Rebound Burst Firing Using a Hodgkin-Huxley Model of a Neuron, Biomed. Sci. Instrum. 31 (1995) 53–58.

[16] J.D. Enderle, J.W. Wolfe, Time-Optimal Control of Saccadic Eye Movements, IEEE Trans. Biomed. Eng. BME-34 (1) (1987) 43–55.

[17] J.D. Enderle, J.W. Wolfe, Frequency Response Analysis of Human Saccadic Eye Movements: Estimation of Stochastic Muscle Forces, Comp. Bio. Med. 18 (1988) 195–219.

[18] J.D. Enderle, W. Zhou, Models of Horizontal Eye Movements. Part 2: A 3rd-Order Linear Saccade Model, Morgan & Claypool Publishers, San Rafael, CA, 2010, p. 144.

[19] J.D. Enderle, E.J. Engelken, R.N. Stiles, A comparison of static and dynamic characteristics between rectus eye muscle and linear muscle model predictions, IEEE Trans. Biomed. Eng. 38 (1991) 1235–1245.

[20] J.D. Enderle, J.W. Wolfe, J.T. Yates, The Linear Homeomorphic Saccadic Eye Movement Model — A Modification, IEEE Trans. Biomed. Eng. BME-31 (11) (1984) 717–720.

[21] E.J. Engelken, K.W. Stevens, J.W. Wolfe, J.T. Yates, A Limbus Sensing Eye-Movement Recorder, USAF School of Aerospace Medicine, Brooks AFB, TX, 1984, USAFSAM-TR-84-29.

[22] W.O. Fenn, B.S. Marsh, Muscular Force at different speeds of shortening, J. Physiol. (London) 35 (1935) 277–297.

[23] A.F. Fuchs, C. Kaneko, C. Scudder, Brainstem Control of Saccadic Eye Movements, Annu. Rev. Neurosci. 8 (1985) 307–337.

[24] A.F. Fuchs, E.S. Luschei, Firing patterns of abducens neurons of alert monkeys in relationship to horizontal eye movement, J. Neurophysiol. 33 (3) (1970) 382–392.

[25] N.J. Gandhi, E.L. Keller, Spatial distribution and discharge characteristics of the superior colliculus neurons antidromically activated from the omnipause region in monkey, J. Neurophysiol. 76 (1997) 2221–2225.

[26] H. Goldstein, The neural encoding of saccades in the rhesus monkey (Ph.D. dissertation), The Johns Hopkins University, Baltimore, MD, 1983.
[27] M.R. Harwood, L.E. Mezey, C.M. Harris, The Spectral Main Sequence of Human Saccades, J. Neurosci. 19 (20) (1999) 9098–9106.
[28] A.V. Hill, The heat of shortening and dynamic constants of muscle, Pro. Royal Soc., London (B) 126 (1938) 136–195.
[29] A.V. Hill, The development of the active-state of muscle during the latent period, Pro. Royal Soc., London (B) 137 (1950a) 320–329.
[30] A.V. Hill, The series elastic component of muscle, Pro. Royal Soc., London (B) 137 (1950) 273–280.
[31] A.V. Hill, The transition from rest to full activity in muscles: the velocity of shortening, Pro. Royal Soc., London (B) 138 (1951a) 329–338.
[32] A.V. Hill, The effect of series compliance on the tension developed in a muscle twitch, Pro. Royal Soc., London (B) 138 (1951b) 325–329.
[33] F.K. Hsu, A.T. Bahill, L. Stark, Parametric sensitivity of a homeomorphic model for saccadic and vergence eye movements, Comp. Prog. Biomed. 6 (1976) 108–116.
[34] X. Hu, H. Jiang, C. Gu, C. Li, D. Sparks, Reliability of Oculomotor Command Signals Carried by Individual Neurons, PNAS (2007) 8137–8142.
[35] M.C. Khoo, Physiological Control Systems: Analysis, Simulation, and Estimation, IEEE Press, Piscataway, NJ, 2000.
[36] R.J. Leigh, D.S. Zee, The Neurology of Eye Movements, third ed., Oxford University Press, New York, 1999.
[37] A. Levin, J. Wyman, The viscous elastic properties of a muscle, Pro. Royal Soc., London (B) 101 (1927) 218–243.
[38] L. Ling, A. Fuchs, C. Siebold, P. Dean, Effects of initial eye position on saccade-related behavior of abducens nucleus neurons in the primate, J. Neurophysiol. 98 (6) (2007) 3581–3599.
[39] S. Ramat, R.J. Leigh, D. Zee, L. Optican, What Clinical Disorders Tell Us about the Neural Control of Saccadic Eye Movements, Brain (2007) 1–26.
[40] D.A. Robinson, Oculomotor control signals, in: G. Lennerstrand, P. Bach-y-Rita (Eds.), Basic Mechanisms of Ocular Motility and their Clinical Implication, Pergamon Press, Oxford, 1975, pp. 337–374.
[41] D.A. Robinson, The Mechanics of Human Saccadic Eye Movement, J. Physiol., London 174 (1964) 245.
[42] D.A. Robinson, Models of mechanics of eye movements, in: B.L. Zuber (Ed.), Models of Oculomotor Behavior and Control, CRC Press, Boca Raton, FL, 1981, pp. 21–41.
[43] R.C. Seidel, Transfer-function-parameter estimation from frequency-response data – A FORTRAN program, (1975). NASA TM X-3286 Report.
[44] D.L. Sparks, R. Holland, B.L. Guthrie, Size and distribution of movement fields in the monkey superior colliculus, Brain Res. 113 (1976) 21–34.
[45] P.A. Sylvestre, K.E. Cullen, Quantitative analysis of abducens neuron discharge dynamics during saccadic and slow eye movements, J. Neurophysiol. 82 (5) (1999) 2612–2632.
[46] J.A. Van Gisbergen, D.A. Robinson, S. Gielen, A quantitative analysis of generation of saccadic eye movements by burst neurons, J. Neurophysiol. 45 (3) (1981) 417–442.
[47] R.B. Weber, R.B. Daroff, Corrective movements following refixation saccades: Type and control system analysis, Vision Res. 12 (1972) 467–475.
[48] G. Westheimer, Mechanism of saccadic eye movements, AMA Arch. Ophthalmol. 52 (1954) 710–724.
[49] D.R. Wilkie, Muscle: Studies in Biology, vol. 11, Edward Arnold Ltd., London, United Kingdom, 1968.
[50] W. Zhou, X. Chen, Enderle, An Updated Time-Optimal 3rd-Order Linear Saccadic Eye Plant Model, Int. J. Neural. Syst. 19 (5) (2009) 309–330.

CHAPTER

14

Biomedical Transport Processes

Gerald E. Miller, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Understand the function of the human lung in terms of mass balance and mass transfer.
- Understand the fundamental equations for mass transfer, including the concepts of diffusion and osmosis.
- Understand the fundamental concepts of mass transfer across pores in systemic capillaries.
- Understand the function of the human kidney and its most basic mass transfer element: the nephron.
- Understand how dialysis systems function and are controlled.
- Understand the role of blood cells in whole blood.
- Understand how the fundamental equations describing the conservation of mass and momentum equations are derived and the underlying physics.
- Understand the three basic elements of heat transfer, including conduction, convection, and radiation.
- Understand the process of heat exchange using a double pipe heat exchanger for heating and cooling of blood.
- Understand how heat is transferred from the core of the body to the periphery and then to the environment.
- Understand how heat transfer occurs in extreme heat environments.

14.1 BIOMEDICAL MASS TRANSPORT

Mass transport in the human body is a vital process that affects how the lungs function in transferring air and its components to the bloodstream, how the capillaries function in transferring nutrients and gases to surrounding body tissues, and how the kidneys function in transferring metabolic waste products and excess water from the blood into the urine. Mass transfer processes also occur in artificial devices such as artificial kidneys (dialysis) and artificial ventilators and respirators. Mass transfer in the body also affects transport across cell membranes, which controls processes in millions of cells affecting every area of the body. The majority of mass transfer occurs across small membranes of thin surfaces in order to shorten the distance over which substances must travel from point A to point B. This is true of cells in the body, which are quite thin, as well as artificial devices whose components are manufactured to be very thin.

14.1.1 Analysis of Respiration and Gas Transport

The human lungs control gas exchange from our environment into the bloodstream by means of pressure and concentration gradients. When we breath, air enters the lungs through a large entrance, the trachea, and eventually branches into smaller and smaller segments until reaching the smallest elements of the lungs, the alveoli. Each of these thin elements is in close proximity to blood in pulmonary capillaries, which are the smallest and thinnest of the blood vessels. With each of the alveoli in close proximity to a pulmonary capillary, the distance for gas exchange is very short, which thus shortens the time by which complete gas exchange occurs. A diagram of the lungs and its branching system is shown in Figure 14.1.

The two human lungs contain approximately 300 to 500 million alveoli, having a total surface area of about 75 m^2 in adults, the size of a tennis court. The branching of the airways into the alveoli represents millions of tiny sacs, which not only represents thinner membranes to speed gas exchange but also more surface area to speed gas exchange. The alveolar network is shown in Figure 14.2.

The lungs may be separated (for the purposes of mass transfer/gas exchange) as a dead space and an alveolar space. The dead space consists of the trachea, the bronchioles, and the bronchi, which are all large segments of the airway where there is air flow but no gas exchange with the bloodstream. The alveolar space is where the actual gas exchange occurs. However, both zones heat the inspired air to body temperature (37°C), as well as humidifying the air. Thus, expired air is heated to body temperature and is usually fully saturated with water vapor. In fact, it is possible to lose up to a half pound per day merely from losing water from the body via respiration. Figure 14.3 shows the relationship between the dead space and the alveolar space.

The amount of air that one inhales and exhales (without exertion) is called the *tidal volume*, which is 500 ml (about a fluid pint) per breath. The typical breathing rate is 12 breaths per minute at rest. The tidal volume represents only that portion of the lung volume where relatively easy breathing (in and out) occurs. Figure 14.4 shows all of the lung volumes and how the tidal volume compares to other (forced) breathing volumes.

Ambient air is approximately 79 percent nitrogen and 21 percent oxygen on a dry basis (not including any water vapor/humidity in the air). With a single breath, oxygen is transferred

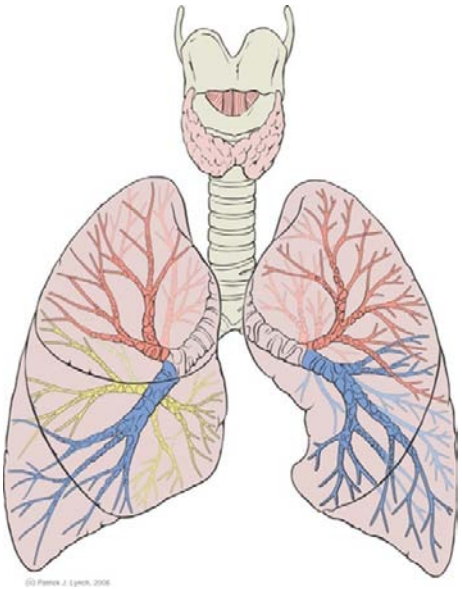


FIGURE 14.1 Branching of the airway network in the human lungs from the trachea ending in alveoli. *Patrick J. Lynch, illustrator; C. Carl Jaffe, MD, cardiologist; Yale University Center for Advanced Instructional Media Medical, Illustrations by Patrick Lynch, generated for multimedia teaching projects by the Yale University School of Medicine, Center for Advanced Instructional Media, 1987–2000.*

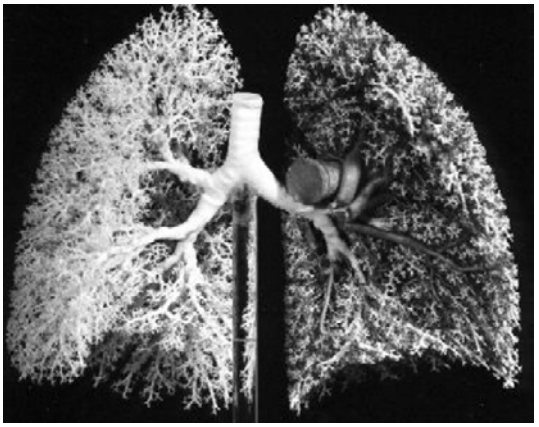


FIGURE 14.2 A cast of the human lungs showing the alveolar sacs. *From the Anatomical Institute, Bern.*

from the air to the pulmonary blood, and carbon dioxide is transferred from the blood to the air, across the alveolar membranes. Nitrogen is not transferred under normal conditions. Figure 14.5 shows the gas exchange between the alveoli and the pulmonary capillaries.

The close spatial relationship between the pulmonary capillaries is necessary to shorten the diffusion distance between the alveoli and the capillaries to ensure adequate gas transfer between the lungs and the capillaries. Figure 14.6 depicts this close spatial relationship.

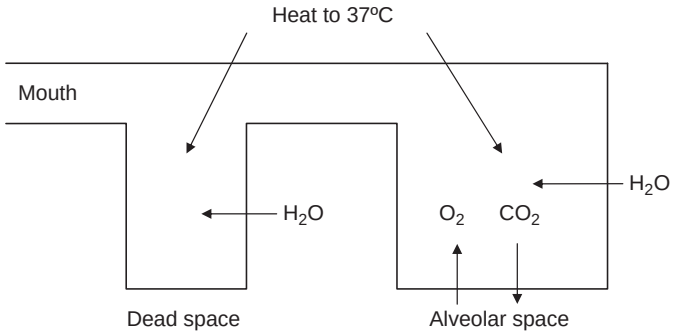


FIGURE 14.3 The dead space and alveolar space in the human lungs. Gas exchange occurs only in the alveolar space, while both zones heat the air and saturate with water vapor.

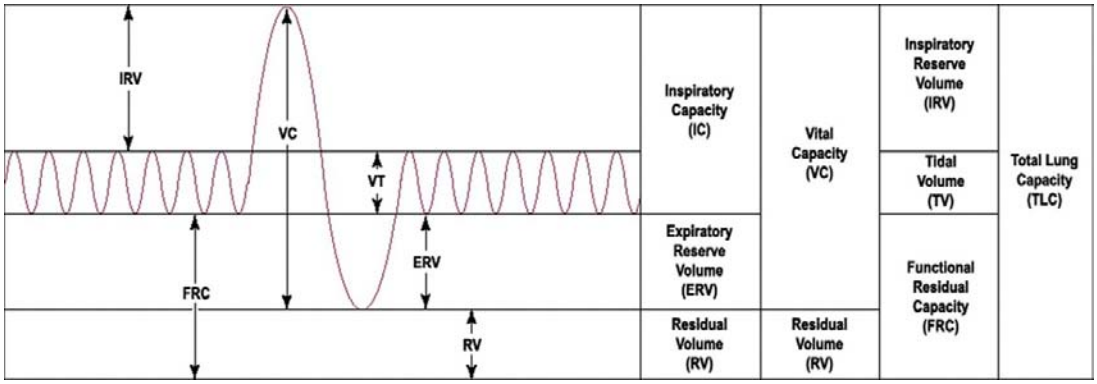


FIGURE 14.4 Lung capacity with various breathing volumes.

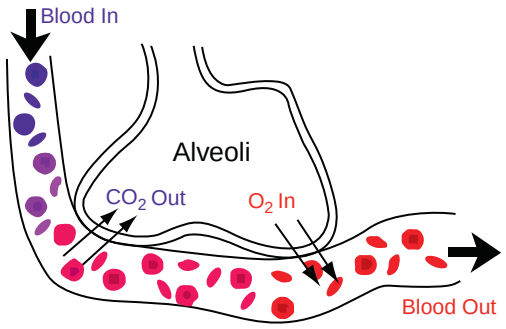


FIGURE 14.5 Gas exchange between the alveoli and the pulmonary capillaries. This event occurs across all of the millions of alveoli and associated capillaries.

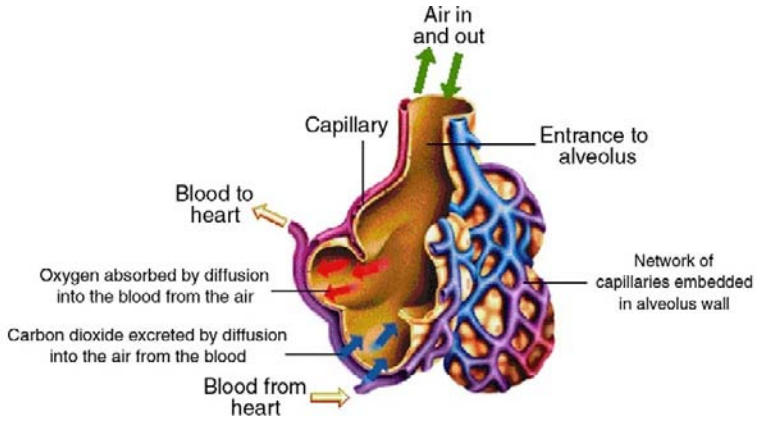


FIGURE 14.6 The close spatial relationship between alveoli and pulmonary capillaries.

The resulting single breath gas exchange between the alveoli and the pulmonary capillaries can be simplified via a mass balance. This is summarized in Table 14.1. The gases (nitrogen, oxygen, and carbon dioxide) are shown with "wet" percentages, which incorporate the water vapor in the air. Note that not all of the oxygen in the air is taken up by the lungs and that the expired gases from the lungs are heated to body temperature and saturated with water vapor—thus the higher water vapor content of the expired air. There is a trace amount of carbon dioxide in the ambient air but a higher percentage in the expired air. Thus, oxygen replaces carbon dioxide in the pulmonary bloodstream.

For the purposes of gas exchange, the components of the inspired air are described by means of their partial pressures. This is the fractional amount of total gas pressure due to the substance being measured. For example, at sea level, the total atmospheric pressure is 760 mm Hg. The amount of this pressure that is due to oxygen is approximately $0.21 \times 760 = 160$ mm Hg. We would say that the partial pressure of oxygen at sea level in dry air (no vapor in the air) is 160 mm Hg. The partial pressure of carbon dioxide in dry air at sea level is $0.03 \times 760 = 22.8$ mm Hg. The maximum amount of water vapor in the air varies with temperature and relative humidity. At body temperature (37°C) air can be saturated up to 47 mm Hg of water vapor pressure. Therefore, in the lung, where air is totally water-saturated, the partial pressure of water vapor would be 47 mm Hg.

TABLE 14.1 Single Breath Analysis for the Components of Air

Air Component	Atmospheric Air %	Expired Air %
Nitrogen and Inert Gases	78.62	74.9
Oxygen	20.85	15.3
Carbon Dioxide	0.03	3.6
Water Vapor	0.5	6.2
Total	100	100

Thus, in terms of partial pressures, and with the ambient humidity equal to a partial pressure of 10 mm Hg (as an example), Table 14.1 can be rewritten as follows: The partial pressure of water vapor is 10 mm Hg, and the partial volume would be equal to $(10/760) \times 500$ ml tidal volume = 6.5 ml. The dry gases (nitrogen and oxygen) represent the remainder of the tidal volume, or 500 – 6.5 = 493.5. Thus, the nitrogen partial volume would be $493.5 \times 79\% = 389.9$ ml, and the partial volume of oxygen would be $493.5 \times 21\% = 103.6$ ml.

The nitrogen is not exchanged in the lungs, and thus its partial volume does not change. However, the oxygen is taken up by the bloodstream, and the carbon dioxide is released by the bloodstream to be exhaled from the lungs. The partial volumes of the exhaled gases are based on the metabolic load of these gases in the body. At rest, these are 284 ml/min of oxygen (at body temperature) and 226 ml/min for carbon dioxide. These are values at body temperature and pressure (BTP). The STP values are 250 ml/min for oxygen and 200 ml/min for carbon dioxide.

The BTP values are calculated by multiplying the STP values by the ration of the temperatures in absolute scale in °K. The STP temperature is 273°K, while the BTP temperature is 273 + 37 = 310°K. Thus, the BTP value for oxygen is $250 \times (310/273) = 284$, and the BTP value for carbon dioxide is $200 \times (310/273) = 226$.

The BTP values equal 23.67 ml/breath for oxygen (284/12 breaths per minute) and 18.83 ml/breath for carbon dioxide (226/12). The water vapor during exhalation equals the partial pressure of water vapor in the body (47 mm Hg) compared to the dry gas pressure (760 – 47 = 713 mm Hg) times the dry gas expired volume. Thus, the water vapor attaches to the dry gas during expiration.

Note that in the calculations shown in Table 14.2, the expired dry gas volume is the sum of nitrogen, oxygen, and carbon dioxide. The water vapor attaches to the dry gas. Also note that the dry gas pressure is only 760 – 47 = 713 mm Hg, since the water vapor accounts for the remaining 47 mm Hg. Thus, the ratio of the partial pressures of water vapor to dry gas is 47/713.

The preceding analysis does not consider expansion and contraction due to temperature differences, which do, in fact, exist. This is beyond the scope of this discussion but is available from several other sources.

In summary, gas exchange between the lungs and the pulmonary circulation occurs at the smallest elements of each, which are also in extremely close proximity to each other. With so many alveoli, the surface area for gas exchange is very large, which promotes faster

TABLE 14.2 Single Breath Analysis Based on Partial Volumes

Air Component	Inspired Gases Partial Volume	Expired Gases Partial Volume
Water Vapor	6.5	$6.5 + (47/713) \times 488.63 = 38.7$ ml
Dry Gas	493.5	389.9 + 79.9 + 18.83 = 488.63
Nitrogen	389.9	389.9
Oxygen	103.6	103.6 – 23.67 = 79.9
Carbon Dioxide	0.0	0.0 + 18.83 = 18.83

mass transfer. The mass transfer of each of the components of air depends on the partial pressure of each, as well as the metabolic load of the body, along with the diffusion distance between alveoli and pulmonary capillaries. Oxygen is taken up by the bloodstream, and carbon dioxide is given off to the lungs from the bloodstream. The exhaled gases are heated to body temperature and are fully saturated with water vapor. Thus, when you place your hand over your mouth as you exhale, the air feels warm and moist.

14.1.2 Membranes, Pores, and Diffusion

The gas exchange between the alveoli and the pulmonary capillaries is affected by the partial pressures of each gas, the distance over which mass transfer occurs, and the surface area for mass transfer. We can think of mass transfer as occurring across a membrane in a steady-state fashion. This can be described by the one-dimensional, steady-state form of Fick's Law:

$$\text{Mass exchange rate} = D\ A\ dC/dx$$

where D is the *diffusivity*, A is the surface area for mass transfer, and dC/dx is the concentration gradient for mass transfer.

For gases, the concentration gradient is the difference in the partial pressure. For a liquid or for substances dissolved in a liquid (such as in blood or extracellular fluid), the concentration gradient can be approximated by dC/dx, the difference in concentration over a specified distance. The diffusivity (D) is a parameter that describes the relative ease by which a substance moves throughout the "medium"—the fluid through which the substance moves. An example of this is how sodium ions move through extracellular fluid. The diffusivity is thus affected by both the medium and the substance to be transferred; it is a material property of both.

For gases, mass transfer occurs across the entire membrane, since gases are lipid soluble, with membranes consisting of a layer of lipids and proteins. However, when the exchange is of a liquid or a substance dissolved in a liquid (such as an ion), then the mass transfer must occur through pores in the membrane. That is because most liquids are not lipid soluble. When substances do indeed travel across pores, then there is selective mass transfer due to the size of the pore as compared to the size of the substance that might transfer. This selective mass transfer by means of the relative sizes is called *filtration*. Fick's Law is modified to account for such selective mass transfer by adding a permeability term (P), which relates the area of the pores as compared to the total area of the membrane. At times, the area for mass transfer (A) is instead described as A_P rather than as the product of $A \times P$. An example of filtration is shown in Figure 14.7.

In general, the process by which mass transfer occurs via a concentration gradient is called *diffusion*. Diffusion is the macroscopic result of random thermal motion on a microscopic scale. For example, in Figure 14.8, oxygen and nitrogen molecules move in random directions, with kinetic energy on the order of kT, where k is the thermal conductivity and T is the absolute temperature. If there are more oxygen molecules on the left side of the plane A-A than on the right, more molecules will cross to the right than to the left; there will be a net movement even though the motion of each individual molecule is completely random. Diffusion in an open environment such as within extracellular fluid, and not constrained by a membrane for mass transfer, is shown in Figure 14.8.

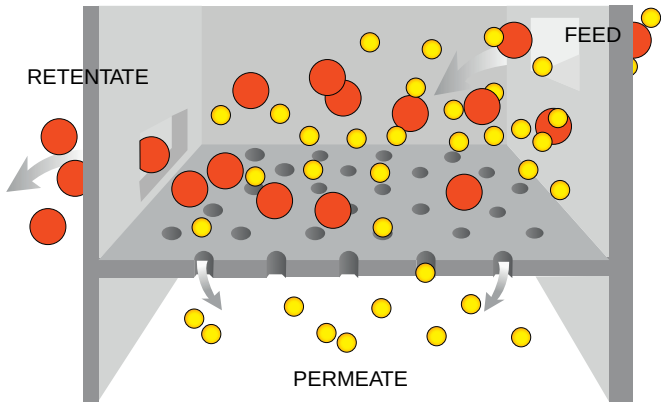


FIGURE 14.7 Filtration—selective mass transfer by means of pore size and size of substances compared to pore size.

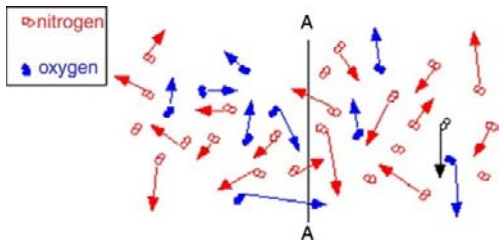


FIGURE 14.8 Random movement of particles resulting in a net movement across a dividing line based on relative concentrations on both sides of the line. *From Time Domain CVD.*

Again, the net movement of any substance (within a mixture of substances as just shown) across line A-A is a function of the concentration of that substance on each side of the line. It is certainly possible that the net movement of one substance is in the opposite direction of another, each based on its own concentration gradient. This is also true for mass transfer across pores. The concentration gradients control the amount and the direction of mass transfer for each substance. When a substance, such as an ion, is in a solution (such as within extracellular fluid), then the substance is known as a *solute* and the fluid as a *solvent*. Mass transfer of water is known as *osmosis.*

Further, the relative size of the solute as compared to the pore size may affect the potential mass transfer by either (1) completely restricting the movement (if the species is larger than the pore size) or (2) merely restricting it because it may be nearly equal to the pore size. In such cases, Fick's Law can be modified to account not only for the limited pore area (via the permeability) but also via the pore to species size, which is given by the *restrictive diffusivity* (D_R). Thus, Fick's Law could be rewritten as

$$\text{Mass exchange rate} = D_R\, A_P\, dC/dx$$

A typical cell membrane is thin (80–100 Å thick) and elastic. Membranes typically consist of 60 to 65 percent proteins, 30 to 35 percent lipids, and 5 to 10 percent polysaccharides and

TABLE 14.3 Relative Diffusion Rates as a Function of Size of Species Compared to the Size of a Pore

Substance	Diameter (Å)	Ratio to Pore D	Relative Diffusion Rate
Water	3	0.38	50,000,000
Urea	3.6	0.45	40,000,000
Chloride ion	3.86	0.48	36,000,000
Potassium ion	3.96	0.49	200
Sodium ion	5.12	0.64	100
Glycerol	6.2	0.77	1
Galactose	8.4	1.03	0
Glucose	8.6	1.04	0
Lactose	10.8	1.35	0

From Cooney (1976).

other elements. The lipids include phospholipids (65 percent), cholesterol (25 percent), and other lipids (10 percent). As was previously mentioned, substances that are not lipid soluble must travel through pores within the cell membrane wall. Such transport of these substances depends on their relative size with respect to the pore size. Table 14.3 lists various substances that can only travel through pores and their relative diffusion rates, which are related to their permeability. An 8 Å pore diameter is chosen for the purposes of comparison of diffusion rates.

Note that potassium and sodium have smaller diffusion rates despite their size, due to a positively charged electrical potential. The pores also have a positive charge that affects the transport of positively charged ions. The transport of substances is primarily due to the substance's individual concentration gradients across the cell membrane. However, substances such as ions, which are suspended in water, are affected not only by their concentrations and charge but also by their diffusivities. The *Nernst equation* can be used to compute the relative electrical potential across a cell as a function of the concentrations of the ions as well as their diffusivities. The Nernst equation for the three primary ions (Na, K, Cl) is as follows:

$$-\mathrm{E} = \frac{\mathrm{RT}}{\mathrm{F}} \ln \frac{\mathrm{D_{Cl}Cl_1} + \mathrm{D_K K_2} + \mathrm{D_{Na}Na_2}}{\mathrm{D_{Cl}Cl_2} + \mathrm{D_K K_1} + \mathrm{D_{Na}Na_1}}$$

where R is the gas constant, T is the absolute temperature, and F is Faraday's constant.

Using the resting concentrations of the ions inside and outside a cell membrane, it is possible to compute the resting electrical potential across a cell. Note that the inside and outside concentrations of the positive ions (Na, K) are opposite to those of the negatively charged ion (Cl). With RT/F = 26.5 mv (at body temperature), the resting transmembrane potential is

$$\mathrm{E} = -26.5 \ln \frac{(1/60)(103) + (1)(141) + (1/75)(10)}{(1/60)(4) + (1)(5) + (1/75)(142)} = -74 \text{ mv}$$

As can be seen, the transport of ions across a cell membrane is affected by their charges and relative diffusivities as much as by their concentration gradients. Note that the diffusivities of all ions are normalized to that of potassium (1).

14.1.3 Mass Transport in Systemic Capillaries

Capillaries, the smallest and most numerous of the blood vessels, form the connection between the vessels that carry blood away from the heart (arteries) and the vessels that return blood to the heart (veins). The primary function of capillaries is the exchange of materials between the blood and tissue cells. Gas exchange occurs across capillary walls in a manner similar to that which occurs across alveoli and pulmonary capillaries. A partial pressure gradient exists for each gas (primarily oxygen and carbon dioxide) that controls the gas exchange process. As with pulmonary capillaries, the gas transfer across systemic capillaries occurs across the entire membrane wall, since the gases are lipid soluble.

However, unlike the pulmonary capillaries, mass transfer across systemic capillaries also includes liquids and ions (dissolved or free floating within the liquid) that travel across pores in the capillary walls. In addition to forming the connection between the arteries and veins, systemic capillaries have a vital role in the exchange of gases, nutrients, and metabolic waste products between the blood and the tissue cells. Substances pass through the capillary wall by diffusion, filtration, and osmosis. Oxygen and carbon dioxide move across the capillary wall by diffusion regulated by the partial pressure differences. Fluid movement across a capillary wall via the pores is determined by a combination of hydrostatic and osmotic pressure. The net result of the capillary microcirculation created by hydrostatic and osmotic pressure is that substances leave the blood at one end of the capillary and return at the other end, as the hydrostatic pressure drops along the length of the capillary, and thus the pressure difference (hydrostatic–osmotic) is different from the beginning of the capillary (arteriole side) to the end (venule side).

Thus, the driving mechanism for this mass transfer is twofold. The pressure (called the *hydrostatic pressure*) is a pushing pressure. Opposite in direction is the *osmotic pressure*, which is a pulling pressure that results from the concentration difference of substances that cannot fit through the pores. These are the nondiffusible components. These two pressure differences acting across the tube-like pores of the capillary are shown in Figure 14.9.

The capillaries form a network of minute vessels between arterioles and venules so as to maximize mass transfer by (a) shortening the distance for mass transfer and (b) maximizing the overall surface area for mass transfer (Figure 14.10). The anatomy of capillaries is well suited to the task of efficient exchange. Capillary walls are composed of a single layer of endothelial cells. The thin nature of the capillary wall facilitates efficient diffusion of oxygen and carbon dioxide, as well as containing short pore lengths to facilitate the bulk motion of liquids and dissolved ions (Figure 14.11).

As was just stated, the hydrostatic pressure in a capillary is higher on the arteriole end of the capillary and lower on the venule end. This is due to the normal pressure gradient along a pipe or tube that produces the axial (down the vessel) flow rate. In fact, a pressure drop is required in order to propel the blood downstream. As a result, the higher hydrostatic pressure near the arteriole end is greater than the osmostic pressure, and the net pressure is outward—pushing fluid out of the capillary. As the hydrostatic pressure drops

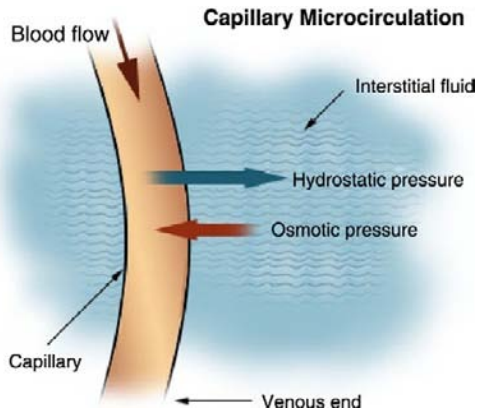


FIGURE 14.9 Mass transfer across systemic capillaries.

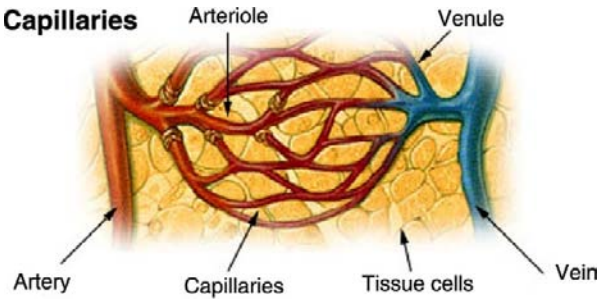


FIGURE 14.10 Capillary network allowing close proximity to tissues, which shortens the distance for mass transfer.

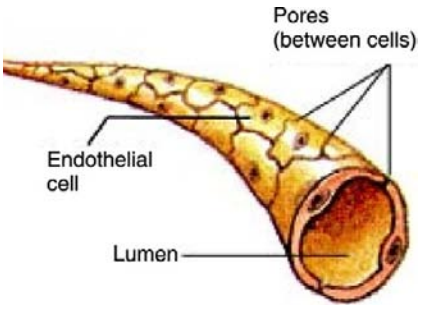


FIGURE 14.11 Thin wall of a systemic capillary that facilitates mass transfer by creating a short distance for transport.

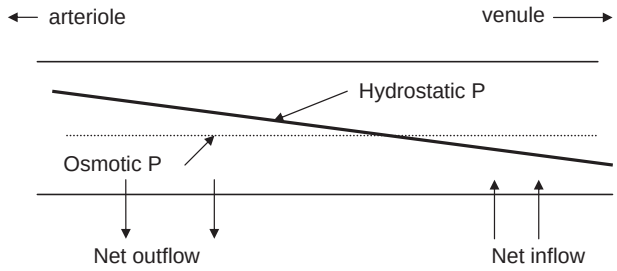


FIGURE 14.12 Relative inflow and outflow of liquids across capillary pores as a function of the relationship between the decreasing hydrostatic pressure and the constant osmostic pressure.

along the length of the capillary, the pressure difference (hydrostatic–osmotic) is reduced and eventually reverses as blood reaches the venule end. Thus, there is a net inflow of fluid at the venule end (into the capillary) (Figure 14.12).

In addition, arterioles are known as the resistance vessels in the circulatory system in that they can constrict or dilate in response to neural and/or hormonal feedback, given their large smooth muscle composition. Arterioles constrict in response to a decrease in arterial pressure and dilate in response to a rise in arterial pressure. When arterioles constrict, they produce an added resistance to blood flow, which drops the hydrostatic pressure downstream. This is because the pressure gradient is proportional to the blood vessel resistance. When the downstream pressure at the end of arterioles is reduced, this corresponds to the inlet pressure to capillaries. Therefore, with arteriolar constriction, the hydrostatic pressure is reduced all along the capillary, which results in less pushing pressure out of the capillary. This produces a net increase in fluid entering the capillary and eventually leaving the capillary via blood flow into the venules. With the increase in fluid level, the blood volume increases, which results in greater venous return of blood into the heart and a resulting increase in cardiac output and arterial pressure (Figure 14.13).

With dilation of the arterioles, the opposite effect is created. The resistance of the arteriole is reduced, which produces a smaller pressure gradient across the arteriole. This then

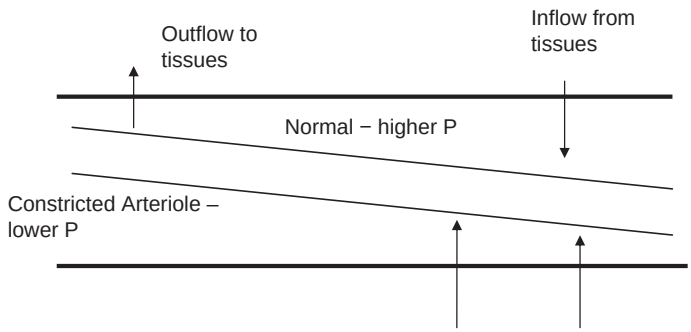


FIGURE 14.13 Hydrostatic pressure gradient inside a capillary for a normal condition and for a constricted arteriole. The hydrostatic pressure is reduced all along the capillary from the resulting reduction at the inlet. This produces a greater inflow of fluid into the capillary.

TABLE 14.4 Relationship between Arteriole Diameter, Capillary Hydrostatic Pressure, Transcapillary Fluid Flow, and Arterial Pressure

Initial Arterial Pressure	Arteriole Diameter Change	Capillary Hydrostatic Effect	Net Fluid	Venous Return	Cardiac/Arterial Output/ Pressure
Low	constrict	reduced	into capillary	increase	increase
High	dilate	increased	out of capillary	decrease	decrease

produces a rise in the inlet hydrostatic pressure to capillaries, which results in a greater pushing pressure for fluid moving out of the capillary and into the tissues and extracellular fluid. With a net increase in fluid leaving the capillaries, there is a reduction in blood volume and in venous return. This results in a reduction in cardiac output and in arterial pressure. Table 14.4 summarizes the cause and effect of changes in arteriole diameter.

When there is an increase in fluid leaving the capillary as a result of arteriole dilation, then the extracellular fluid pressure rises. The increase in fluid volume will eventually cause edema, a buildup of fluid that produces swelling and potential damage to cells and tissues. To alleviate this condition, there is a network of lymphatic vessels that parallels the circulatory system. The peripheral lymphatic vessels are similar in size to capillaries and are networked alongside the capillaries. Their function is to collect excess fluid from extracellular spaces and transport it along the lymphatic vessel network of larger and larger vessels until the lymph reaches the vena cava, where it is returned to the bloodstream. The lymph flow is normally quite small compared with the blood flow. Even the terminal lymph flow is quite small (1 ml/hr) compared with the capillary axial blood flow (1 ml/sec). The terminal lymphatic vessels located in close proximity to systemic capillaries are shown in Figure 14.14.

Propulsion of lymph through the larger vessels is produced by contraction of the walls of the vessels in a fashion similar to that provided for flow through the human intestines and somewhat similar to the action of muscles on veins in the cardiovascular system. Lymph vessels, like the venous system, have valves within the vessels to limit backflow and to promote forward flow along the lymphatic circulation. This is in contrast with flow through the bloodstream, which is produced by a driving hydrostatic pressure gradient from the aorta through arteries, arterioles, capillaries, venules, and veins. The lymph flow originates in the terminal vessels and therefore has no embedded driving pressure gradient. The contraction of the lymph vessels in concert with the added fluid volume (from all terminal lymph vessels joining together within the network) produces the required lymph flow headed to the vena cava.

14.1.4 Mass Transport in the Kidneys and Dialysis

As with the alveoli in the lungs, mass transfer in the kidneys is also processed via a large array of extremely small elements. Each of these elements is a *nephron*. Although nephrons and alveoli represent the essence of mass transfer in the body—small size and large overall surface area—the nephron is quite different from an alveolus. The alveolus is solely

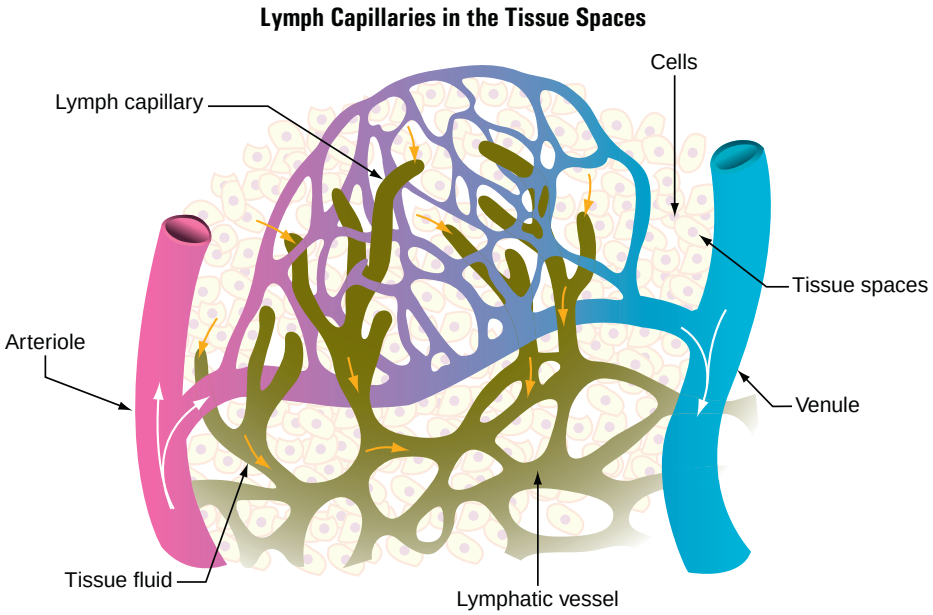


FIGURE 14.14 Lymphatic vessels embedded within the systemic capillary bed.

responsible for mass transfer of gases in one dimension and one direction. Mass transfer in the nephron includes the movement of ions, water, and metabolic waste products, such as urea, uric acid, and creatinine. Furthermore, the nephron initially removes more fluid and ions than the body can safely lose. It is the mechanism by which much of this fluid and ions are reabsorbed into the bloodstream that makes the nephron and the kidneys a unique mass transfer system. There are millions of nephrons in both kidneys (Figure 14.15).

The nephron consists of several segments. The initial filter is the glomerulus, which is embedded inside Bowman's capsule (Figure 14.16).

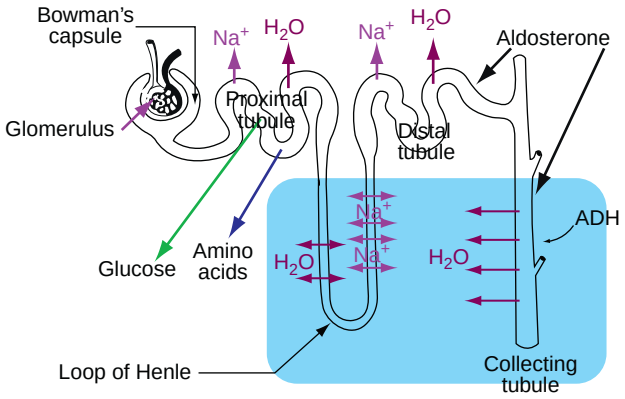


FIGURE 14.15 A nephron in the human kidneys.

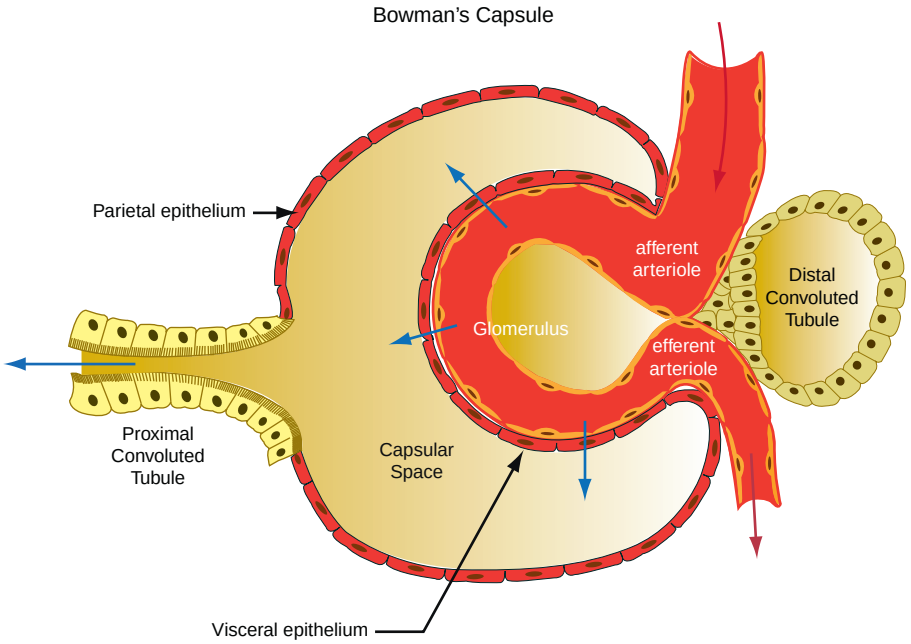


FIGURE 14.16 The glomerulus inside Bowman's capsule.

The glomerulus contains very large pores, each 50 angstroms in diameter and 500 angstroms long. These large pores allow a cumulative 125–150 ml/minute of filtered blood through to the remainder of the nephron. These pores allow water, ions, and metabolic waste products through, but they are too small to allow blood cells, proteins, and large sugars to pass. Once the fluid passes through the glomerulus, it is then known as *filtrate*. Thus, the term *glomerular filtration rate* (*GFR*) corresponds to the 125–150 ml/min traveling through glomerular pores into the tubules of the nephron. The pressure gradient across the glomerulus that causes the filtrate to pass through the glomerular pores is 50 mmHg with a back pressure (osmotic pressure) of 25 mm Hg for a net 25 mm Hg. One can estimate the number of glomeruli in both kidneys by using a model for steady flow through a tube: Poiseuille flow. The Poiseuille flow equation, which will be described later, is

$$Q\,(\mathrm{GFR}) = -n\pi r^4 \Delta P/8\mu L$$

where r is the pore radius, μ is the fluid viscosity, L is the pore length, and n is the number of total pores.

The pore radius is 25 angstroms (25×10^{-8} cm), the pressure gradient is 25 mm Hg (3.33×10^4 g/cm-sec^2), the pore length is 500 angstroms (500×10^{-8} cm), and the filtrate viscosity is 0.012 g/cm-sec.

Obviously, one cannot survive if 125 ml/minute of fluid were to leave the body. Thus, most of this fluid is reabsorbed back into the bloodstream, with the concentrated remainder resulting in urine. Normally, up to 124 ml are reabsorbed, resulting in a urine output of 1 ml/min, which is stored in the bladder.

The remainder of the nephron includes the proximal tubule, the loop of Henle, the distal tubule, and finally the collecting duct, where urine is formed and collected. At each segment of the nephron, water and ions are removed from the tubules to be reabsorbed into the bloodstream. The end result is more concentrated filtrate culminating in urine. Blood hormones control the amount of ions and water that are reabsorbed into the bloodstream, which results in more or less dilute urine. A more detailed description of mass transfer of ions and water from the tubules of the nephron is shown in Figure 14.17.

Parallel to the loop of Henle in the nephron is a loop of capillaries, called the *peritubular capillaries* or *vasa recta*. It is here where the water and ions in the filtrate are reabsorbed into the bloodstream. This parallel network is shown in Figure 14.18.

Mass transfer within the tubules of the nephron is a combination of concentration-driven diffusion as well as active transport. The loop of Henle is configured in a countercurrent approach that allows for some substances that leave the descending loop of Henle to be reabsorbed into the ascending loop in order to maximize the concentration of waste products by controlling water transport. The water transport follows ion transport so as to equalize the concentrations of both ions and water on either side of the tubule. Thus, the concentration at the bottom of the loop of Henle is very different from either top ends (both descending—near the proximal tubule and ascending—near the distal tubule). This is shown in Figures 14.19a and b.

The process by which waste products, water, and excess ions are removed from a failing kidney is *dialysis*. Dialysis is a means of artificially cleaning the blood by accessing the

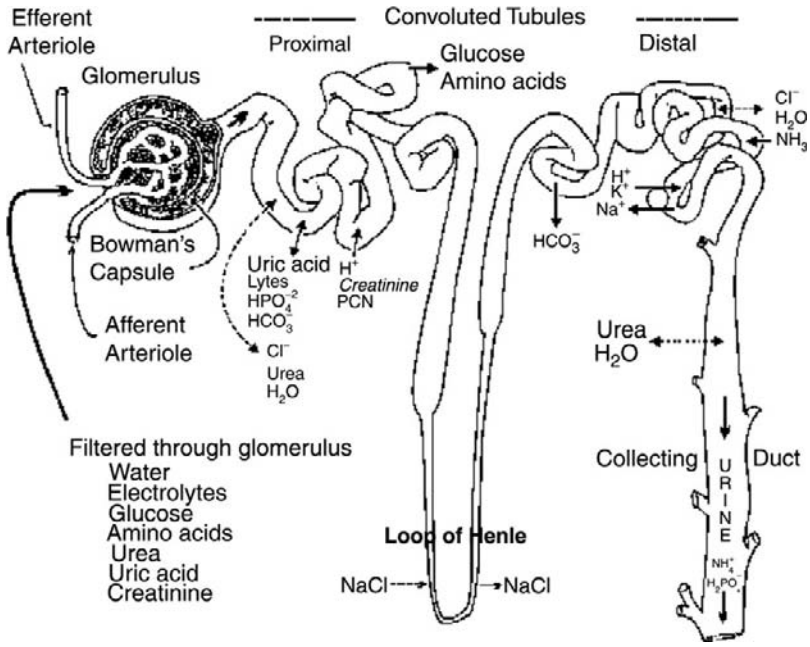


FIGURE 14.17 Mass transfer of ions, water, and waste products along the length of the nephron and its segments.

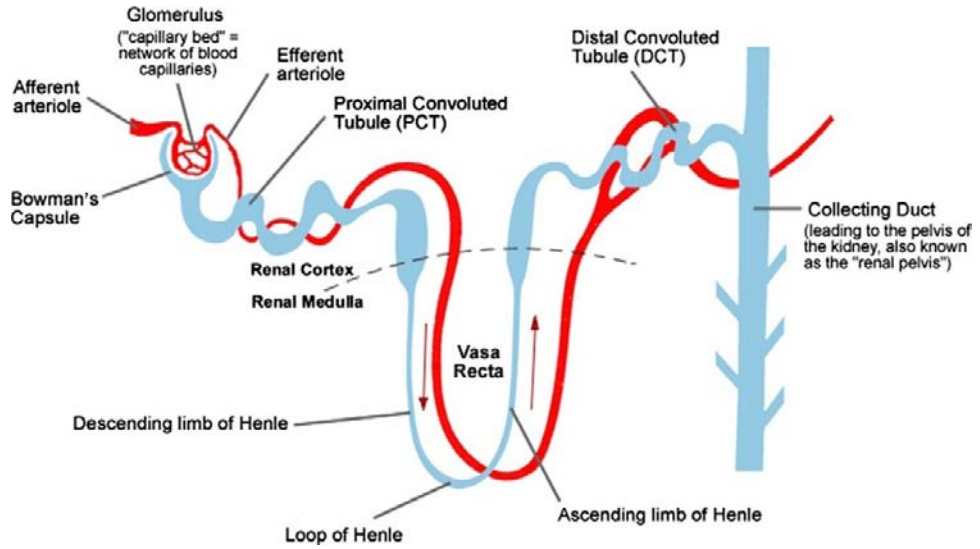


FIGURE 14.18 Vasa recta parallels the loop of Henle for reabsorption into the bloodstream.

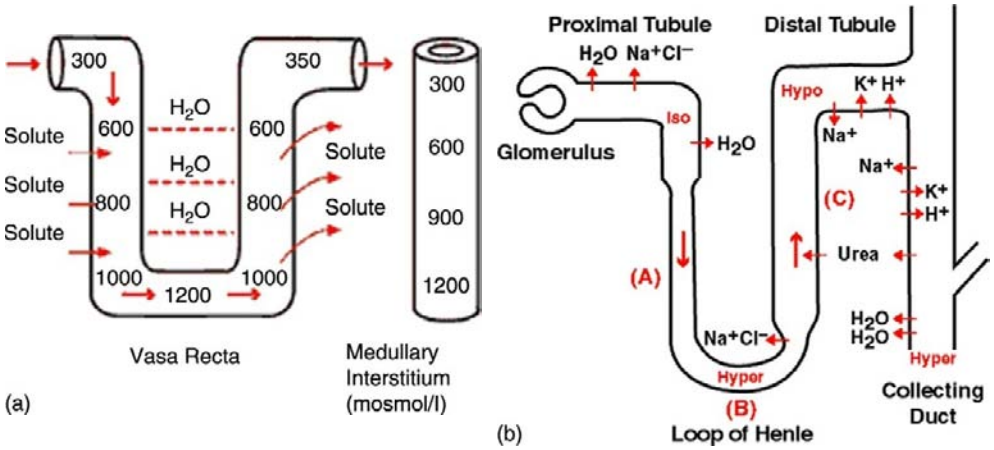


FIGURE 14.19 (a) Changes in osmolarity. (b) Countercurrent system.

radial artery for blood flow into a dialyzer cartridge embedded in a dialysis machine. The resulting clean blood is then returned through the cephalic vein into the body. This is shown Figure 14.20a, with a plastic shunt between the artery and the vein to protect the vessels from continual puncture during chronic dialysis shown in Figure 14.20b.

Within the dialysis cartridge are thousands of cellulose capillary tubes, each with pores. Blood flows inside these capillary tubes, and a concentration-driven diffusion occurs across the pores, which are large enough for transport of water, ions, and the metabolic wastes but too small for blood cells, proteins, or other vital blood components. Inside the cartridge is a

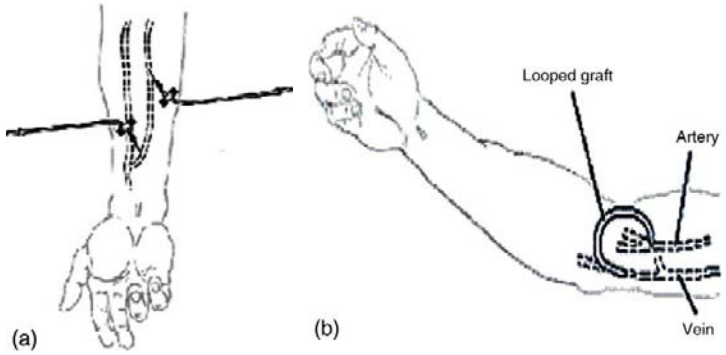


FIGURE 14.20 (a) Blood access for dialysis through the radial artery and the cephalic vein. (b) Plastic shunt connecting the radial artery to the cephalic vein.

fluid that accepts the transported substances. This fluid is called *dialysate*. The process of dialysis is shown in Figure 14.21. The dialysis cartridge is shown in Figure 14.22, which depicts both the blood flow and the dialysate flow pathways.

The mass transport in the radial direction across the dialyzer cartridge capillary tube pores is a concentration-driven diffusion process from blood to dialysate. Blood has a high

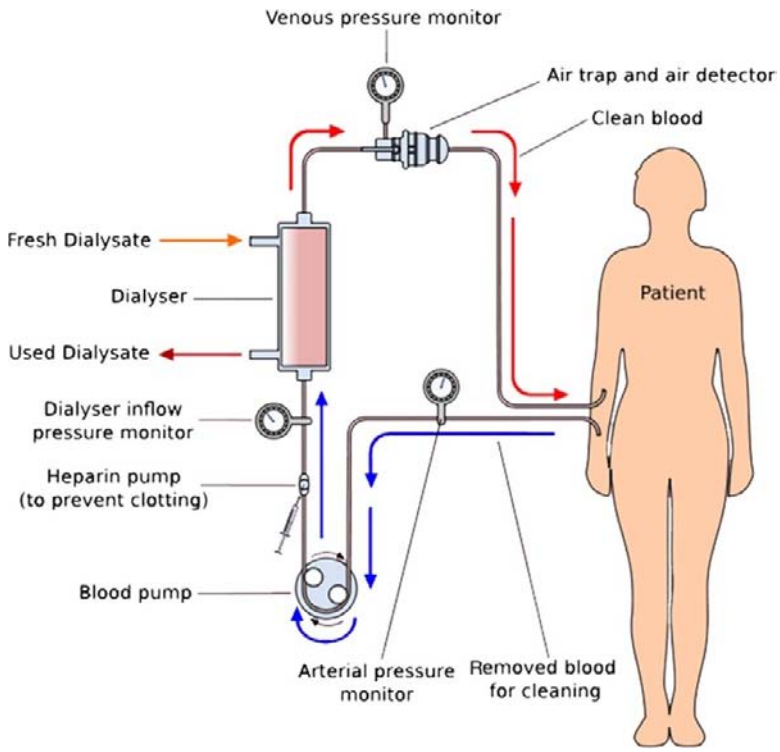


FIGURE 14.21 Dialysis system.

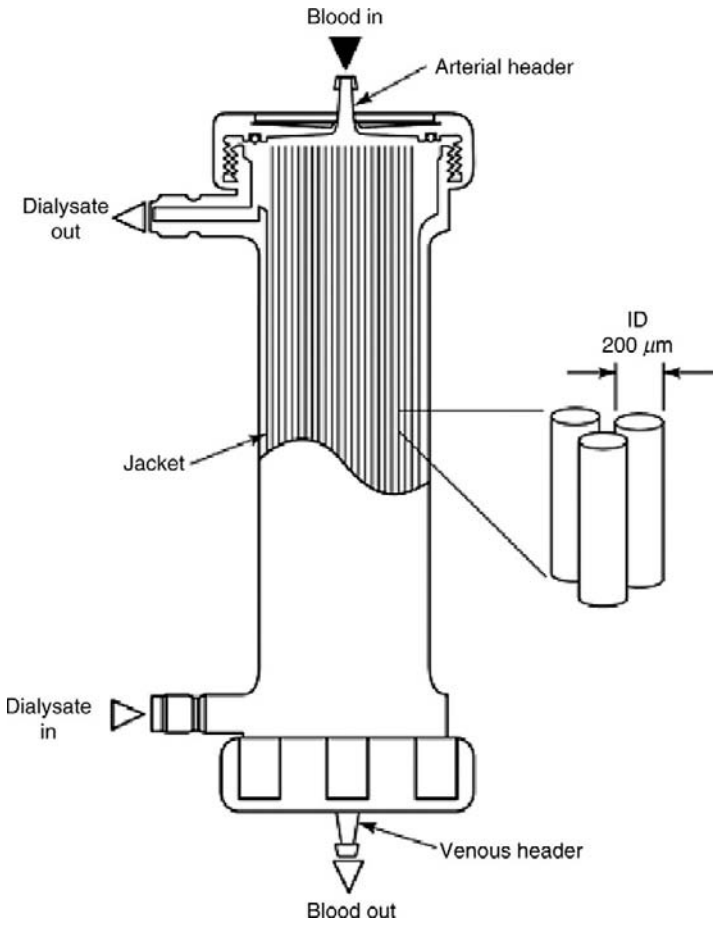


FIGURE 14.22 Dialyzer cartridge depicting blood and dialysate pathways.

concentration of wastes, and dialysate is waste free. Blood has a high concentration of ions, and dialysate has a normal concentration (Figure 14.23).

The solubility of diffused substances within the capillary tube pore (labeled membrane in Figure 14.23) is different from that of the freely moving substances within blood and/or dialysate. Thus, there is a discontinuity in the concentration of substances within the membrane, which is noted by a membrane distribution coefficient. Note that the diffusion is from high concentration within the blood toward the lower concentration within the dialysate. There is also a free flow zone for both blood and dialysate (where the concentration is flat, indicating no mass transfer), as well as a boundary layer near the wall of the capillary tube (where the concentrations are now altered due to mass transfer). The boundary layer produces slower axial flow rates and allows for the radial mass transfer to proceed. The slower flow rates near the wall of the capillary tube within the boundary layer are due to shearing forces within flowing blood. This will be described in the subsequent

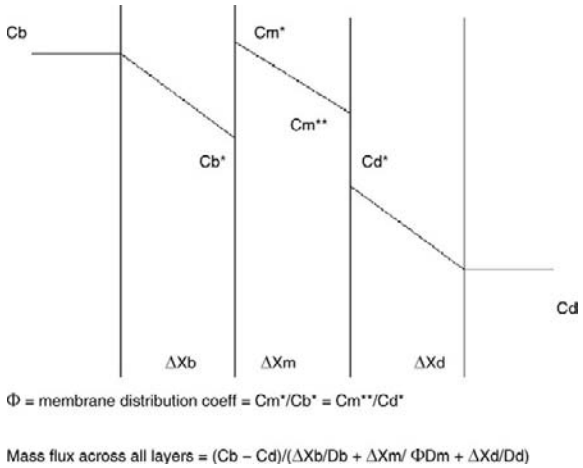


FIGURE 14.23 Diffusion of wastes, ions, and water in the radial direction across a dialysis capillary tube pore.

section on fluid flow (momentum transport). The dialysate flow rate is typically much higher than the blood flow rate in order to move the "dirty" dialysate (now having received the wastes and excess ions from the blood) out of the dialyzer cartridge and bring in clean dialysate. This keeps the concentration gradient for mass transfer large. A typical blood flow rate through the cartridge is 200 ml/min, while the dialysate flow rate is 500–800 ml/min.

Blood entering the top of the dialyzer cartridge is cleared of wastes and excess ions by the time blood leaves through the bottom of the cartridge. However, this blood is then returned to the bloodstream to be mixed with the remainder of the blood. Therefore, in order to completely clean the entire blood supply within the body, dialysis requires four hours to complete.

Fick's Law describes the concentration-driven diffusion between blood and dialysate across the pores of the capillary tubes. As such, each layer of Figure 14.23 indicates a diffusivity (D), a thickness of the diffusion zone (∇X), and a concentration gradient ($C_b - C_b^*$), which are all factors associated with one-dimensional Fick's Law. Each layer has its own diffusivity, since D is a function of both the substance to be transported along with the material through which it moves. Each layer has its own diffusion thickness and concentration gradient. It is true that the mass transfer of a given substance is the same for all three layers, since there is no accumulation of any substance in an individual layer. Therefore, an overall mass transfer rate across all three layers can be derived as shown on the bottom of Figure 14.23 that uses the overall concentration gradient but incorporates the individual diffusivities and individual thicknesses.

Water transport via dialysis cannot be conducted by means of concentration-driven diffusion, since the water concentration gradient is in the wrong direction (from dialysate toward blood). Therefore, water transport from blood to dialysate is controlled by means of a hydrostatic pressure gradient, as shown in Figure 14.24.

The water transport is actually controlled via the dialysate flow rate, per the Bernoulli effect, similar to the way a vacuum cleaner works. By having a large flow rate near an

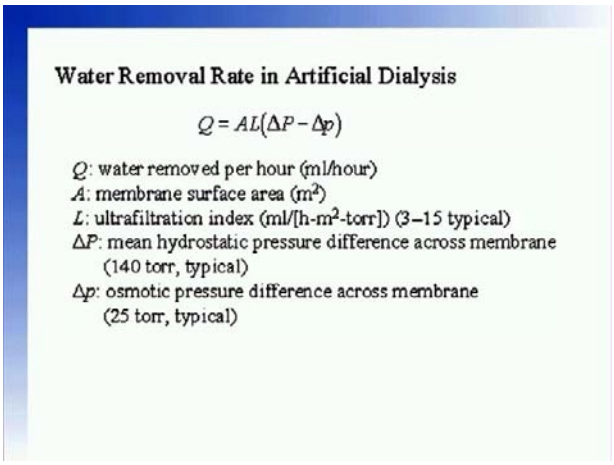


FIGURE 14.24 Water removal during dialysis via a hydrostatic pressure gradient and an osmotic pressure gradient.

opening at right angles to that flow, there is a vacuum created at the opening, thus pulling the water out of the blood and into the dialysate.

As can be seen from this section, mass transfer across membranes, either within the body or within artificial organs, is vital to keeping us alive. Mass transfer allows cells to control nervous impulses, blood chemistry, kidney and lung function, and hemostasis of bodily functions. Mass transport occurs in virtually every cell in the body and is thus evaluated on a smaller scale than that of fluid flow and fluid momentum, which is the subject of the next section.

14.2 BIOFLUID MECHANICS AND MOMENTUM TRANSPORT

We have already seen examples of mass transport of substances across membranes within the human body and through those of artificial organs, such as dialyzers. In addition to mass transport, there are numerous examples of *momentum transport*, with momentum being the product of mass times velocity. This momentum transport is manifested in fluid, such as air flow in the lungs and blood flow through the human circulatory system. Much as there were guiding equations, such as Fick's Law, to govern mass transfer, there are also guiding equations to govern fluid flow. In addition, much as there were parameters that affected mass transfer and were properties of a given material (diffusivity is an example), such is also the case for momentum transfer/fluid mechanics. Like diffusivity is a proportionality constant for mass transfer, so is *viscosity* a proportionality constant for fluid mechanics. In mass transfer, the driving force was either a concentration gradient (for liquids or solutes embedded in solvents) or a partial pressure gradient (for gases). For fluid flow and momentum transport, the driving force is a pressure gradient. This is not a partial pressure gradient but rather a standard pressure gradient, often called the hydrostatic pressure.

However, the total pressure gradient for a flowing fluid is the sum of the hydrostatic pressure, the dynamic pressure (related to the fluid velocity), and the pressure associated with gravity and a height differential. The gravity component is easily described by the height difference in a fluid manometer. An example of a manometer, which can be used to measure a pressure difference as a result of the density and gravitational difference between two fluids, is shown in Figure 14.25.

The total pressure of a flowing fluid is given by

$$P_T = P\,(\text{hydrostatic}) + \tfrac{1}{2}\rho V^2 + \rho g h$$

Where the hydrostatic pressure is what is normally thought of as "pressure," the dynamic pressure is given by ½ ρV^2, and the last term is the pressure as a result of a height differential and is measured by a manometer. A pressure transducer can be used to measure the hydrostatic pressure, a flowmeter is used to measure the fluid velocity (flow rate divided by vessel cross-sectional area) resulting in the dynamic pressure, and a manometer for the gravity term. Depending on the density of the fluid and the value for fluid velocity, it is possible that one of the three pressure components can far outweigh the others. The Bernoulli equation is a modification of the total pressure equation. It states that the total pressure at two points is the same or that the sum of the three pressures is equal, as shown following.

$$P_1 = \tfrac{1}{2}\rho {V_1}^2 + \rho g h_1 = P_2 + \tfrac{1}{2}\rho {V_2}^2 + \rho g h_2$$

14.2.1 Blood and Fluid Viscosity

Viscosity is a measure of the resistance of a flowing fluid that is being deformed by either shear stress or normal stress. In general terms, it is the resistance of a liquid to flow or its "thickness." Viscosity describes a fluid's internal resistance to flow and may be thought

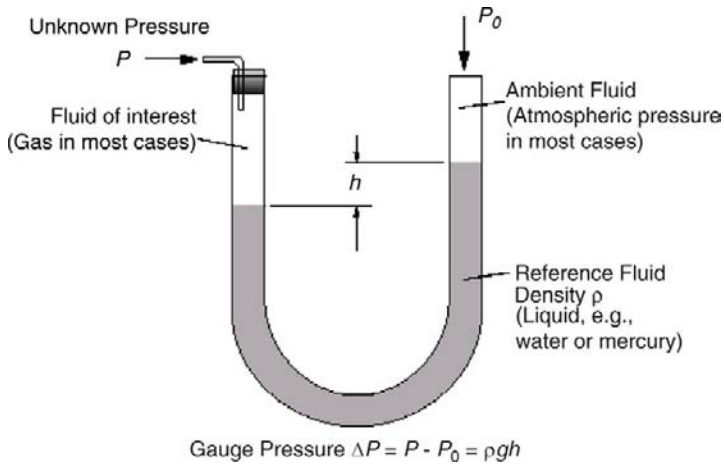


FIGURE 14.25 Fluid manometer used to measure the density/gravitational pressure difference between two fluids with varying heights within the manometer. The pressure gradient is a function of the density multiplied by the height difference in the two fluids.

of as a measure of fluid friction. Thus, water is "thin," having a lower viscosity, while vegetable oil is "thick," having a higher viscosity. All real fluids (except superfluids, such as liquid nitrogen) have some resistance to stress and fluid flow, but a fluid that has no resistance to shear stress is known as an ideal fluid or inviscid fluid. A real fluid with a non-zero viscosity is called a viscous fluid. The study of viscosity of a real fluid is known as rheology.

Isaac Newton postulated that for straight, parallel, and uniform flow, the shear stress, τ, between fluid layers is proportional to the velocity gradient, $\partial u/\partial y$, in the direction perpendicular to the layers.

$$\tau = \mu \frac{\partial u}{\partial y}$$

Here, the constant μ is known as the *coefficient of viscosity*, the *viscosity*, the *dynamic viscosity*, or the Newtonian viscosity. Many fluids, such as water and most gases, satisfy Newton's criterion and are known as Newtonian fluids. Non-Newtonian fluids exhibit a more complicated relationship between shear stress and velocity gradient than simple linearity. Many such non-Newtonian fluids have a high concentration of suspended particles. Blood may be one such fluid, since blood contains almost 50 percent of its volume as blood cells. In the metric system, also known as the cgs (centimeter-gram-second) system, the unit of viscosity is the *poise*, named after Jean Louis Marie Poiseuille, who formulated Poiseuille's law of viscous flow. The value of viscosity is so small for many fluids (liquids and gases) that their viscosity is often listed in cP, centipoise, or one-hundredth of a poise. A poise is equal to one dyne/cm-sec.

As an example, the viscosity of water is approximately 1 cP with a density of 1 g/cc. Blood has a viscosity of 3–4 cP, with a density of 1.05 g/cc. As can be seen, although the densities of water and blood are fairly similar, their viscosities are not. This is due to the large suspension of cells within blood, which adds to the resistance to blood flow. The more solid/liquid boundaries in a fluid, where the shearing stresses tend to be higher, the higher the viscosity. The cells within blood add numerous "solid"/liquid boundaries. Denser pure fluids, such as honey and molasses, are thicker fluids with a much higher viscosity than water or blood. When a pure liquid (without a suspension of particles) has a higher density, its viscosity also tends to be proportionally higher.

Fluids with large suspensions, such as blood, have viscosities that are proportional to the concentrations of particles. For blood, this is the hematocrit, which is the percentage of red cells by volume in whole blood. Normally, the hematocrit is 45 percent. Figure 14.26 depicts the relationship between blood hematocrit and viscosity.

The blood consists of a suspension of cells in a liquid called *plasma.* In an adult man, the blood is about $\frac{1}{12}$ the body weight, and this corresponds to 5–6 liters. Blood consists of 55 percent plasma and 45 percent cells, sometimes called *formed elements*. The blood performs many important functions. By means of the hemoglobin contained in the erythrocytes (red cells), it carries oxygen to the tissues and collects the carbon dioxide. It also conveys nutritive substances (e.g., amino acids, sugars, mineral salts) and gathers the excreted material that will be eliminated through the kidney. The blood also carries hormones, enzymes, and vitamins. It performs the defense of the body by means of the phagocytic activity of the leukocytes (white cells).

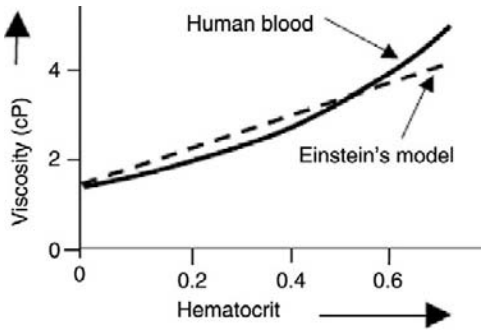


FIGURE 14.26 Blood viscosity as a function of blood hematocrit. Albert Einstein developed a model of viscosity of a fluid with a large concentration of substances where $c > 10\%$.

The erythrocytes are the most numerous blood cells, about 4–6 millions/mm^3. They are also called red cells. In humans and all mammals, erythrocytes are devoid of a nucleus and have the shape of a biconcave lens. The red cells are rich in hemoglobin, a protein able to bind to oxygen. Hence, these cells are responsible for providing oxygen to tissues and partly for recovering carbon dioxide produced as waste. However, most CO_2 is carried by plasma in the form of soluble carbonates. Red cells, which are the most prolific cells in the bloodstream, form the basis of the hematocrit, which is a factor in the determination of blood viscosity. A typical red cell is shown in Figure 14.27. The other blood cells are the leucocytes (white cells), which are designed to combat infections, and the thrombocytes (platelets), which are designed to assist in the clotting process.

White blood cells are colorless without hemoglobin. They contain a nucleus and have an irregular shape. Though there are fewer white blood cells than red blood cells, they are much bigger in size. They can change their shape easily, and this allows them to squeeze through the walls of the blood vessels into the intercellular spaces. Unlike red blood cells or platelets, there are five different types of white blood cells, each serving a different purpose in our body's immune system. Neutrophils make up 55 to 70 percent of the total white

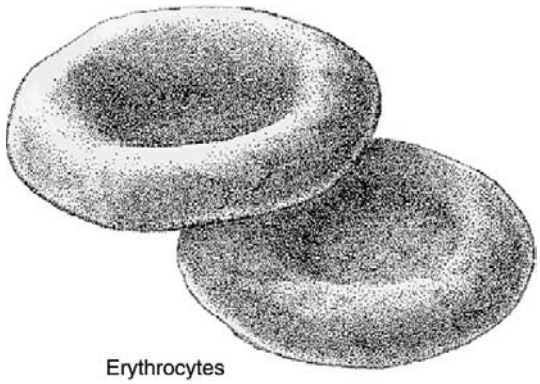


FIGURE 14.27 Red cells consisting of biconcave disks, each 8 microns in diameter and 2 microns thick.

blood count in the bloodstream. They have segmented nuclei and are "C" shaped. Neutrophils can be most commonly found near sites of infection or injury, where they will stick to the walls of the blood vessels and engulf any foreign particles that try to enter the bloodstream. They can also be found in the pus of wounds.

Eosinophils make up 2 to 5 percent of the total blood count and mainly attack parasites and any antigen complexes. These cells are also responsible for allergic response within the blood. Basophils make up less than 1 percent of the total white blood count. They secrete anticoagulant and antibodies, which mediate hypersensitivity reactions within the blood. They are known to have phagocytory features, although they are more often related to immediate immune reaction against external germs and diseases. Monocytes, having only 5 to 8 percent of the total white blood count, are the largest of the five types of white blood cells. They act as tissue macrophages and remove foreign particles and prevent the invasion of germs that cannot be effectively dealt with by the neutrophils. They have been known to have phagocytic functions.

Lymphocytes produce antibodies against toxins secreted by bacteria and infecting germs. These antibodies will be excreted into the plasma to kill bacteria in the blood as well as act as antitoxins. These antibodies will cause the foreign particles to cluster together, rendering them easily engulfed by the phagocytes. However, the nature of lymphocytes is highly specific, and they can recognize only certain antigens. Figure 14.28 shows several white cell subtypes.

Platelets are only about 20 percent of the diameter of red blood cells, the most numerous cell of the blood. The normal platelet count is 150,000–350,000 per microliter of blood, but since platelets are so small, they make up just a tiny fraction of the blood volume. The principal function of platelets is to prevent bleeding. Platelets are produced in the bone marrow, the same as the red cells and most of the white blood cells. Platelets are actually not true cells but merely circulating fragments of cells. But even though platelets are merely cell

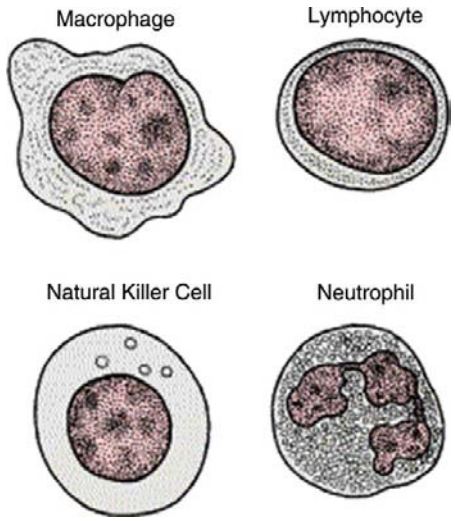


FIGURE 14.28 Sizes and shapes of various white cell types.

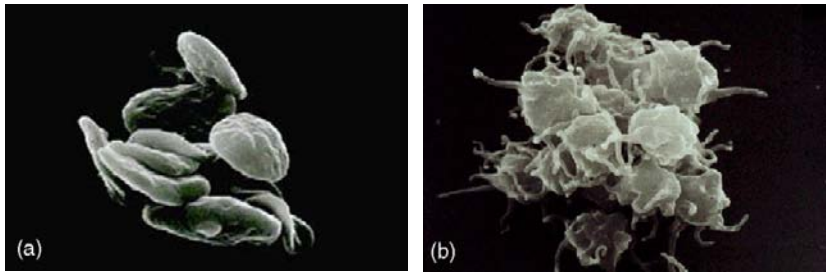


FIGURE 14.29 (a) Normal platelet. (b) Activated platelet.

fragments, they contain many structures that are critical to stop bleeding. They contain proteins on their surface that allow them to stick to breaks in the blood vessel wall and also to stick to one another. They contain granules that can secrete other proteins required for creating a firm plug to seal blood vessel breaks. Also, platelets contain proteins similar to muscle proteins that allow them to change shape when they become sticky. Figures 14.29a and b depict a normal platelet and an activated platelet. The normal platelet is shaped like a plate, thus lending itself to the name. An activated platelet extends tentacles to allow it to adhere to surfaces where bleeding exists.

The measurement of fluid viscosity is conducted with a device known as a viscometer. For liquids with viscosities that vary with flow conditions, an instrument called a rheometer is used. Viscometers only measure viscosity under one flow condition. In general, either the fluid remains stationary and an object moves through it, or the object is stationary and the fluid moves past it. The drag caused by relative motion of the fluid and a surface is a measure of the viscosity.

A falling sphere viscometer is one of the simpler types of viscometers. It is based on Stokes's law, which describes the flow of a fluid past a sphere in a vertical tube, where the flow is very slow. A sphere of known size and density is allowed to descend through the liquid. If correctly selected, it reaches terminal velocity, which can be measured by the time it takes to pass two marks on the tube. Knowing the terminal velocity, the size and density of the sphere, and the density of the liquid, Stokes's law can be used to calculate the viscosity of the fluid. A series of steel ball bearings of different diameters are normally used in the classic experiment to improve the accuracy of the calculation. Stokes's law is given as

$$F = 6\pi r\eta v$$

where

F = the frictional force
r = the radius of the spherical object
η = the fluid viscosity
v = the particle's velocity

An example of a functional relationship for a falling sphere viscometer is shown in Figure 14.30.

Since blood is opaque, sensors within the tube indicate when the sphere passes points along the tube. Another type of viscometer is the *Couette* or concentric cylinder viscometer.

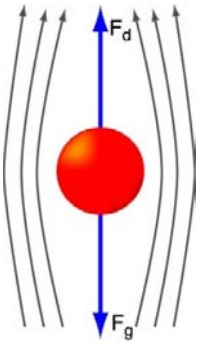


FIGURE 14.30 Slow flow past a falling sphere, which forms the basis of a falling sphere viscometer.

The fluid is placed between two cylinders with large radii and with a gap width between the cylinders that is very small compared to the radii. One of the cylinders rotates while the other remains stationary. With a small gap width, the velocity between the cylinders approaches a linear profile. The torque required to spin the cylinder as a function of the cylinder rotation rate provides the fluid viscosity. Different fluids would have different linear flow profiles between the cylinders, indicating different values of viscosities. A typical concentric cylinder viscometer is shown in Figure 14.31.

14.2.2 Conservation of Mass

The equation that describes fluid flow as a function of the pressure gradient in a tube is used to describe blood flow in a blood vessel or air flow in the trachea and bronchioles. As was the case with diffusion occurring in the direction of decreasing concentration gradient, so is the case of fluid flow in the direction of decreasing pressure in a blood vessel. In the human systemic circulation, the aortic pressure is 120 systolic pressure over 80 diastolic pressure (in mm Hg). This is because blood flow is pulsatile when it comes from the aortic valve as the ventricles undergo systolic contraction followed by diastolic relaxation. The blood pressure is 100 mm Hg mean and eventually decreases to about 5–10 mm Hg at the vena cava. Thus, there is a 90–95 mm Hg pressure gradient that pushes blood around the systemic circulation. Blood flow in the aorta is highly accelerated and time varying and continues around the sharp bend of the aortic arch. The equations that describe blood flow and fluid mechanics are time varying in nature, as well as affected by spatial variation due to the changing geometry/anatomy of the circulatory system. These equations include the equation relating to the conservation of mass (sometimes called the continuity equation) and the equations of fluid momentum (sometimes called the equations of fluid motion).

As was just noted, the type of fluid (Newtonian or non-Newtonian) can affect the equations related to fluid mass and fluid momentum. For simplicity, we can assume that blood acts Newtonian and can then alter that assumption after deriving the preceding equations. The continuity equation for a Newtonian fluid is shown in the following equation. We are using the version of this equation in cylindrical coordinates because blood vessels are cylinders. Thus, the velocities associated with this coordinate system are in the z direction (axially downstream),

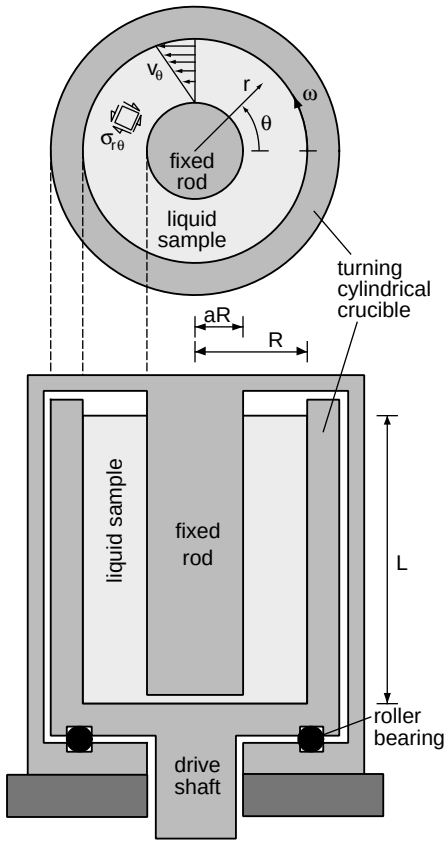


FIGURE 14.31 Concentric cylinder viscometer.

the r direction (radially from the center to the edge of the vessel), and the theta direction (circumferentially). This equation assumes that the density is a constant. A constant density assumes the fluid to be incompressible, which is a realistic assumption for most liquids (such as blood) and is unrealistic for most gases, which can compress at higher pressures.

$$\frac{1}{r}\frac{\partial}{\partial r}(ru_r) + \frac{1}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{\partial u_z}{\partial z} = 0.$$

Because mass is a scalar quantity, there is only one conservation of mass equation, although there are three terms in that equation, relating the derivatives of the three orthogonal velocities.

14.2.3 Fluid Momentum, the Equations of Motion, and the Navier Stokes Equations

The Navier-Stokes equations, named after Claude-Louis Navier and George Gabriel Stokes, describe the motion of viscous fluids such as liquids and gases and therefore the

conservation of fluid momentum. These equations arise from applying Newton's second law to fluid motion, together with the assumption that the fluid stress (τ in the preceding section describing viscosity) is the sum of a diffusing viscous term (proportional to the gradient of velocity), plus a pressure term. As was the case for the continuity equation, the Navier-Stokes equations may be written in various coordinate systems, but the cylindrical coordinate system is most appropriate for blood flow, since blood vessels are cylinders.

$$\rho\left(\frac{\partial u_r}{\partial t}+u_r\frac{\partial u_r}{\partial r}+\frac{u_\theta}{r}\frac{\partial u_r}{\partial \theta}+u_z\frac{\partial u_r}{\partial z}-\frac{u_\theta^2}{r}\right)=-\frac{\partial p}{\partial r}+\mu\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u_r}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u_r}{\partial \theta^2}-\frac{\partial^2 u_r}{\partial z^2}-\frac{u_r}{r^2}-\frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta}\right]+\rho g_r$$

$$\rho\left(\frac{\partial u_\theta}{\partial t}+u_r\frac{\partial u_\theta}{\partial r}+\frac{u_\theta}{r}\frac{\partial u_\theta}{\partial \theta}+u_z\frac{\partial u_\theta}{\partial z}-\frac{u_r u_\theta}{r}\right)=-\frac{1}{r}\frac{\partial p}{\partial \theta}+\mu\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u_\theta}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u_\theta}{\partial \theta^2}+\frac{\partial^2 u_\theta}{\partial z^2}+\frac{2}{r^2}\frac{\partial u_r}{\partial \theta}-\frac{u_\theta}{r^2}\right]+\rho g_\theta$$

$$\rho\left(\frac{\partial u_z}{\partial t}+u_r\frac{\partial u_z}{\partial r}+\frac{u_\theta}{r}\frac{\partial u_z}{\partial \theta}+u_z\frac{\partial u_z}{\partial z}\right)=-\frac{\partial p}{\partial z}+\mu\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u_z}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u_z}{\partial \theta^2}-\frac{\partial^2 u_z}{\partial z^2}\right]+\rho g_z$$

The Navier-Stokes equations also assume a constant density as well as a constant viscosity, indicating not only an incompressible fluid but a Newtonian one as well. Note that there are three fluid flow equations, since fluid momentum is a vector quantity, the product of mass and velocity, with velocity being a vector quantity with three components. Also note that each of the three fluid flow equations also contain multiple elements owing to the three derivatives of the fluid velocity.

14.2.4 Poiseuille Flow

As we just saw, even with the assumption of a Newtonian fluid, these equations are time varying with partial derivatives and are difficult to solve. As such, the Poiseuille assumptions, named after Jean Louis Marie Poiseuille, who formulated Poiseuille's law of viscous flow (as mentioned previously), are as follows:

1. The flow is steady ($\partial(\ldots)/\partial t = 0$).
2. The radial and swirl components of the fluid velocity are zero ($u_r = u_\theta = 0$).
3. The flow is axisymmetric ($\partial(\ldots)/\partial\theta = 0$) and fully developed ($\partial u_z/\partial x = 0$).

As a result, the first two of the three Navier-Stokes momentum equations and the continuity equation are identically satisfied. The third (z) momentum equation reduces to

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u_z}{\partial r}\right)=\frac{1}{\mu}\frac{\partial p}{\partial z}$$

The two boundary conditions for the flow of a viscous fluid (such as blood) are $(dV_z/dr) = 0$, when $r = 0$ (center of blood vessel), and $V_z = O$ at $r = R$ (edge of tube), which are the result of axisymmetric flow and viscous "no slip at the wall," respectively.

The solution is

$$u_z=\frac{1}{4\mu}\frac{\partial p}{\partial z}r^2+c_1\ln r+c_2$$

Since u_z needs to be finite at $r = 0$, $c_1 = 0$. The no slip boundary condition at the pipe wall requires that $u_z = 0$ at $r = R$ (radius of the pipe), which yields

$$c_2 = -\frac{1}{4\mu}\frac{\partial p}{\partial z}R^2.$$

Thus, we have finally the following parabolic velocity profile:

$$u_z = -\frac{1}{4\mu}\frac{\partial p}{\partial z}(R^2 - r^2).$$

The maximum velocity occurs at the pipe centerline ($r = 0$):

$$u_{z\,max} = -\frac{R^2}{4\mu}\left(-\frac{\partial p}{\partial z}\right).$$

The blood flow is thus a paraboloid with a maximum velocity at the center of the blood vessel with a zero velocity at the wall, as shown in Figure 14.32. The velocity, varying in the radial direction, can be thought of as flowing within concentric layers, each with a different velocity at a different radial value. The zero value at the wall corresponds to the no slip condition associated with Poiseuille flow. The fluid velocity is symmetric about the center of the tube, corresponding to the axisymmetric assumption associated with Poiseuille flow. The no slip condition is a result of the flow of a viscous fluid against a nonmoving boundary.

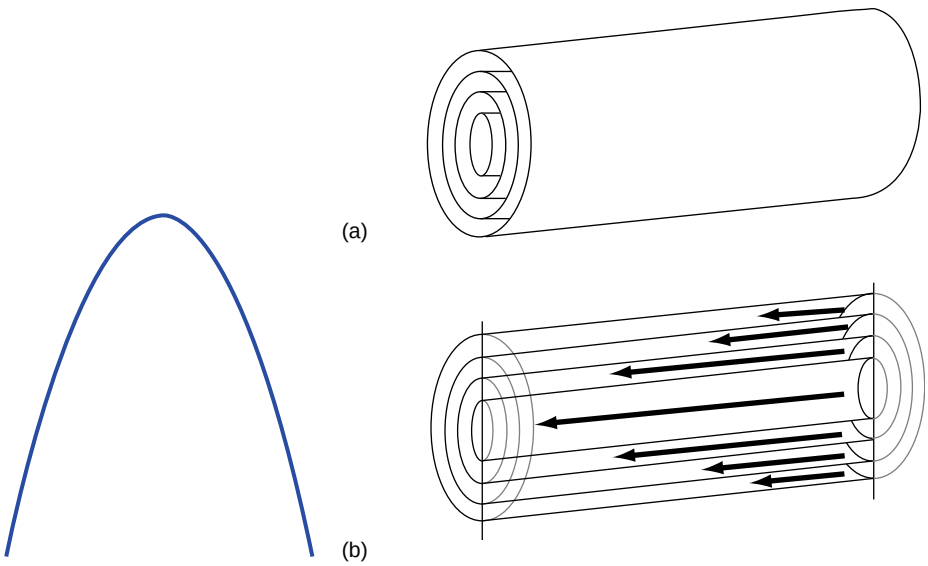


FIGURE 14.32 The parabolic velocity profile associated with the solution of the Navier-Stokes equations for Poiseuille flow. The velocity is maximal at the center of the vessel and zero at the wall. The velocity is axisymmetric, corresponding to the associated Poiseuille assumption.

If one integrates the velocity over the cross-sectional area of the blood vessel, this results in the well-known Poiseuille law for steady, incompressible, axisymmetric flow in a cylindrical vessel

$$Q = \pi R^4 \nabla P / 8\, \mu l$$

where Q is the volume flow rate (cardiac output for blood), R is the vessel radius, ∇P is the pressure gradient, μ is the blood viscosity, and l is the length of the blood vessel over which the pressure gradient is acting.

Poiseuille's law (sometimes called the Hagen-Poiseuille law) states that assuming laminar flow in a tube, flow (Q) is proportional to its radius to the fourth power (R^4) and inversely proportional to its length (l) and viscosity of the fluid (μ). The general principle of this law is that small changes in the internal diameter of a blood vessel lumen can make a big difference to the rate of flow. The flow in the aorta is large—in part because the radius of the aorta is large. The flow in smaller vessels is less, not only because the branching blood vessels segment the available flow but also because the vessel radius is smaller. As already stated, when the hematocrit is smaller, the blood viscosity is as well. As a result, an anemic patient has a smaller blood viscosity, and therefore, according to the Poiseuille law, the flow rate is larger. The resistance to blood flow (Poiseuille) corresponds to the relationship between volume flow rate (Q) and the pressure gradient (∇P) by the equation

$$Q = \nabla P / R_s$$

where R_s is the resistance to flow. If one uses the Poiseuille flow equation, then the flow resistance takes the form of

$$R_s = 8\mu l / \pi R^4$$

Thus, the resistance to flow inside a given blood vessel is proportional to the fluid viscosity and inversely related to the vessel radius to the fourth power.

Throughout the analysis of fluid flow in a vessel, the driving mechanism has been the pressure gradient (∇P). An example of the relationship between pressure and flow is the relationship between left ventricular pressure in the heart, left atrial pressure, and aortic pressure. The heart valves associated with the left heart (mitral valve between the left atrium and left ventricle, and aortic valve between the left ventricle and the aorta) open and close via the difference in pressure across each valve. This is evident in Figure 14.33, which shows the pressure and timing relationship across the two left heart valves.

Another manifestation of pressure drop producing a fluid flow is the relationship between pressure across the systemic circulation, resulting from flow across the aortic valve, through the aorta, arteries, arterioles, capillaries, venules, veins, and the vena cava. The pressure across the systemic circulation begins as a pulsatile pressure in the aorta (at a mean pressure of about 100 mm Hg) and ends at the vena cava with a mean pressure of about 5 mm Hg. This is shown in Figure 14.34, along with the smaller pressures associated with the right heart and the pulmonary circulation.

14.2.5 Boundary Layers

The Poiseuille assumption of fully developed flow ($\partial u_z / \partial z = 0$) refers to the fluid boundary layer, the region of the flow field where viscous forces predominate. What is "fully

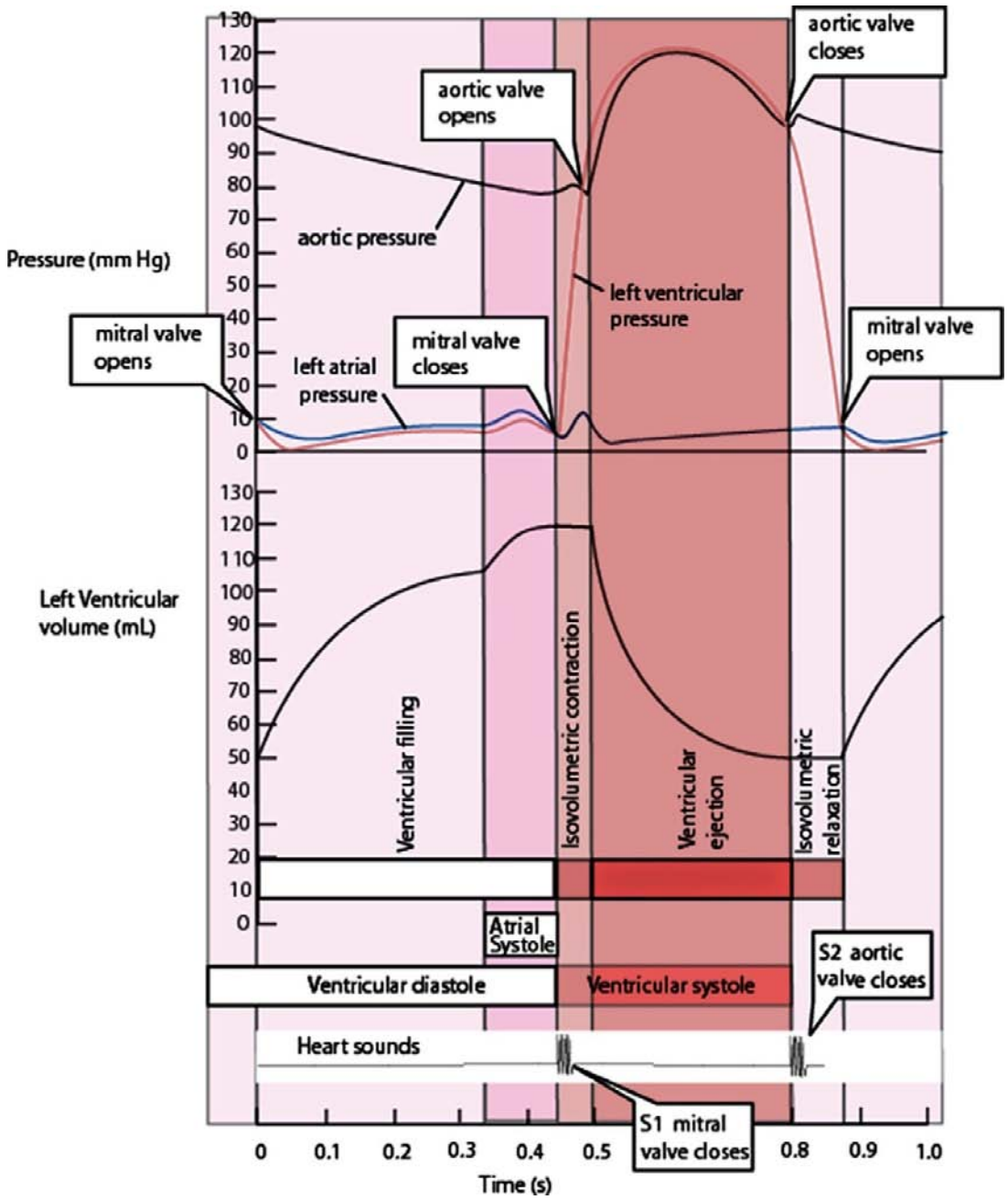


FIGURE 14.33 Pressure relationships across the left heart, which produce flow through the left atrium and left ventricle (across the mitral valve), or through the left ventricle and aorta (across the aortic valve).

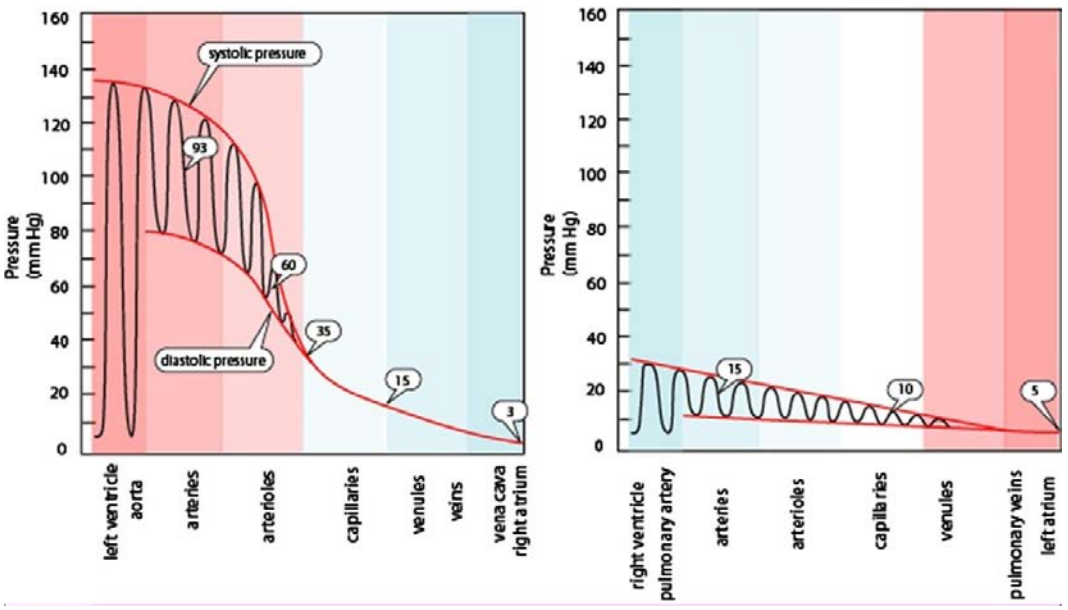


FIGURE 14.34 Pressure reduction across the systemic circulation (left side) and the pulmonary circulation (right side) indicating the continual pressure gradient that drives blood flow along the circulatory systems.

developed" is the boundary layer. Thus, for Poiseuille flow, the entire flow field is viscous dominated. This is manifested by the parabolic flow profile just mentioned. However, in reality, the fluid flow in the aorta is not truly parabolic but is more of a blunt profile. The flow profile becomes more rounded as the boundary layer grows from the outside of the vessel toward the center.

From a momentum transport perspective, the outside of the blood vessel, where the boundary layer is growing, is said to undergo convective deceleration, while in the center of the vessel, the flow is undergoing convective acceleration. Thus, there is a transfer of fluid momentum from the outside of the vessel toward the inside until finally the entire flow pattern is completely rounded and parabolic. The growth of the boundary layer is shown in Figure 14.35 for laminar flow, which is flow of a fluid that is smooth and orderly.

The boundary layer thickness is given by the symbol δ, whose value is quite small during "plug" flow at the entrance to a blood vessel but that slowly grows during viscous flow until it equals the vessel radius. At that point, the velocity profile is fully parabolic and the flow is said to be "fully developed." This is the point where the Poiseuille assumptions are now valid. Note that the pulsatile nature of blood flow in an artery has been neglected in this analysis and in Figure 14.35.

When the fluid flow is turbulent, the flow profile is more chaotic. This results in more fluid mixing from the areas of higher-velocity flow at the center with lower-velocity flow at the periphery of the vessel (toward the blood vessel wall). Thus, turbulent flow has more mixing and therefore more momentum transport between the center of the vessel and the wall. In laminar flow, the fluid shear stress is highest at the wall and zero at the center. Even so,

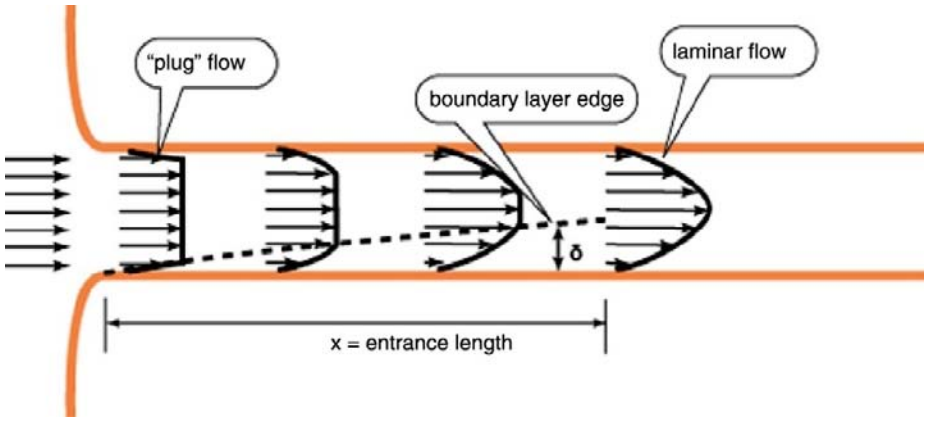


FIGURE 14.35 Growth of the laminar boundary layer from a blunt profile on the left toward a more rounded profile on the right, where the boundary layer becomes fully developed.

the laminar wall shear stress is not huge. However, in turbulent flow, the higher velocities are convected toward the wall, and the wall shear stress is higher than that of laminar flow. The turbulent boundary layer grows at a different rate than that for laminar flow, and the fully developed flow profile is not parabolic but more blunt in shape (Figure 14.36).

Turbulent flow has more energy and therefore a more chaotic flow as compared to the relatively smooth profile of laminar flow. With the added energy of turbulent flow, there are fluctuating components of the flow that can be periodically large. Turbulent flow is normally three-dimensional and time varying, while laminar flow is normally two-dimensional and time varying only during boundary layer development in the cardiovascular system. As such, it is possible to hear turbulent flow, and this approach often forms the basis to determine the prevalence of fluid turbulence in the cardiovascular system.

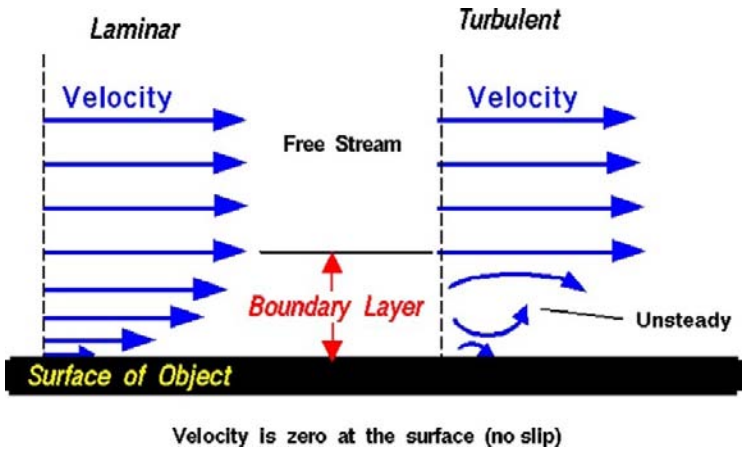


FIGURE 14.36 Laminar and turbulent boundary layers with laminar version more rounded than turbulent version. Turbulent version has greater mixing and greater momentum transport.

If the boundary layer of either type of flow separates from the wall, there is a zone of stagnant fluid. For blood, this would be a zone where thrombosis (clotting) would occur. Such a clot might become dislodged, where it would be transported downstream to lodge in a smaller blood vessel. This might result in a stroke if the dislodged clot (called an embolism) were to become trapped in the cerebral (brain) circulation. Such boundary layer separation might occur at sharp bends where the fluid flow cannot adequately follow the curvature (Figure 14.37). Boundary layer separation might also occur at branches of blood vessels as shown in Figure 14.38.

There is certainly a danger in boundary layer separation owing to the potential for blood clotting. Separation also occurs downstream of an atherosclerotic plaque that adheres to the vessel wall and narrows the vessel diameter. As the diameter becomes smaller with growth of the plaque, there is separation downstream of the stenosis (narrowing of the vessel), which creates clotting and further promotes the stenosis downstream. Thus, at some point in the progression of arterial disease, there is a tipping point where the stenosis becomes

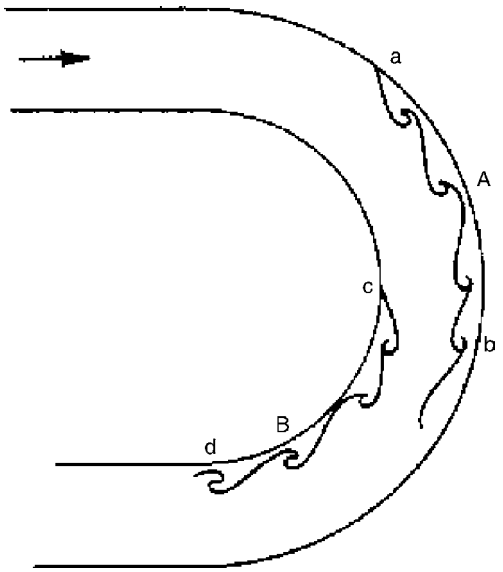


FIGURE 14.37 Boundary layer separation at a tight curve, such as the aortic arch.

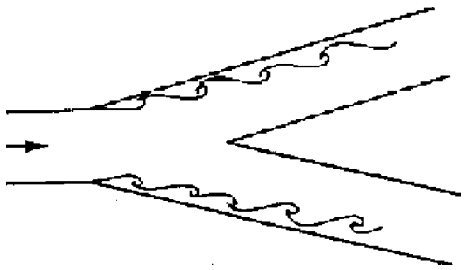


FIGURE 14.38 Boundary layer separation in branching blood vessels.

large enough to extend and grow via fluid separation. At that stage, the disease grows rapidly. However, since turbulence may exist downstream of the separation region, it is possible to diagnose the disease by listening for turbulence with a stethoscope or other diagnostic tool, such as ultrasound-based echocardiography.

14.2.6 Reynolds Number and Types of Fluid Flow

The differences between laminar and turbulent flow are considerable. Laminar flow, sometimes known as streamline flow, occurs when a fluid flows in parallel layers, with no disruption between the layers. It is the opposite of turbulent flow. In nonscientific terms, laminar flow is "smooth," while turbulent flow is "rough." The dimensionless Reynolds number is an important parameter in the equations that describe whether flow conditions lead to laminar or turbulent flow. It indicates the relative significance of the viscous effect compared to the inertia effect, with laminar flow being slower and more viscous in nature, while turbulent flow can be faster and more inertial (accelerating) in nature. The Reynolds number is proportional to the ratio of the inertial force (acceleration) to the viscous force (fluid deceleration). The values of the Reynolds number for various types of flow are as follows:

- **laminar** if $\mathrm{Re} < 2{,}000$
- **transient** if $2{,}000 < \mathrm{Re} < 3{,}000$
- **turbulent** if $3{,}000 < \mathrm{Re}$

These are approximate values. The Reynolds number can be affected by the anatomy/geometry of the fluid flow field, the roughness of the vessel wall, and irregularities in pressure or external forces acting on the fluid. When the Reynolds number is much less than 1, creeping motion or Stokes flow occurs. This is an extreme case of laminar flow where viscous (friction) effects are much greater than the virtually nonexistent inertial forces. Stokes flow was previously described regarding the falling sphere viscometer and is also typical of blood flow in capillaries.

As already noted, the Reynolds number is dimensionless and gives a measure of the ratio of inertial forces ($V\rho$) to viscous forces (μ/L), and, consequently, it quantifies the relative importance of these two types of forces for given flow conditions. The equation for the Reynolds number is

$$\mathrm{Re} = \frac{\rho \mathrm{VD}}{\mu} = \frac{\mathrm{VD}}{v} = \frac{\mathrm{QD}}{vA}$$

where

V= the mean fluid velocity in (cm/s)
D = the diameter (cm)
μ = the dynamic viscosity of the fluid (g/cm-sec)
$\nu = \mu/\rho$ = the kinematic viscosity ($\nu = \mu/\rho$) (cm^2/s)
r = the density of the fluid ($\mathrm{g/cm}^3$)
Q = the flow rate (cm^3/s)
A= the pipe cross-sectional area (cm^2)

The first form of the equation is the most common form. In the circulatory system, the Reynolds number is 3,000 (mean value) and 7,500 (peak value) for the aorta, 500 for a

typical artery, 0.001 in a capillary, and 400 for a typical vein. This corresponds to turbulent flow in the aorta (where the aortic wall is strengthened to overcome the turbulent forces), laminar flow in arteries and veins, and creeping flow in capillaries. Thus, there are added forces, mixing, and momentum transport in the aorta and virtually no momentum transport in capillaries. However, as was noted in the discussion on mass transfer in systemic capillaries, it is the low axial blood flow that allows for the radial mass transfer to occur. This could not happen in arteries, since the fluid velocity is too large to allow any significant radial mass transfer to occur. However, with turbulent flow in the aorta, there is considerable mixing of fluid layers and added wall shear stresses. As a result, there is added mass transfer across the artery walls, which is why arterial disease is prominent in the aorta. Such a disease is caused by mass transfer of lipids across the arterial wall and by distortion of the endothelial cells lining the vessel wall, allowing for such mass transfer to occur in the gaps.

A depiction of the systemic capillaries is shown in Figure 14.39. The Reynolds number is less than 1 and the flow is creeping flow. Such flow produces entirely viscous flow with no inertial effects, no turbulence, and no flow separation. As such, there are no negative effects due to the circuitous nature of the capillary bed nor to sudden sharp turns or bifurcations. In larger vessels, such vessel geometry would produce turbulence and/or flow separation.

14.2.7 Blood Pressure Measurement

The measurement of blood pressure noninvasively is conducted by the well-known blood pressure cuff assembly: the sphygmomanometer. Most individuals have had their blood pressure taken with this mechanism. It involves inflating a pressure cuff just above the elbow to a pressure significantly exceeding systolic blood pressure (120 mm Hg). At such a high pressure, the artery just below the surface (brachial artery) collapses as the externally applied pressure from the cuff exceeds the internal blood pressure inside the artery. A stethoscope is placed just below the cuff. As the pressure is reduced, the external pressure will eventually fall slightly below the systolic blood pressure, and the brachial artery will open slightly. The blood at that high pressure will jet through the small opening, producing turbulent flow. Such jetting turbulent flow can be heard via the stethoscope. This is similar to a heart murmur heard through a stethoscope, which is indicative of a leaky heart valve, also producing turbulence, as well as separated flow downstream of a stenosis.

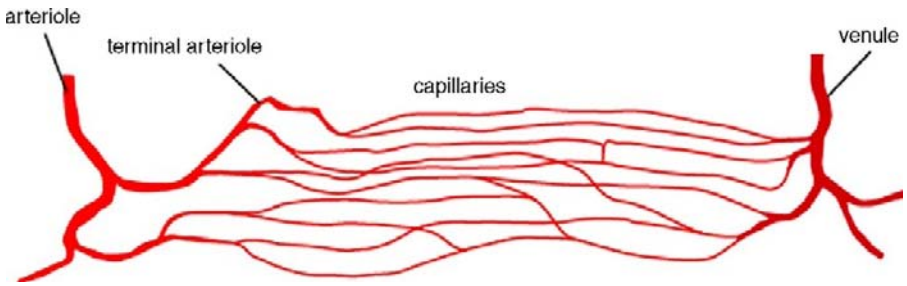


FIGURE 14.39 The network of the systemic capillaries leading from arterioles and ending in venules, with bending and bifurcating elements in the capillary bed that are possible with creeping flow.

As the cuff pressure continues to fall, the duration that the brachial artery pressure exceeds the cuff pressure will increase. As such, the (small) opening will increase in size and duration, until finally the vessel remains open as the cuff pressure falls below diastolic pressure in the artery. At that point, the turbulent sounds, called Korotkoff sounds, fade away.

A manometer is connected to the pressure cuff so the physician can see what the pressures are when the sound begins (systolic pressure) and when the sound finally fades away (diastolic pressure). Thus, the sphygmomanometer can be used as a noninvasive means of determining the two extremes of blood pressure, normally at 120/80 mm Hg. The relationships between cuff pressure, vessel pressure, and sounds via the stethoscope are shown in Figure 14.40.

The measurement of blood pressure depends greatly on the location of the measurement. As shown in Figure 14.34, the blood pressure falls as blood circulates through the systemic

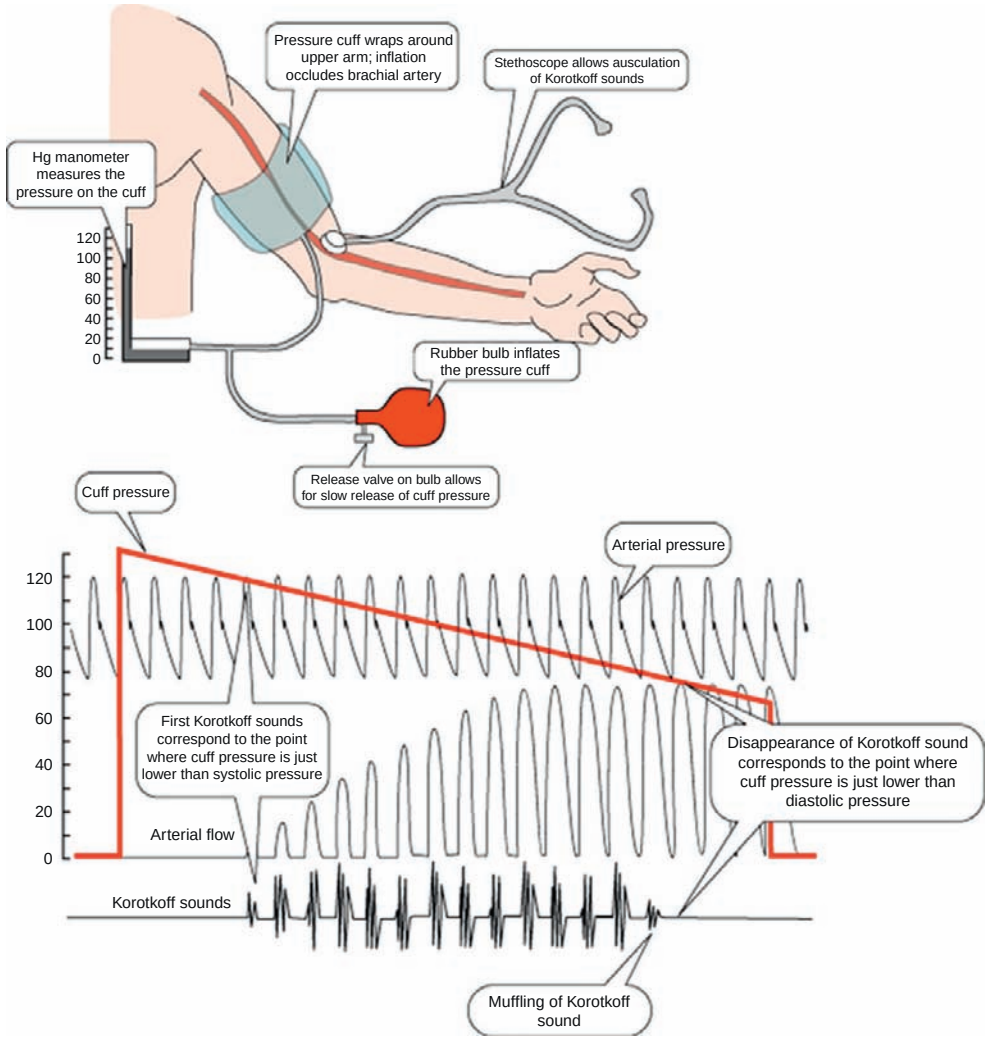


FIGURE 14.40 The sphygmomanometer and the generation of Korotkoff sounds indicating systolic and diastolic arterial pressure.

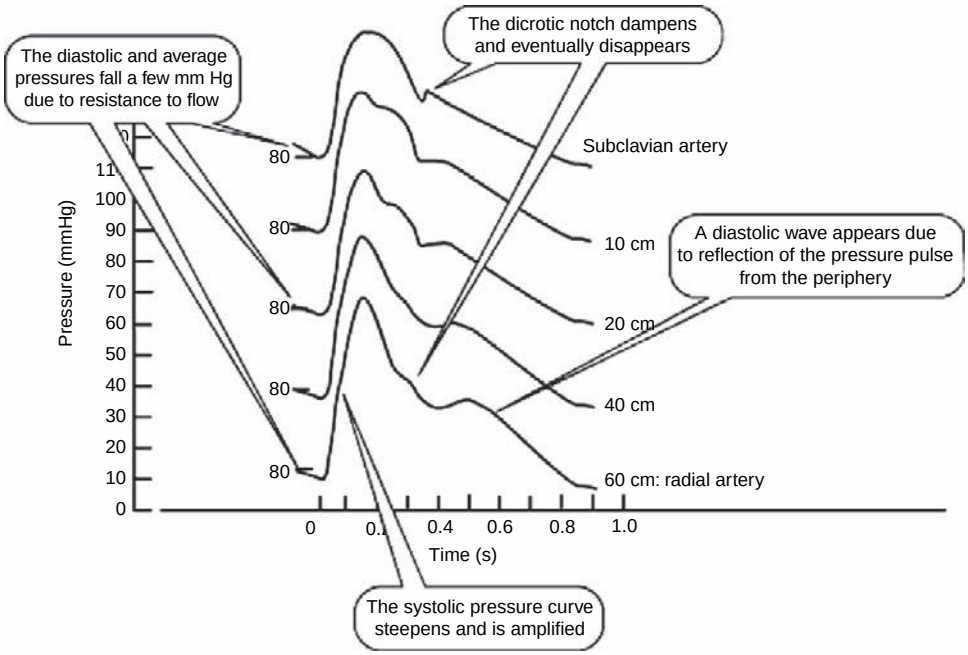


FIGURE 14.41 The blood pressure waveform and the resulting pulse pressure (difference between systolic and diastolic) can vary in differing midlevel arteries. This would result in different measured systolic and diastolic values of blood pressure via a sphygmomanometer.

circulatory system. However, even though blood pressure is normally taken in the arterial network, the location of that measurement can affect the systolic and diastolic values. As shown in Figure 14.41, the blood pressure waveform can be steeper with a corresponding larger difference between systolic and diastolic pressure, depending on the location within one of the midsized arteries. The increase in the pulse pressure (difference between systolic and diastolic) is sometimes due to a reflection of the pressure wave, which is the propagation of the pulse wave (moving far faster than the fluid time-varying pressure).

As can be seen, blood flow and velocity patterns can be quite complex and varying in direction, time, and dimensions. Simplifying assumptions such as Poiseuille flow, a Newtonian fluid, and an incompressible fluid aid in our ability to develop first-order solutions to rather complex waveforms and flow systems. Fluid mechanics is a more macroscopic description of body functions, where mass transfer was a more cellular description of body functions. Both mass transfer and fluid flow (momentum transfer) can affect heat transfer within and outside the body. This is the topic of the next section.

14.3 BIOMEDICAL HEAT TRANSPORT

Heat transport within the human body provides the mechanism for the body's core temperature to remain constant by shunting heat from the core to the periphery, where heat is lost through the skin and through respiration. Heat is generated inside the body as it metabolizes food, which produces heat as a by-product. Heat transfer from the body to the

environment is also affected by outside temperature and wind speed. The latter affects windchill, which can produce a substantial heat loss from the body. Heat transport also occurs in blood heating and cooling devices, which are often used during surgical procedures and as part of heart-lung machines. Finally, heat transport is vital for extreme conditions such as a firefighter's suit or a spacesuit, both used to protect the body from extreme temperatures in the environment.

The heat generated by the body at rest is called the *basal metabolic rate*, equal to 72 kcal/hr. Obviously, as metabolic activity increases, such as during exercise, the level of heat production rises.

The equations describing heat transport are similar in form to the equations relating transport of mass (Fick's Law) and momentum. Such transport is always in the direction of a decreasing gradient. For mass, the gradient is the concentration difference between two points or across membranes. For momentum transport of fluids, it is the pressure differential causing fluid flow. For heat transfer, the driving gradient is the temperature difference between two points.

14.3.1 Conduction, Convection, and Radiation

Heat transport can occur by three basic mechanisms. Thermal *conduction* is the process by which heat transfer occurs by molecular interaction. It can occur in gases, liquids, and solids. An example is heat transfer across a closed window or across a wall. In the human body, the example of conductive heat transfer is across tissue from the core toward the periphery (skin). Thermal *convection* is the process by which heat transfer occurs via bulk motion of a fluid. An example might be forced air flow from an air conditioner vent or windchill on a windy day. In the body, examples of convective heat transfer include blood flow from the core toward the periphery or air flow in the lungs from the alveoli through the trachea and out the mouth and nose. Both conduction and convection require that there be a material involved, although convection does not occur in solids, whereas conduction can. The third method of heat transfer is thermal *radiation*, which occurs as a result of electromagnetic transport processes. An example is heat gain from the sun to the earth (and your own skin). Radiation is a surface-to-surface phenomenon and does not require a material interaction. The greenhouse effect is an example of thermal radiation. As an example, Figure 14.42 depicts the three types of heat transfer from a campfire.

14.3.2 Thermal Conduction

Conduction (or heat conduction) is the transfer of thermal energy between neighboring molecules in a substance due to a temperature gradient. It always takes place from a region of higher temperature to a region of lower temperature and acts to equalize temperature differences. Conduction takes place in all forms of matter, including solids, liquids, and gases, but does not require any bulk motion of matter. In solids, it is due to the combination of vibrations of the molecules in a lattice and the energy transport by free electrons. In gases and liquids, conduction is due to the collisions and diffusion of the molecules during their random motion. Conduction is the movement of heat from a warmer object to a cooler one when they are in direct contact with one another. This serves to even the temperature

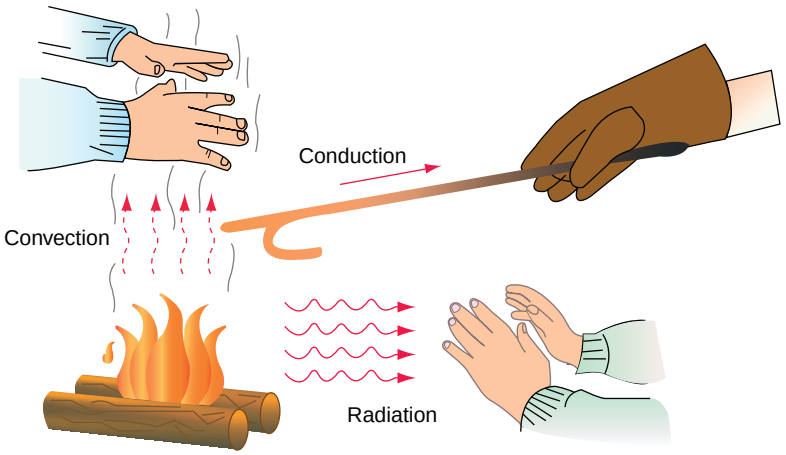


FIGURE 14.42 The three types of heat transfer: conduction, convection, and radiation.

difference between them over time. The rate of heat transfer between two objects of different temperatures depends on several factors:

- The temperature difference between the two objects
- The total surface area where the two objects are in contact
- The efficiency of the insulation that is between the objects

The greater the temperature difference between two objects in contact, the more heat that is transferred between them in a given time. For example, when you place your hand on a very hot stove top, you will quickly receive a great heat input from the stove to your hand. If the stove top is only warm, it will take much longer to receive the same amount of heat into your hand.

The more surface area in contact between two objects, the more quickly heat is transferred between them. Stick your finger on an icicle for a minute and it feels cold, but you will probably not feel too uncomfortable. Strip naked and lay on a block of ice for a minute and you will most likely be very uncomfortable indeed as the ice absorbs heat from your body at a very fast rate.

The amount of heat being transferred between two objects of different temperatures can be slowed by the use of effective insulation. Insulation retards the movement of heat between them by creating pockets of dead air space that trap the flow of heat or by otherwise slowing the overall heat transfer rate by adding a low-conductance/high-resistance layer.

The law of heat conduction, also known as Fourier's law, states that the time rate of heat transfer through a material is proportional to the negative gradient in the temperature and to the area at right angles to that gradient through which the heat is flowing. We can state this law in two equivalent forms: the integral form, in which we look at the amount of energy flowing into or out of a body as a whole, and the differential form, in which we look at the flows or fluxes of energy locally. The differential form is

$$\vec{q} = -k\nabla T$$

where $\vec{q}$ is the local heat flux, $[W \cdot m^{-2}]$, k is the material's conductivity, $[W \cdot m - 1 \cdot K^{-1}]$, and ∇T is the temperature gradient, $[K \cdot m^{-1}]$. "K" refers to °Kelvin, an absolute scale for temperature. For materials with a high thermal conductivity, such as metals, there is little if any insulating capacity. For materials with a low thermal conductivity, the material has more insulating characteristics. Air and many gases have low thermal conductivities. Human tissue also has a relatively low thermal conductivity and is thus relatively insulating, which aids in reducing heat transfer to the core of your body on extremely hot or cold days. Clothing helps to further insulate the body by providing extra layers for heat conduction and by trapping air, which is itself a thermal insulator.

For many simple applications, a one-dimensional form of Fourier's law is written as

$$q_x = -k\frac{dT}{dx}$$

The thermal conductivity, k, is often treated as a constant, though this is not always true. While the thermal conductivity of a material generally varies with temperature, the variation can be small over a significant range of temperatures for some common materials. In anisotropic materials, the thermal conductivity typically varies with orientation. For the human body, trapped air in the lungs has a different conductivity as opposed to human tissue, but both have smaller values than denser materials, such as human bone. Water (and in many ways human blood) has a larger conductivity value, but it is also a thermal capacitor, able to retain heat.

Since the surface area can also affect heat conduction occurring orthogonal to a plane, Fourier's law is also written as

$$\frac{\Delta Q}{\Delta t} = -kA\frac{\Delta T}{\Delta x}$$

where A is the cross-sectional surface area, ΔT is the temperature difference between the ends of a thickness, and Δx is the distance between the ends.

The thermal conductivities for various materials, including human tissue, are as follows:

Material	k (W/mK)
Human tissue (organs/muscle)	0.5
Human tissue (fat)	0.2
Human tissue (skin)	0.3
Blood	0.5
Bone	0.5–0.6
Water	0.62
Air	0.03
Glass	1.1
Aluminum	200
Copper	400

As can be seen from the preceding list, metals have high thermal conductivities, which is to be expected. Most human tissue has relatively low thermal conductivities, suggesting that conduction is not rapid within the human body. The skin has a lower conductivity than

human tissue, suggesting that conduction into the body (as on a hot day) is also limited. It is interesting to note that air, like most gases, has a very low thermal conductivity. Thus, trapped air within clothing provides excellent resistance to thermal conduction in either direction—into or out of the body. On an extremely cold day, a parka with trapped air within a nylon covering provides excellent thermal resistance. A three-layer model can be utilized to examine heat transfer via thermal conduction, as shown in Figure 14.43.

By applying Fourier's law to each layer and assuming that there is no accumulation of heat within any layer, the equations become

$$Q1 = K1(T1\text{-}T2)/\Delta X1 \qquad Q2 = K2(T2\text{-}T3)/\Delta X2 \qquad Q3 = K3(T3\text{-}T4)/\Delta X3$$

Since there is no accumulation of heat, then $Q1 = Q2 = Q3 = Q$. Rearranging the preceding equations gives

$$T1\text{-}T2 = Q\Delta X1/K1 \qquad T2\text{-}T3 = Q\,\Delta X2/K2 \qquad T3\text{-}T4 = Q\,\Delta X3/K3$$

Summing these equations results in

$$T1\text{-}T4 = Q/(\Delta X1/K1 + \Delta X2/K2 + \Delta X3/K3) \text{ or}$$
$$Q = (T1\text{-}T4)(\Delta X1/K1 + \Delta X2/K2 + \Delta X3/K3)$$

Since the thermal conductivities are in the denominator, any one term can affect the overall heat transfer rate. Thus, if air is one of the layers, its thermal conductivity is so low that with 1/K as a factor, it would dominate the other layers. As a result, it is usually the layer with the lowest thermal conductivity that affects the overall heat transfer rate via conduction. This is why trapped air is such a good thermal insulator. Note that T1-T4 will eventually equal out if there is no continual heat source, but the rate of transfer will be slower with air as one of the layers.

If the body maintains a certain level of heat production, then a coat with trapped air will help to maintain skin and body temperature without a danger of hypothermia. Figure 14.44 shows the temperature distribution inside and at the body surface.

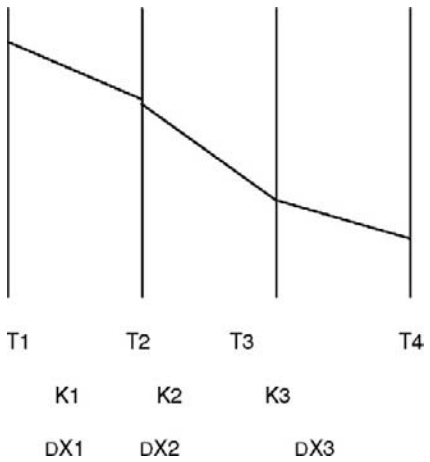


FIGURE 14.43 A three-layer model for steady-state heat conduction.

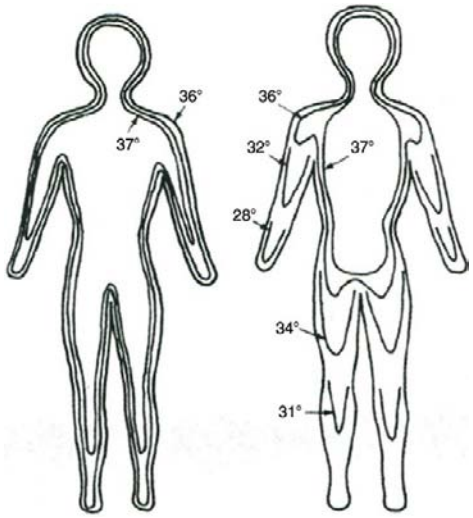


FIGURE 14.44 Temperature distribution in the human body, with the left figure depicting the temperature distribution in a warm environment and the right figure for a cold environment. *From Ruch and Patton (1965).*

Although the temperature nearest the skin is lower in a cold environment, the use of clothing with trapped air prevents the skin temperature from becoming dangerously cold. For someone in an extremely cold environment, such as outer space, the use of an encapsulated suit is needed because the exterior temperature is extremely low. In an earth-bound environment, the issues associated with insulating clothing do not address the added issue of heat loss (or gain) via respiration. We will soon see how respiration can be an important element in heat loss from the body.

14.3.3 Thermal Convection

Convection is heat transfer via bulk motion of a fluid, which can be a liquid or a gas. Examples of convective heat transfer include air conditioning coming from a compressor, blower, and ductwork; windchill; heat exchangers; and blood warming/cooling as part of heart/lung machines (also called cardiopulmonary bypass).

Convective heat loss can be forced or via natural (unforced) mechanisms. Forced convection is commonly due to a pressure gradient or other forcing function that would move a fluid (with its associated heat) to another position or location. Thus, a forced air system consisting of an air conditioner, blower, and vents is an example of a forced convection system. Another is windchill, which results from a pressure gradient moving air across a surface and removing heat from that surface. If an individual experiences windchill, he or she would feel colder because the windchill would remove heat from the skin.

Natural or unforced convection is due to thermal gradients in a fluid that result in fluid motion due to buoyancy differences. Thus, hot air rises because it is lighter than cold air.

The movement of the hot air by differences in fluid density is a prime example of unforced convection. Normally, free convection is much smaller in magnitude than forced convection, since the movement of fluid via density differences is often far smaller than the movement of fluid via pressure gradients. Hot air rises but at a rate far smaller than forced hot air from a heat pump through vents or via windchill from a strong wind.

Convective heat loss can generally be approximated by a convective heat loss coefficient used in the generalized equation

$$Q_c = h_c \, A \, \Delta T$$

where

Q_c = convection heat transfer rate
h_c = convective heat transfer coefficient
A = area over which heat transfer occurs
ΔT = temperature gradient between the surface and the environment

Typically, the convective heat transfer coefficient is a function of fluid velocity for forced convective heat transfer. There are several versions of the equation relating the convective heat transfer coefficient to fluid velocity. One such example is

$$h_c = 5.6 \, v^{0.67}$$

where v is in m/sec and h_c is in kcal/m^2 – hr – ^oC.

As the fluid velocity increases, so does the convection heat transfer, which increases the convective heat transfer rate. Thus, as the wind velocity increases, so does the windchill, which is an example of convection cooling. The convection heat transfer rate for free (natural) convection is normally a constant at about 2–2.3. If one equates the free convective coefficient with the forced coefficient equation, then the air velocity required for the forced term to exceed the free term is

$$2.3 = 5.6 \, v^{0.67} \text{with } v^{0.67} = 2.3/5.6 = 0.4107$$

and therefore

$$v = 0.263 \text{ m/sec or } v = 0.588 \text{ mph}$$

Thus, any wind speed over about a half mile per hour would exceed the free convective heat loss.

However, the more accurate method of computing the forced convective heat loss is to use the Reynolds number (indicative of laminar vs. turbulent flow), the Prandtl number (relating viscous effects to thermal effects/conduction), and the Nusselt number (convective heat transfer to thermal effects/conduction). This allows for the specific calculation of the convection heat transfer rate for laminar or turbulent flow. Because these two types of fluid flow are very different in terms of velocities and velocity profile, it is not surprising that the heat transfers associated with these two types of fluid flow are quite different.

In general, the Nusselt number (Nu) = h_c D/k, where D is the tube diameter or representative length, k is the thermal conductivity, and h_c is the convective heat transfer coefficient.

However, the Nusselt number is related to the Reynolds number (Re) and the Prandtl number (Pr) by the following representative equations:

$$\begin{aligned} Nu &= 0.322\ Re^{0.5}\ Pr^{0.33} \quad \text{for laminar flow} \\ Nu &= 0.023\ Re^{0.8}\ Pr^{0.33} \quad \text{for turbulent flow} \end{aligned}$$

where

$$\begin{aligned} Re &= \rho\ V\ D/\upsilon \\ Pr &= c_p\ \upsilon/k \\ \rho &= \text{density} \\ \upsilon &= \text{viscosity} \\ c_p &= \text{specific heat} \end{aligned}$$

Since the Nusselt number is related to both the Reynolds number and the Prandtl number, as well as to the convective (forced) heat transfer coefficient, by computing these three parameters, one can determine the convective heat transfer for the appropriate type of flow. Obviously, this is a more complex set of calculations than the simple form of $h_c = 5.6\ v^{0.67}$, but it is more accurate, since the simplistic form does not designate the type of fluid flow.

14.3.4 Heat Exchangers

Heat exchangers are commonly used in medical settings as blood heaters/coolers for open heart-lung machines or for patient heating/cooling for standard surgeries. The purpose of cooling the patient during surgery is to reduce the metabolic load, which lowers the need for larger blood flow rates or ventilation rates. The patient is then heated (or the blood heated) toward the end of surgery. A typical type of heat exchanger for blood is the double pipe heat exchanger, which consists of two concentric pipes with blood in one pipe and water in the other. By using cold water, the blood is cooled, while using warm water heats the blood. A double pipe heat exchanger combines convective heat transfer, as the fluids are moving, along with heat conduction through the pipe walls. A typical double pipe heat exchanger is shown in Figure 14.45. Because the blood and the water can either be moving in the same direction (cocurrent flow) or in opposite directions (countercurrent flow), there are two versions of the heat exchanger. The physical device is shown in Figure 14.46.

The design equations for a double pipe heat exchanger combine Fourier's law for heat conduction through the wall of the inner pipe, together with the complex form of the convective heat transfer coefficients with the Reynolds, Prandtl, and Nusselt numbers.

This equation is $Q = U\ A\ \Delta T_{lm}$, where U is the overall heat transfer coefficient and is a function of the thermal conductivity and the convective heat transfer coefficient, A is the surface area of the pipe, and ΔT_{lm} is the effective temperature difference between the two fluids from one end of the exchanger to the other. This term is known as the log mean of the temperature gradient. Since the surface area of the pipe could be the inner or outer surface, then the overall heat transfer coefficient is synched with the surface area as U_o and A_o, or U_i and A_i. The temperature gradient is unaffected by inner or outer diameter but is affected by whether the flow is cocurrent or countercurrent. In general, countercurrent flow is more efficient for overall heat transfer.

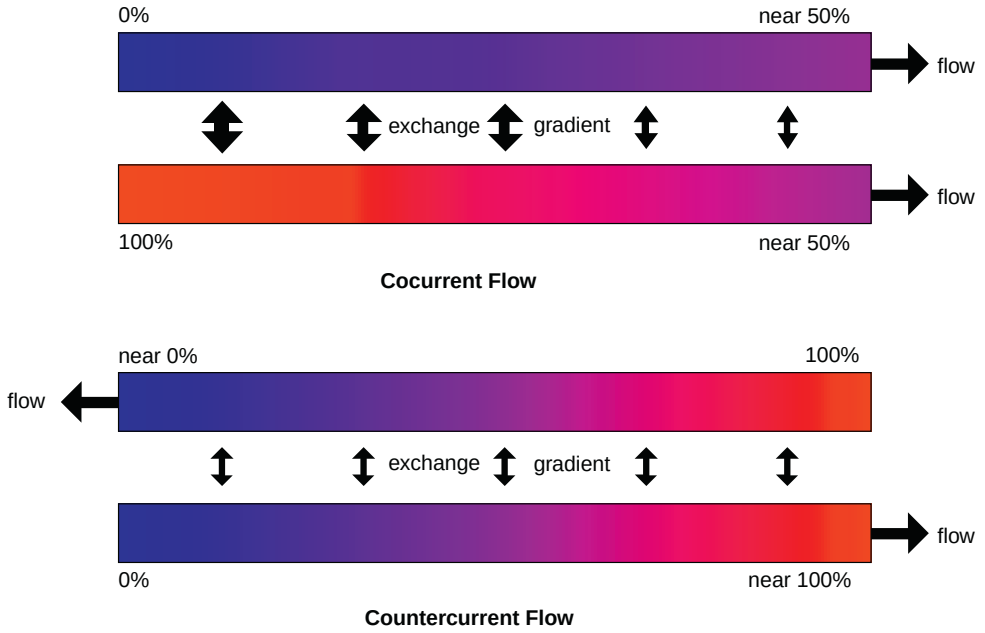


FIGURE 14.45 Double pipe heat exchangers for cocurrent and countercurrent flows.

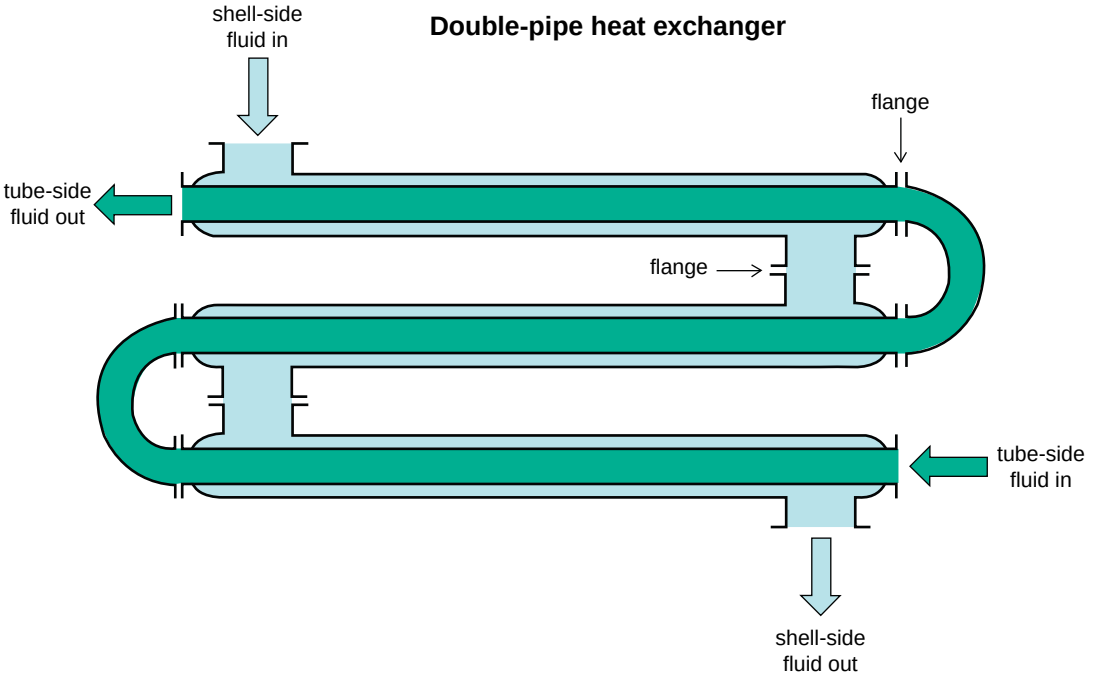


FIGURE 14.46 Double pipe heat exchanger design.

14.3.5 Thermal Radiation

Thermal radiation is electromagnetic radiation emitted from a material that is due to the heat of the material, the characteristics of which depend on its temperature. An example of thermal radiation is the infrared radiation emitted by a common household radiator or electric heater. A person near a raging bonfire will feel the radiated heat of the fire, even if the surrounding air is very cold. Thermal radiation is generated when heat from the movement of charges in the material (electrons and protons in common forms of matter) is converted to electromagnetic radiation. Sunshine, or solar radiation, is thermal radiation from the extremely hot gasses of the sun, and this radiation heats the earth. The earth also emits thermal radiation but at a much lower intensity because it is cooler. The balance between heating by incoming solar thermal radiation and cooling by the earth's outgoing thermal radiation is the primary process that determines the earth's overall temperature. As such, radiation is the only form of heat transfer that does not require a material to transmit the heat. Radiative heat is transferred from surface to surface, with little heat absorbed between surfaces. However, the surfaces, once heated, can release the heat via conduction or convection to the surroundings.

Thermal radiation is conducted via electromagnetic waves. As such, this form of heat transfer is not only a function of the temperature difference between the two surfaces but also the frequency range of the emitted and received energy. As an example, sunlight is composed of the visible light spectrum as well as infrared energy and ultraviolet energy. Figure 14.47 depicts the effects of temperature and wavelength of the thermal energy on the heat transfer rate.

When radiant energy reaches a surface, the energy can be absorbed, transmitted (through), or reflected (or any combination). The sum of these three effects equals the total energy transmitted, and the parameters that describe these three phenomena are given by

$$\alpha + \rho + \tau = 1,$$

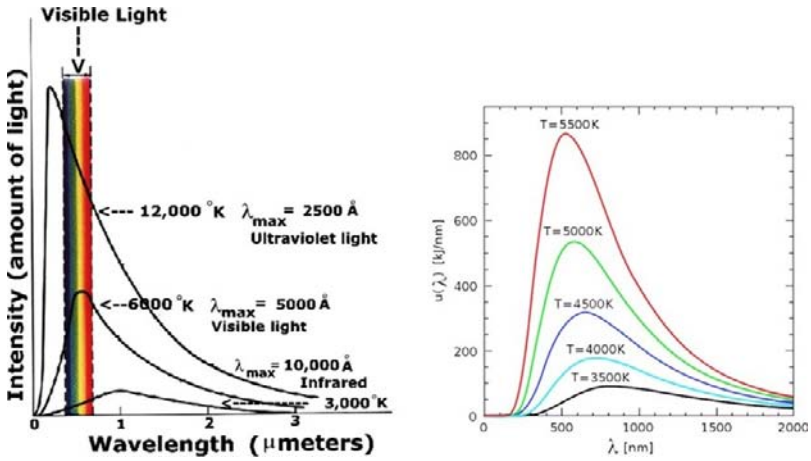


FIGURE 14.47 Peak wavelength and total radiated amount vary with temperature. Although the plots show relatively high temperatures, the same relationships hold true for any temperature down to absolute zero. Visible light is between 380 and 750 nm.

where α represents spectral absorption factor, ρ represents the spectral reflection factor, and τ represents the spectral transmission factor.

The radiative heat transfer rate is given by the Stefan-Boltzmann law

$$Q_r = \sigma A\, T^4$$

where σ is the Boltzmann constant and A is the surface area of the radiating source. The temperature is in an absolute scale (°Kelvin, corresponding to °C, or °Rankin, corresponding to °F).

To predict the exact amount of radiative heat transfer between two surfaces, the preceding equation is expanded as

$$Q_r = \sigma F_1 A_1 (T_1^4 - T_2^4)$$

where F is the facing factor that represents the amount of the emitting surface (1) facing the receiving surface (2) with the surface area A representing surface 1. Correspondingly, this equation can use F_2 and A_2 to represent the facing factor for the receiving surface toward the emitting surface with the surface area of 2. Boltzmann's constant and the temperature gradient are unchanged for either form of the equation. The facing factor can be approximated as a disk of radius R if the distance between the two surfaces is large, such as the earth to the sun. For shorter distances, the facing factor is a complex interaction between the angles of the two surfaces that face each other.

As can be seen, thermal radiation is affected by the frequency of the emitted energy. This is why sunscreen ointments have ultraviolet protection, since this type of energy can be damaging to skin. In addition, it is common to feel warmer on the sunny side of the street as opposed to the shady side, given the radiative heat transfer. Radiation can be a significant source of heat as compared to the other forms (conduction and convection) because radiation is composed primarily of sunlight and the respective heating of the earth.

14.3.6 The Greenhouse Effect

Emitted radiation from the sun consists of a spectrum of electromagnetic radiation of varying frequencies and wavelengths. The sun's surface temperature is approximately 10,000°C. When this energy hits the earth's surface, it can be absorbed, reflected, or transmitted through. The amount of each of these factors depends on the surface temperature of the earth and the exact material (water, soil, man-made materials). However, the reflectivity, absorptivity, or transmissivity is frequency dependent. In addition, the frequency spectrum of the incoming thermal energy is shifted due to the different surface temperature of the earth as compared to the sun.

As the energy reached the earth through its atmosphere, the transmissivity was high (allowing it to pass through the atmosphere) and the reflectivity and absorptivity were low. However, as the energy hit the earth's surface, the new temperature shifted the spectrum such that the new reflectivity was high and the transmissivity was low. Thus, when some of the thermal energy was reflected back from the earth's surface toward the atmosphere, it did not transmit through toward outer space but instead was reflected back toward the earth's surface. At each reflection back, more energy was absorbed into the earth's surface, heating the earth. Normal conduction and convection would heat the

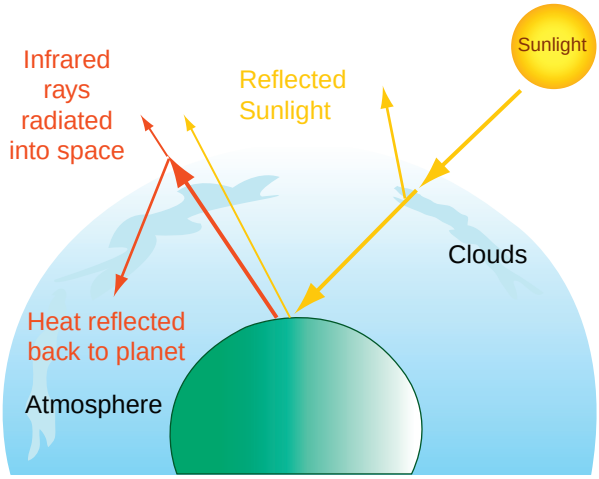


FIGURE 14.48 Depiction of the greenhouse effect resulting in thermal radiation being reflected back toward the earth's surface.

surrounding air and buildings, resulting in the warm temperatures that allow us to live on this planet. This phenomenon is known as the *greenhouse effect* (Figure 14.48).

You have probably experienced the greenhouse effect firsthand if you ever left your car unattended on a sunny day. The thermal radiation is transmitted through the windshield and hits the seats and dashboard. The frequency shift occurs, and the reflected energy is then reflected back to the windshield, only now at a high reflectivity and low transmissivity. As such, the inside of the car heats up, often up to 20–30 degrees higher than the ambient air. A sunscreen on the window will reduce the amount of radiation entering the car and keep it cooler.

Commercial greenhouses purposefully trap the heat via the greenhouse effect and use some limited screening to control the amount of trapped heat. The increased heat allows the plants to remain at a relatively warm temperature and avoid the changes in temperature with changing air currents and air temperature. By keeping the doors to the greenhouse closed, the temperature can be somewhat maintained overnight. A commercial greenhouse is shown in Figure 14.49.

Heat loss at night via radiation occurs as reemitted energy from the earth's surface to the dark night sky. The amount of radiation energy released at night depends on whether the sky is clear or cloudy.

14.3.7 Heat Loss via Respiration

When air is inspired into the lungs, heat and water vapor are transferred to the air by convection and evaporation from the surface lining the respiratory tract. By the time the air has reached the deepest parts of the lungs, the air is at deep body temperature (37°C) and saturated with water vapor (47 mm Hg partial pressure). As the air moves outward through the respiratory tract during expiration, some heat is transferred back to the body

FIGURE 14.49 Commercial greenhouses.

and some water is condensed. However, the inspired air still contains significantly more heat and water than the inspired air.

Respiration results in a latent heat loss and a sensible heat loss. The latent heat loss is based on the latent heat of vaporization of water by the equation

$$Q_{el} = (dm_a/dt)(Y_o - Y_i)\lambda$$

where

dm_a/dt = the kilograms of air breathed in and out per hour
$Y_o - Y_i$ = the difference in the expired and inspired air water content
λ = the latent heat of vaporization of water at the expired air temperature

The pulmonary ventilation rate (dm_a/dt) is primarily a function of the body metabolic rate via the following relationship:

$$(dm_a/dt) = 0.006\ M$$

where M is the metabolic rate in Kcal per hour.

It is somewhat rare to breath in bone-dry air, so the inspired air temperature and water content can affect the expired air from the lungs. Fanger and colleagues (1968) found that there is a relationship between the inspired water content of air and the expired water content

$$Y_o = 0.029 + 0.20\ Y_i$$

for normal conditions. Thus, if the entering air is very dry ($Y_i = 0$), the expired air would be less humid than if the inspired air were humid. McCutchan and Taylor (1951) determined that the temperature of the expired air is dependent on the inspired air in the following fashion:

$$T_o = 32.6 + 0.006\ T_i + 32\ Y_i$$

Although the relationship between T_o and T_i is expected, the inspired air humidity also plays a role. The air, upon being expired, is somewhat cooler than it would have been if the inspired air had been humid.

Besides the evaporative latent heat loss associated with respiration, there is also a sensible heat loss (when $T_o > T_i$) or heat gain (when $T_o < T_i$) that is described by

$$Q_{sl} = (dm_a/dt)\, C_p\, (T_o - T_i)$$

where C_p is the specific heat. This is the heat loss due to heating up the dry air component of the inspired air. An example of calculating heat loss follows.

Assume that 6 liters/min of bone-dry air at 20°C is inspired (12 breaths per minute at 500 ml tidal volume per breath = 6 liters per minute). Also assume that the expired air is fully saturated with water vapor (47 mm Hg) and is at 37°C. The physical properties are

$$C_{p,\ air} \text{ at } 20^\circ\text{C} = 0.25 \text{ cal/g}^\circ\text{C}$$
$$\lambda \text{ at } 37^\circ\text{C} = 577 \text{ cal/g}$$
$$\text{Vapor pressure of water at } 37^\circ\text{C} = 47 \text{ mm Hg}$$

The dry air mass flow rate in grams per minute can be calculated via the ideal gas law:

$$\begin{aligned}(dm_s/dt) &= (6 \text{ liters/min})(1 \text{ mol}/22.4 \text{ liters})(273^\circ\text{K}/293^\circ\text{K})(28.9 \text{ g/mol}) \\ &= 7.2 \text{ g/mol of dry air}\end{aligned}$$

The amount of water vapor in the expired air is

$$(dm_w/dt) = (7.2/28.9)(47/760 - 47)(18) = 0.295 \text{ g/min}$$

Note that the true ratio of the partial pressure of water vapor in the lungs is not 47/760 but rather 47/760 – 47, since the water vapor is attached to the dry gas, which is the remainder of the partial pressure (713). The molecular weight of water is 18. It is necessary to compute the moles of water vapor when multiplying by the partial pressure ratio, since the partial pressures are related to the moles rather than the mass. Multiplying by the molecular weight returns the calculation back to units of mass.

The latent heat loss is

$$Q_{el} = 0.295\ (577) = 170 \text{ cal/min}$$

The sensible heat loss is

$$Q_{sl} = 7.2\ (0.25)(37 - 20) = 30.4 \text{ cal/min}$$

Note that the latent heat loss is far more significant than the sensible heat loss. The total heat loss of approximately 200 cal/min equals 12.0 kcal/hr, which is 17 percent of the basal metabolic rate of the body—an appreciable quantity. During exertion, one's respiration increases, which provides even greater heat loss by this mode. The overall rise in the metabolic heat production rate is also greater, such that an additional heat loss mode is necessary, such as sweating. Evaporation of sweat from the body produces the same latent heat loss as was evident via respiration. However, if one sweats profusely, such that sweat rolls off the body, then dehydration may occur, which can be quite dangerous.

An interesting note is that the panting of dogs is the primary mechanism of heat rejection, since dogs cannot sweat. By interchanging air with the dead space on a rapid basis, rather than via deep breathing, the dogs avoid hyperventilation yet produce heating of inspired air as well as humidification. Thus, there is latent and sensible heat loss. Many dogs have thick coats that provide insulation against heat gain or loss via conduction. Thus,

if the dog were to get overheated, the solution is to hose down the dog, which would reduce the trapped air insulation via the fur. The same effect occurs when skiing in a cold environment. A ski parka has trapped air that serves as an insulator. However, if the parka were to become wet, the trapped air would be mixed with water, thus reducing its effectiveness as an insulator.

14.3.8 Heat Loss inside the Body

Many of the mechanisms for heat transfer that have been discussed refer to heat loss from the body to the environment. However, heat must be transferred from the body core to the skin or lungs. The two mechanisms available for such transfer are via conduction and via convection. Conduction is a slow process. Convection is accomplished via blood flow. As blood reaches the body core, it absorbs heat. Blood then is channeled toward the periphery, where heat is released. Ruch and Patton (1965) discuss the mechanisms by which the blood circulation affects internal heat distribution in three ways:

1. It minimizes temperature differences within the body. Tissues having high metabolic rates (e.g., the liver) are more highly perfused and are thus kept at nearly the same temperature as less active tissues. Cooler tissues are warmed by blood coming from the active organs.
2. It controls the effective body insulation in the skin region. Warm blood is increased to the skin via vasodilation, when the body wishes to reject heat. Blood is bypassed from arteries to veins via deeper channels through vasoconstriction when conservation of body heat is vital.
3. Countercurrent heat exchange between major arteries and veins often occurs to a significant extent. If heat conservation is necessary, arterial blood flowing along the body's extremities is precooled by loss of heat to adjacent venous streams.

A model of heat transfer from the core toward the skin can be split into two sections: from the core to a muscle region and from the muscle region to the skin (Cooney, 1976). The heat transfer includes both a conductive term using Fourier's law and a convective term using a forced convective form as follows:

$$Q = k_{cm}\, A\, \frac{(T_c - T_m)}{\Delta Z_{cm}} + (dm_b/dt)\, C_{pB}\, (T_c - T_m)$$

conduction convection

$$Q = k_{ms}\, A\, \frac{(T_m - T_s)}{\Delta Z_{ms}} + (dm_b/dt)\, C_{pB}\, (T_m - T_s)$$

Since the heat transfer rate is assumed to be the same for both zones and the convection term is likely far greater than the conduction term, then an overall heat transfer rate from core to skin is of the form

$$Q = (dm_b/dt)\, C_{pB}\, (T_c - T_s)$$

This form does not account for any shunting of blood to allow for heat gain or loss via vasoconstriction or vasodilation.

14.3.9 Heat Loss in Extreme Environments

One example of heat loss (or gain) in an extreme environment is in a fire, which firefighters must deal with. As such, fire protection clothing and gear must protect against excessive heat conduction and therefore must have considerable insulating effect. It must also protect against radiation, since an open flame produces significant radiant energy, as was shown in Figure 14.42. As water is used at high flow rates, the firefighter's suit must also be water resistant. A typical firefighter's clothing, called turnout gear, is shown in Figure 14.50.

FIGURE 14.50 Typical firefighter's turnout gear. (1) Nomex hood. (2) Cotton T-shirt with department logo, small on chest and large on back. (3) Suspenders with retroflective striping, connecting to the pants at eight points. (4) Insulated pants with retroflective striping. They are held closed with Velcro and spring hooks. They are reinforced with leather at the knees and bottoms and have two large side pockets and one smaller rear pocket.

Continued

FIGURE 14.50, cont'd (5) Steel-toed insulated rubber boots with Vibram soles. These boots have handles at their tops to help pull them on and come up to just below the knees. (6) Helmet, with goggles and department logo. These helmets have a Velcro/buckle chinstrap, adjustable headband, and a protective cloth flap that hangs over the collar, further protecting the neck and preventing embers from falling down the collar. (7) Goggles, attached to helmet. Used for eye protection when not wearing breathing apparatus. (8) Radio with clip-on microphone. These radios are waterproof and inherently safe (nonsparking) for use in explosive atmospheres if necessary. (9) D-ring carabiner, used to clip additional equipment to the coat (not standard issue). (10) Flashlight. Department jackets have custom loops to hold the flashlights at center chest. (11) Insulated leather gloves. (12) Insulated jacket with retroflective striping. Oversized pockets hold radio, gloves, a hose strap, and other equipment. Like the pants, it is held closed with Velcro and spring hooks. (13) Air-line and pressure gauge. On this particular brand of SCBA, there are two air gauges: one at the bottom of the tank in back (for checking the pressure when the tank isn't being worn) and one in front. (14) SCBA harness, comprised of shoulder and waist straps. (15) PASS device. Current issue is an integrated PASS/SCBA, which activates automatically when the air supply is turned on. (16) Name label on back of helmet. (17) SCBA shoulder straps. (18) Department identification. (19) Air tank bottle and backpack frame. The bottle is quick-swappable because at a working fire a firefighter often goes through several bottles. (20) Regulator and main supply valve. (21) Name label (under tank). *Courtesy of the Santa Clara, California, Fire Department.*

The coat and pants are insulated and reinforced. They are made out of a fire-resistant fabric called PBI (other materials like Nomex are also used). They have reflective stripes to make them reflect when a light is pointed at them so they can be better seen in the dark, as well as glow-in-the-dark patches. They also have the firefighter's name and department printed on the back, like a football player, to help identify the individual, since when everyone is suited up and wearing masks, it is difficult to tell who is who. They are also equipped with several large pockets for holding gloves, tools, radios, and so forth. Rubber or leather waterproof steel-toed boots protect the firefighter's feet. The rubber boots are usually stored in the "turned-out" pants so they can be quickly donned—hence the term *turnouts*. A fire-retardant hood covers the firefighter's head and neck, protecting the ears and other parts that would be exposed under a helmet. When properly worn, no part of the firefighter's skin is exposed or unprotected.

Helmets are color-coded so the wearer can be quickly identified at a fire scene. For many fire departments, the color codes shown in Figure 14.51 are used.

Heat transfer can affect the human body in terms of sustaining life, protecting human life in hazardous settings, and maintaining comfort. Without thermal convection via blood flow, the body core temperature would rise to dangerous levels, since conduction is too

Helmet Color Coding

Yellow:	Firefighter/Paramedic
Black:	Volunteer Firefighter
Red:	Captain
White:	Chief (Battalion or District)
Blue:	Contract EMT/Paramedic (AMR)

FIGURE 14.51 Firefighter protective helmet with color coding to indicate function. *From Santa Clara, California, Fire Department.*

slow a process to remove all heat from the body core. Without the greenhouse effect, the earth's temperature would fall, since most of the thermal radiation from the sun would be reflected back into space. Even heat exchangers for blood cooling and reheating make extended surgeries possible.

14.4 EXERCISES

1. What would be the new values in Table 14.2 if the outside water partial pressure were 25 mm Hg?
2. During cell depolarization, the relative diffusivity of sodium increases 1,000-fold. What would be the new electrical potential across the cell membrane in mv?
3. How long would it take for changes in the arteriolar diameter that affect transcapillary fluid exchange to affect the venous return? The cardiac output is 5,000 ml/min. Assume an average diameter of a blood vessel and the average time course of blood through the cardiovascular system.
4. If the glomerular filtration rate (GFR) is 125 ml/min (125 cc/min), then how many pores are there in all glomeruli?
5. Based on the relative densities, compute the ratio of pressure differences in the manometer (shown in Figure 14.25) as a function of the same height difference for the following fluids (use water as a reference):
   Mercury/water ratio:
   Glycerin/water ratio:
   Oil to water ratio:
6. Research and describe the nonlinear behavior (stress vs. strain rate) for several non-Newtonian fluids such as a dilatant fluid, a pseudoplastic fluid, and a Bingham plastic fluid. Use the form of the stress versus strain rate equation for a Newtonian fluid as a reference point.
7. In Figure 14.33, there is a notch in the aortic pressure waveform where the aortic valve closes. Explain why that might happen in terms of valve movement.
8. Using the Reynolds number values listed in Section 14.2.6 for various blood vessels, compute the ratios of $V \times D$ for these vessels, assuming that the blood density and viscosity do not change. Then, researching the average vessel diameters, compute their average velocities.
9. Using the combination equation relating Q to (T1-T4) in the three-layer model for heat conduction, compute the overall term consisting of the thermal conductivities for a thermal window with glass on the outer sides and air in the middle. Use the K values in the list of thermal conductivities and assume that all thicknesses (ΔX) are the same.
10. Over a ten-day period, note the nighttime low temperature on each day, as well as whether the sky was clear at night, partly cloudy, or cloudy. How does that relate to the greenhouse effect?
11. Using a breathing rate of 18 breaths per minute instead of 12 and an external temperature of 30°C (303°K) instead of the 20° used in the example in the text, compute the new latent and sensible heat losses from the body. Assume all other parameters are unchanged.
12. Research Nomex and PBI and determine their thermal conductivities. Is one a better insulator than the other for use in firefighter protective clothing? Gortex is also used as a material. Compare that as well.

Suggested Readings

D.O. Cooney, Biomedical Engineering Principles: An Introduction to Fluid, Heat, and Mass Transport Processes, Marcel Dekker Publisher, New York, 1976.

P.O. Fanger, P.E. McNall, R.G. Nevins, Predicted and measured heat losses and thermal comfort conditions for human beings, Symposium on Thermal Problems in Biotechnology, ASME, New York, 1968.

J.W. McCutchan, C.L. Taylor, Respiratory heat exchange with varying temperatures and humidity of inspired air, J. Appl. Physiol. 4 (1951) 121.

T.C. Ruch, H.D. Patton, Physiology and Biophysics, ninteenth ed., Saunders Publishers, Philadelphia, 1965.

Santa Clara Fire Department, California.

Wikipedia.

This page intentionally left blank

CHAPTER

15

Radiation Imaging

Joseph D. Bronzino, PhD, PE

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Understand the fundamental principles of radioactivity.
- Understand that ionizing radiation, as generally employed for medical imaging, is either externally produced and detected after it passes through the patient or introduced into the body, making the patient the source of radiation emissions.
- Understand the operation of basic nuclear instrumentation and imaging devices such as scintillation counters and gamma cameras.
- Become familiar with the basic concepts of computerized tomography systems.

15.1 INTRODUCTION

Throughout history, a persistent goal of medicine has been the development of procedures for determining the basic cause of a patient's distress. As a result, the search for tools capable of "looking into" the human organism with minimal harm to the patient has always been considered important. However, not until the later part of the twentieth century were such devices made available. Today, modern imaging devices, based on fundamental concepts

in physical science (e.g., x-ray and nuclear physics, optics, acoustics, etc.) and incorporating the latest innovations in computer technology and data processing techniques, have not only proved extremely useful in patient care, but they have revolutionized health care.

Over a century ago (1895), the first Nobel laureate, physicist Wilhelm Conrad Roentgen, described a new type of radiation, *x-rays*, that ultimately led to the birth of the new medical specialty of *radiology*, along with the medical imaging industry. Initially, these radiographic imaging systems were very rudimentary, primarily providing images of broken bones or contrast-enhanced structures such as the urinary or gastrointestinal systems (Figure 15.1). However, since the 1970s, advances in imaging techniques have revolutionized the application of radiographic imaging techniques in medical diagnosis. The same excitement that surrounded Roentgen followed Allan Macleod Cormack and Godfrey Newbold Hounsfield. Cormack, working in the Physics Department of Tufts University, elegantly demonstrated the mathematical rudiments of image reconstruction in a remarkable paper published in 1963. Less than a decade later, Hounsfield, of the British firm ETD Limited, unveiled an incredible engineering achievement: the first commercial instrument capable of obtaining digital axial images with high-contrast resolution for medical purposes. In recognition of the significant advances made possible by the development of computerized tomography, Cormack and Hounsfield shared the award of the Nobel Prize in physiology and medicine in 1979. The excitement of their discovery has not yet subsided.

Ionizing radiation (i.e., radiation capable of producing ion pairs) as generally employed for medical imaging involves the following:

- Radiation that is introduced into the body, thereby making the patient the "source" of radiation emissions, or
- Externally produced radiation, which passes through the patient and is detected by radiation-sensitive devices "behind" the patient.

FIGURE 15.1 An early x-ray image, dated 1895. *Courtesy of http://graphics8.nytimes.com/images/2007/06/19/science/19docs.1901.jpg.*

During cellular or organ system function studies, gamma or x-rays are emitted and detected outside the patient, providing physiological (rate of decay or "washout") as well as anatomical (imaging) information. More than 90 percent of the diagnostic procedures in nuclear medicine use emission-imaging techniques. On the other hand, externally produced radiation, which passes through the patient, is the basis of operation for x-ray machines and computerized tomography. These devices and the procedures involving them are housed within departments of radiology. The purpose of this chapter is to present the fundamental principles of operation for each of these radiation imaging modalities.

15.2 EMISSION IMAGING SYSTEMS

Nuclear medicine is that branch of medicine that employs emission scanning for the purpose of helping physicians arrive at a proper diagnosis. An outgrowth of the atomic age and ushered in by the advances made in nuclear physics and technology during World War II, nuclear medicine emerged as a powerful and effective approach in detecting and treating specific physiological abnormalities.

The field of nuclear medicine is a classic example of a medical discipline that has embraced and utilized the concepts developed in the physical sciences. Conceived as a "joint venture" between the clinician and the physical scientist, it has evolved into an interdisciplinary field of activity with its own body of knowledge and techniques. In the process, the domain of nuclear medicine has grown to include studies pertaining to the following:

- The creation and proper utilization of radioactive tracers (or radiopharmaceuticals) that can be safely administered into the body.
- The design and application of nuclear instrumentation devices and systems to detect and display the activity of these radioactive elements.
- The determination of the relationship between the activity of the radioactive tracer and specific physiological processes.

To better understand this radiation imaging modality, it is necessary to discuss radioactivity, its detection, and the instruments available to monitor the activity of radioactive materials.

15.2.1 Basic Concepts

In 1895, when Roentgen announced the discovery of a new type of penetrating radiation that he called "x-rays," he opened a new realm of scientific inquiry. X-rays are a form of electromagnetic (EM) energy just like radio waves and light. The main difference between x-rays and light or radio waves, however, is in their frequency or wavelength. Figure 15.2 shows the EM radiation spectrum. It will be noted that x-rays typically have a wavelength from 100 nm to 0.01 nm, which is much shorter than radio or light waves.

Spurred on by Roentgen's discovery, the French physicist Henri Becquerel investigated the possibility that known fluorescent or phosphorescent substances produced a type of radiation similar to the x-rays discovered by Roentgen. In 1896, Becquerel announced that certain uranium salts also radiated—that is, emitted penetrating radiations, whether or

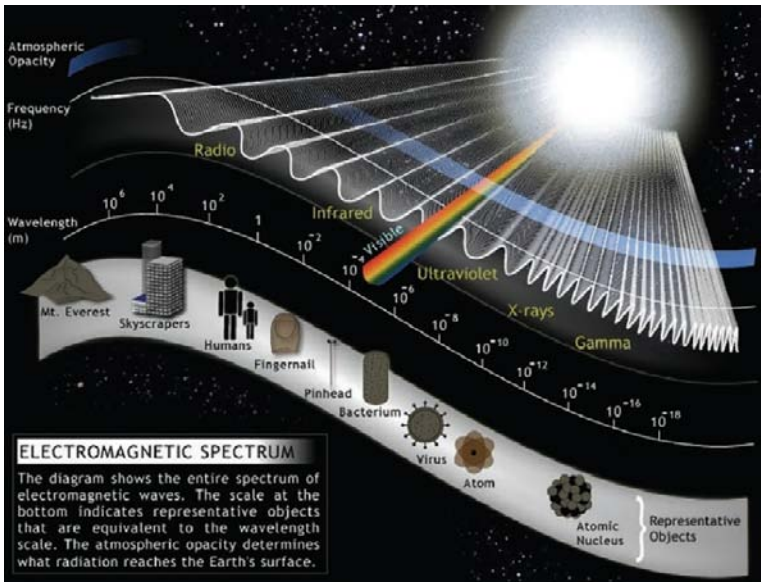


FIGURE 15.2 Electromagnetic wave spectrum. *Compliments of http://ds9.ssl.berkeley.edu/LWS_GEMS/pdfs/em-spectrum.pdf.*

not they were fluorescent. These results were startling and presented the world with a new and entirely unexpected property of matter.

Urged on by scientific curiosity, Becquerel convinced Marie Curie, one of the most promising young scientists at the École Polytechnique, to investigate exactly what it was in the uranium that caused the radiation he had observed. Madame Curie, who subsequently was to coin the term *radioactivity* to describe the emitting property of radioactive materials, and her husband Pierre, an established physicist by virtue of his studies of piezoelectricity, became engrossed in their search for the mysterious substance. In the course of their studies, the Curies discovered a substance far more radioactive than uranium. They called this substance "radium" and announced its discovery in 1898.

While the Curies were investigating radioactive substances, there was a tremendous flurry of activity among English scientists who were beginning to identify the constituent components of atoms. Since the turn of the nineteenth century, Dalton's chemical theory of atoms in which he postulated that all matter is composed of atoms that are indivisible reigned supreme as the accepted view of the internal composition of matter. However, in 1897, when J. J. Thompson identified the electron as a negatively charged particle having a much smaller mass than the lightest atom, a new concept of the basic elements comprising matter had to be formulated. Our present image of atoms was developed in 1911 by Ernest Rutherford, who demonstrated that the principal mass of an atom was concentrated in a dense positively charged nucleus surrounded by a cloud of negatively charged electrons. This idea became incorporated into the planetary concept of the atom, a nucleus surrounded by very light orbiting electrons, which was conceived by Niels Bohr in 1913.

With this insight into the elementary structure of atoms, Rutherford was the first to recognize that radioactive emissions involve the spontaneous disintegration of atoms. After the general acceptance of this basic concept, Rutherford was awarded a Nobel Prize in chemistry in 1908, and the "mystery" began to unfold. By observing the behavior of these radioactive emissions in a magnetic field, the Curies discovered that there are three distinct types of active radiation emitted from radioactive material. The three, arbitrarily called *alpha, beta,* and *gamma* by Rutherford, are now known to be (1) alpha (α) particles, which are positively charged and identical to the nucleus of the helium atom; (2) beta (β) particles, which are negatively charged electrons; and (3) gamma (γ) rays, which are pure electromagnetic radiation with zero mass and charge.

15.2.2 Elementary Particles

By the 1930s, the research in this field clearly identified three elementary particles: the electron, the proton, and the neutron, usually considered the building blocks of atoms. Consider the arrangement of these particles as shown in Figure 15.3, a common view of the atom. The atom includes a number of particles: (1) one or more electrons, each having a mass of about 9.1×10^{-31} kg and a negative electrical charge of 1.6×10^{-19} coulombs; (2) at least one proton with a mass of 1.6×10^{-27} kg, which is approximately 1,800 times that of the electron; and (3) perhaps neutrons, which have the same mass as protons but possess no charge.

Neutrons and protons exist together in the nucleus and have been given the collective name of *nucleons*. The total number of nucleons in the nucleus of an element is called the atomic mass, or mass number, and is represented by the symbol (A), whereas the number of protons alone is referred to as the atomic number (Z). Atoms having the same A, such as Carbon-12 and Boron-12, are called isobars. Atoms with the same number of neutrons are called isotones.

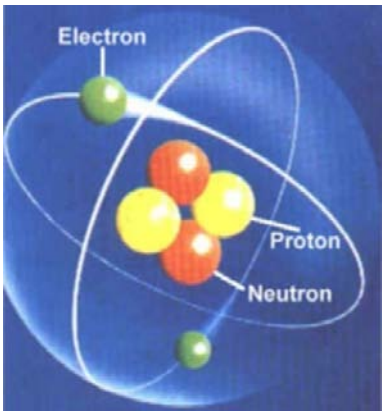


FIGURE 15.3 A common view of the atom. *Compliments of http://www.eskom.co.za/nuclear_energy/fuel/atom.jpg.*

The various combinations of neutrons and protons that may exist in nature are illustrated by examining the composition of the nuclei of the three types of hydrogen atom (hydrogen, deuterium, and tritium). Although hydrogen has a nucleus consisting of a single proton, the combination of one proton and one neutron exists as a single particle called a *demtron* and is the nucleus of the atom called heavy hydrogen, or deuterium. Extending this concept further, the combination of two protons and two neutrons forms a stable particle—the alpha particle—that in nature exists as the nucleus of the helium atom. Alpha radiation emitted from radioactive substances consists essentially of a stream of such particles. Tritium, the third atom of hydrogen, on the other hand, has a nucleus consisting of only one proton and two neutrons.

These three types of hydrogen are examples of atoms whose nuclei have the same number of protons (i.e., have the same atomic number (Z)) and at the same time may have a different number of neutrons and, thereby, a different atomic mass (A). Atoms exhibiting this characteristic were given the name *isotope* (from the Greek, meaning "same place"). The term *isotope* has been widely used to refer to any atom, particularly a radioactive one. However, current usage favors the word *nuclide* to refer to a particular combination of neutrons and protons. Thus, isotopes are nuclides that have the same atomic number. All elements with an atomic number (Z) greater than 83 and/or atomic mass (A) greater than 209 are radioactive—that is, they decay spontaneously into other elements, and this decay causes the emission of active particles.

In order to specifically designate each individual atom, certain symbols have been developed. In the process of reading the literature in this field, it is therefore necessary to become somewhat familiar with them. In the United States, it was the practice to place the atomic number as a subscript before and the atomic mass as a superscript after the chemical symbol of the atom ($_{53}I^{131}$). Since the chemical symbol itself also specifies the atomic number, one often omits it and simply writes I^{131}. In Europe, on the other hand, it was customarily written as a superscript prior to the chemical symbol (^{131}I). In an effort to achieve international standards, it was agreed in 1964 that the atomic mass should be placed as a superscript preceding the chemical symbol (^{131}I). When superscripts are not used, a more literal form of designation, such as cobalt 60, is commonly used. Referring to the helium atom, one simply writes ^{4}He, where 4 is equal to the atomic weight (because of the aggregation of protons and neutrons in the nucleus).

Since the mass of an individual nuclear particle is very small and, when expressed in grams, involves the use of unwieldy negative exponents, a system of atomic mass units (amu) has been developed that uses carbon 12 ($^{12}_{6}\leftarrow C$) as a reference atom. The arbitrary value of 12 mass units has been assigned to carbon-12. The masses of all other atoms are based on a unit, that is $^{1}/_{12}$ the mass of the carbon-12 atom. The mass of the lightest isotope of hydrogen is thus approximately 1 amu.

The mass in grams of an isotope numerically equal to its atomic mass is called a gram-atomic mass. A gram-atomic mass of a substance is also referred to as a mole or equivalent mass of the substance. Since gram-atomic masses have magnitudes proportional to the actual masses of the individual atoms, it follows that one mole of any substance contains a definite number of atoms. The number of atoms in one gram-atomic mass is given by Avogadro's number (N_A) and is equal to 6.023×10^{23}.

The mass of any atom in grams can thus be found by dividing the gram-atomic mass of the isotope by Avogadro's number. For example, the mass of an atom of carbon-12 is its gram-atomic mass (12 grams) divided by Avogadro's number (N_A):

$$\text{mass of } {}^{12}\text{C} = \frac{12 \text{ gm}}{6.023 \times 10^{23}} = 1.99 \times 10^{-23} \text{ gm}$$

Since ^{12}C is equal to 12 atomic mass units (amu), the mass of 1 amu would be $^1/_{12}$ the mass of a carbon-12 atom.

$$1 \text{ amu} = \frac{1.99 \times 10^{-23} \text{gm}/{}^{12}\text{C atoms}}{12 \text{ amu}/{}^{12}\text{C atom}}$$

$$1 \text{ amu} = 1.66 \times 10^{-24} \text{ gm}$$

The electrical charge carried by the electron is a fundamental property of matter, as is the mass of a particle. Since this is the smallest amount of electricity that can exist, it is usually expressed as a negative unit charge (–1). When expressing a unit charge in the metric system of units (meters, kilograms, and seconds), one obtains the value 1.6×10^{-19} coulombs. The charges carried by all other atomic particles are therefore some integral multiple of this value. Consequently, it is impossible for a particle to have, for example, a charge equal to two and one-half times that of the electron.

15.2.3 Atomic Structure and Emissions

As already mentioned, Bohr's model suggested the existence of an atomic structure analogous to the planetary system. In this system electrons rotated in discrete orbits or shells around the nucleus, and the orbital diameters were determined by a quantum number (n) having integer values. These orbits were then represented by K, L, M, and N, corresponding to an increasing number of n, a nomenclature still in use today. This model was further refined in 1925 by Wolfgang Pauli in terms of quantum mechanical principles. Pauli's work on atomic structure explained various observed phenomena, including the estimates for binding energy of the electrons at various orbits of an atom. Pauli further observed that an atom can be defined by four quantum numbers: (1) n is the principal quantum number, which is an integer and scalar quantity; (2) l is the angular momentum quantum number, a vector quantity that has integral values ranging from 0 to $n-1$; (3) m_1 is the magnetic quantum number with integral values ranging from –1 to +1; (4) m_s is the spin magnetic quantum number, which has the values of +½ and –½. According to the Pauli exclusion principle, no two electrons in an atom can have the same set of quantum numbers.

In an electrically neutral atom, the number of orbital electrons exactly balances the number of positive charges in the nucleus. The chemical properties of an atom are determined by the orbital electrons, since they are predominantly responsible for molecular bonding, light spectra, fluorescence, and phosphorescence. Electrons in the inner shells, on the other hand, are more tightly bound and may be removed from their orbits only by considerable energy such as by radiation interaction.

The amount of energy required to eject an electron from an orbit is the *binding energy* for that shell, which is highest for the electrons at the innermost shell. The energy required to move an electron from an inner shell to an outer shell is equal to the difference in binding energies between the two shells. This energy requirement represents one of the natural characteristics of an element. The elementary particle of electromagnetic radiation is the *photon*. When the binding energy is released as a photon, in the case of transition of an electron from an outer shell to an inner shell, it is known as a characteristic x-ray. However, if instead of the emission of a photon the energy is transferred to another orbital electron, called an Auger electron, it will be ejected from orbit. The probability for the yield of a characteristic x-ray in such a transition is known as fluorescent yield.

Protons and neutrons (also known as nucleons) experience a short-range nuclear force that is far greater than the electromagnetic force of repulsion between the protons. The movement of nucleons is often described by a shell model, analogous to orbital electrons. However, only a limited number of motions are allowed, and they are defined by a set of nuclear quantum numbers. The most stable arrangement is known as the ground state. The other two broad arrangements are (1) the metastable state, when the nucleus is unstable but has a relatively long lifetime before transforming into another state; and (2) the excited state, when the nucleus is so unstable that it has only a transient existence before transforming into another state. Thus, an atomic nucleus may have separate existence at two energy levels, known as *isomers* (both have the same Z as well as the same A). An unstable nucleus ultimately transforms itself to a more stable condition, either by absorbing or releasing energy (photons or particles) to a nucleus at ground state. This process is known as radioactive transformation or decay. As just stated, naturally occurring heavier elements, having Z greater than 83, are all unstable.

Assessment of nuclear binding energy is important in determining the relative stability of a nuclide. This binding energy represents the minimum amount of energy necessary to overcome the nuclear force required to separate the individual nucleons. This can be assessed on the basis of mass-energy equivalence as represented by $E = mc^2$, where E, m, and c represent energy, mass, and speed of light, respectively. This has led to the common practice of referring to masses in terms of electron-volts (eV). The mass of an atom is always found to be less than the sum of the masses of the individual components (neutrons, protons, and electrons). This apparent loss of mass (Δm), often called mass defect or deficiency, is responsible for the binding energy of the nucleus and is equivalent to some change in energy ($\Delta\ mc^2$). As mentioned previously, the mass of a neutral carbon-12 atom has been accepted as 12.0 atomic mass units (amu). The sum of the masses of the components of ^{12}C, however, is 12.10223 amu. The difference in masses (0.10223 amu) is equivalent to 95.23 mega-electron-volts (MeV) of binding energy for this nucleus, or 7.936 MeV/nucleon (obtained by dividing MeV by A = 12) for carbon-12. For nuclei with atomic mass numbers greater than 11, the binding energy per nucleon ranges between 7.4 and 8.8 MeV. One atomic mass unit (1 amu), therefore, is equal to $1.6605655 \times 10^{-24}$ g or, using the mass-energy relation, is equivalent to 931.502 MeV. The world record for proton acceleration, set in November 2009 by the Large Hadron Collider in Geneva, stands at an energy level of 1.18 trillion eV. The resting mass of an electron, on the other hand, is very small: only 0.511 MeV.

EXAMPLE PROBLEM 15.1

Verify that the energy released by 1 amu is 931 MeV.

Solution

The conversion can be expressed by Einstein's equation

$$E = mc^2$$

where m is the mass in grams, and c is the velocity of light in cm/sec. Then, the energy equivalent (E) of one atomic mass unit is given by

$$E = (1.66 \times 10^{-24}\ \text{gm})(3 \times 10^{10}\ \text{cm/sec})^2$$
$$E = 1.49 \times 10^{-3}\ \text{gm-cm}^2/\text{sec}^2$$

The unit gm-cm^2/sec^2 is frequently encountered in physics and is termed the *erg*, so

$$E = 1.49 \times 10^{-3}\ \text{ergs}$$

Another convenient unit of energy is the *electron volt*. An electron volt is defined as the amount of kinetic energy acquired by an electron when it is accelerated in an electric field produced by a potential difference of one volt. Since the work done by a difference of potential V acting on a charge e is Ve, and the charge on one electron is 1.6×10^{-19} coulombs, it is possible to calculate the amount of energy in one electron volt as follows:

$$1\ \text{eV} = 1.6 \times 10^{-19}\ \text{coulomb} \times 1\ \text{volt}$$
$$1\ \text{eV} = 1.6 \times 10^{-19}\ \text{joules}$$

or

$$1\ \text{eV} = 1.6 \times 10^{-12}\ \text{ergs}$$

Since the electron volt is a very small amount of energy, it is more commonly expressed in thousands of electron volts (keV) or millions of electron volts (MeV):

$$1\ \text{MeV} = 1,000,000\ \text{eV}$$
$$1\ \text{keV} = 1,000\ \text{eV}$$

It is now possible to express the atomic mass unit in MeV:

$$1\ \text{amu} = 1.49 \times 10^{-3}\ \text{ergs, and}$$
$$1\ \text{Mev} = 1.60 \times 10^{-6}\ \text{ergs; therefore,}$$
$$1\ \text{amu} = 931\ \text{MeV}$$

Therefore, if one amu could be completely transformed into energy, 931 MeV would result. To gain a concept of the magnitude of the electron volt, one million electron volts (MeV) is enough to lift only a milligram weight one-millionth of a centimeter.

An unstable nuclide, commonly known as a *radionuclide*, eventually comes to a stable condition with the emission of ionizing radiation after a specific probability of life expectancy. In general, there are two classifications of radionuclides: natural and artificial. Naturally occurring radionuclides are those nuclides that emit radiation spontaneously

and therefore require no additional energy from external sources. Artificial radionuclides, on the other hand, are essentially man-made and are produced by bombarding so-called stable nuclides with high-energy particles. Both types of radionuclides play an important role in emission scanning and nuclear medicine. The average life or half-life, the mode of transformation or decay, and the nature of emission (type and energy of the ionizing radiation) constitute the basic characteristics of a radionuclide.

All radioactive materials, whether they occur in nature or are artificially produced, decay by the same types of processes: emitting alpha, beta, and/or gamma radiations. As just discussed, the emission of alpha and beta particles involves the disintegration of one element, often called the parent, into another, the daughter. The modes or phases of transformation in the process of a radionuclide passage from an unstable to stable condition may be divided into six different categories: α (alpha) decay/emission, β- (negatron) decay, β+ (positron) decay, electron capture (EC), isomeric transition (IT), and fission.

Alpha (α) Decay

Alpha particles are ionized helium atoms ($^{4}_{2}$H) moving at a high velocity (recall $^{A}_{Z}$I notation). If a nucleus emits an alpha particle, it loses two protons and two neutrons, thereby reducing Z by 2 and A by 4. Several naturally occurring heavy elements undergo such decay. For example, a parent nucleus $^{238}_{92}$U emits an alpha particle, thereby changing to a daughter nucleus $^{234}_{90}$Th (Thorium). In symbolic form, this process may be written as

$$^{238}_{92}\mathrm{U} \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}$$

Note the following facts about this reaction: (1) the atomic number (number of protons) on the left is the same as on the right ($92 = 90 + 2$), since charge must be conserved, and (2) the mass number (protons plus neutrons) on the left is the same as on the right ($238 = 234 + 4$).

When one element changes into another, as in alpha decay, the process is called *transmutation*. In order for alpha emission to occur, the mass of the parent must be greater than the combined mass of the daughter and the alpha particle. In the decay process, this excess mass is converted into energy and appears in the form of kinetic energy in the daughter nucleus and the alpha particle. Most of the kinetic energy is carried away by the alpha particle because it is much less massive than the daughter nucleus. That is, because momentum must be conserved in the decay process, the lighter alpha particle recoils with a much higher velocity than the daughter nucleus. Generally, light particles carry off most of the energy in nuclear decays.

EXAMPLE PROBLEM 15.2

Radium, $^{226}_{88}$Ra, decays by alpha emission. What is the daughter element formed?

Solution

The decay can be written symbolically as

$$^{226}_{88}\mathrm{Ra} \rightarrow \mathrm{X} + {}^{4}_{2}\mathrm{He}$$

where X is the unknown daughter element. Requiring that the mass numbers and atomic numbers balance on the two sides of the arrow, we find that the daughter nucleus must have a mass number of 222 and an atomic number of 86:

$$^{226}_{88}\text{Ra} \rightarrow ^{222}_{86}\text{X} + ^{4}_{2}\text{He}$$

The periodic table shows that the nucleus with an atomic number of 86 is radon, Rn.

EXAMPLE PROBLEM 15.3

In Example Problem 15.2, we showed that the $^{226}_{88}$Ra nucleus undergoes alpha decay to $^{222}_{86}$Rn. Calculate the amount of energy liberated in this decay. Take the mass of $^{226}_{88}$Ra to be 226.025406 amu, that of $^{222}_{86}$Rn to be 222.017574 amu, and that of $^{4}_{2}$He to be 4.002603 amu.

Solution

After decay, the mass of the daughter, m_d, plus the mass of the alpha particle, m_α, is

$$m_d + m_\alpha = 222.017574 \text{ amu} + 4.002603 \text{ amu} = 222.020177 \text{ amu}$$

Thus, calling the mass of the parent nucleus M_p, we find that the mass lost during decay is

$$\Delta m = M_p - (m_d + m_\alpha) = 226.025406 \text{ amu} - 226.020177 \text{ amu} = 0.005229 \text{ amu}$$

Using the relationship 1 amu = 931.5 MeV, we find that the energy liberated is

$$E = (0.005229 \text{ amu})(931.50 \text{ MeV/amu}) = 4.87 \text{ MeV}$$

Negatron (β-) or (β-, γ) Decay

When a radioactive nucleus undergoes beta decay, the daughter nucleus contains the same number of nucleons as the parent nucleus, but the atomic number (Z) is increased by 1. A typical beta decay event is

$$^{14}_{6}\text{C} \rightarrow ^{14}_{7}\text{N} + ^{0}_{-1}\text{e}$$

The superscripts and subscripts on the carbon and nitrogen nuclei follow our usual conventions, but those on the electron may need some explanation. The –1 indicates that the electron has a charge whose magnitude is equal to that of the proton but is negative. The 0 used for the electron's mass number indicates that the mass of the electron is almost zero relative to that of carbon and nitrogen nuclei.

The emission of electrons from a nucleus is surprising because the nucleus is usually thought to be composed of protons and neutrons only. This apparent discrepancy can be explained by noting that the electron that is emitted is created in the nucleus by a process in which a neutron is transformed into a proton. This can be represented by the following equation:

$$^{1}_{0}\text{n} \rightarrow ^{1}_{1}\text{p} + ^{0}_{-1}\text{e}$$

EXAMPLE PROBLEM 15.4

Find the energy liberated in the beta decay of $^{14}_{6}C$ to $^{14}_{7}N$.

Solution

$^{14}_{6}C$ has a mass of 14.003242 amu, and $^{14}_{7}N$ has a mass of 14.003074 amu. Here, the mass difference between the initial and final states is

$$\Delta m = 14.003242 \text{ amu} - 14.003074 \text{ amu} = 0.000168 \text{ amu}$$

This corresponds to an energy release of

$$E = (0.000168 \text{ amu})(931.50 \text{ MeV/amu}) = 0.156 \text{ MeV}$$

Only a small number of electrons have this kinetic energy. Most of the emitted electrons have kinetic energies less than this predicted value. If the daughter nucleus and the electron are not carrying away this liberated energy, then the requirement that energy is conserved leads one to ask the question, "What accounts for the missing energy?"

In 1930, Pauli proposed that a third particle must be present to carry away the "missing" energy and to conserve momentum. Enrico Remi later named this particle the "neutrino" (little neutral one) because it had to be electrically neutral and have little or no resting mass. Although it eluded detection for many years, the neutrino (symbol v) was finally detected experimentally in 1950. The neutrino has the following properties:

1. It has zero electric charge.
2. It has a resting mass smaller than that of the electron.
3. It interacts very weakly with matter and is therefore very difficult to detect.

Phosphorus 32 is a typical example of a pure β- emitter (^{32}P is transformed to ^{32}S) that has been used for therapy.

Very often a nucleus that undergoes radioactive decay is left in an excited energy state. The nucleus can then undergo a second decay to an even lower energy state by emitting one or more photons. The process is very similar to the emission of light by an atom. An atom emits radiation to release some extra energy when an electron "jumps" from a state of high energy to a state of lower energy. Likewise, the nucleus uses essentially the same method to release any extra energy it may have following a decay or some other nuclear event. In nuclear de-excitation, the "jumps" that release energy are made by protons or neutrons in the nucleus as they move from a higher energy level to a lower level. The photons emitted in such a de-excitation process are called gamma rays and have very high energy relative to the energy of visible light. Most of the radionuclides undergoing β- decay also emit γ rays, almost simultaneously. For example, iodine 131 emits several β- and γ rays in this process.

The following sequence of events represents a typical situation in which gamma decay occurs:

$$^{12}_{5}B \rightarrow {}^{12}_{6}C^{*} + {}^{0}_{-1}e$$
$$^{12}_{6}C^{*} \rightarrow {}^{12}_{6}C + \gamma$$

The first process represents a beta decay in which ^{12}B decays to ^{12}C*, where the asterisk is used to indicate that the carbon nucleus is left in an excited state. The excited carbon nucleus then decays to a ground state by emitting an x-ray. Note that an x-ray emission does not result in any change in Z or A.

Positron (β+) (β+, γ) Decay

With the introduction of the neutrino, the beta decay process in its correct form can be written as

$$^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}\mathrm{N} + {}^{0}_{-1}\mathrm{e} + -v$$

where the bar in the symbol $-v$ indicates that this is an antineutrino. To explain what an antineutrino is, consider the following decay process:

$$^{12}_{7}\mathrm{N} \rightarrow {}^{12}_{6}\mathrm{C} + {}^{0}_{1}\mathrm{e} + -\mathrm{v}$$

When ^{12}N decays into ^{12}C, a particle is produced that is identical to the electron except that it has a positive charge of $+e$. This particle is called a positron. Because it is like the electron in all respects except charge, the positron is said to be antiparticle to the electron.

It should be noted that positrons (β+) are created in the nucleus, as if a proton was converted to a neutron and a positron. Positron decay produces a different element by decreasing Z by 1, with A being the same. For example, carbon-11 decays to the predominant stable isotope of boron (i.e., ^{11}C transforms to ^{11}B). Positrons (like negatrons) share their energy with neutrinos. Positron decay is also associated with γ-ray emission. The positron, once emitted, however, is often annihilated as a result of a collision with an electron within one nanosecond. This produces a pair of photons of 0.511 MeV that move in opposite directions. A minimum transition energy of 1.022 MeV is required for any positron decay.

EXAMPLE PROBLEM 15.5

^{12}N beta decays to an excited state of ^{12}C, which subsequently decays to the ground state with an emission of 4.43 MeV gamma ray. What is the maximum energy of the emitted beta particle?

Solution

The decay process for positive beta emission is

$$^{12}_{7}\mathrm{N} \rightarrow {}^{12}_{6}\mathrm{C}^{*} + {}^{+}\mathrm{e} + v \rightarrow {}^{12}_{6}\mathrm{C} + \gamma$$

Conservation of energy tells us that a nucleus X will decay into a lighter nucleus X^1 with an emission of one or more particles (collectively designated as x) only if the mass of X is greater than the total mass $X^1 + x$. The excess mass energy is known as the Q value of the decay. In this process

$$\mathrm{Q} = [\mathrm{m}(^{12}_{7}\mathrm{N}) - \mathrm{m}(^{12}_{6}\mathrm{C}^{*}) - 2\mathrm{m_e}]\mathrm{c}^2$$

to determine Q for this decay, we first need to find the mass of the product nucleus ^{12}C in its excited state. In the ground state ^{12}C has a mass of 12.000000 amu, so its mass in the excited state is

$$12.000000\ \text{amu} + \frac{4.43\ \text{MeV}}{931.5\ \text{MeV/amu}} = 12.004756\ \text{amu}$$

$$\therefore \mathrm{Q} = [12.018613 - 12.004756\ \text{amu} - 2 \times 0.000549\ \text{amu}]\ 931.5\ \text{MeV/amu} = 11.89\ \text{MeV}$$

Electron Capture or K Capture

An orbital electron, usually from the inner shell (K shell), may be captured by the nucleus (as if a proton captured an electron and converted itself into a neutron). An electron capture process produces a different element by decreasing Z by 1, with A being the same (similar to $\beta+$ decay). Indeed, some radionuclides have definite probabilities of undergoing either positron decay or electron capture (such as iron 52, which decays with about 42 percent EC and 58 percent $\beta+$ emission). Further, there could be an associated gamma emission with electron capture. For example, ^{51}Cr (chromium 51) transforms to ^{51}V (vanadium 51), with about 90 percent going directly to the ground state. The remaining 10 percent goes to an excited state of ^{51}V followed by transition to ground state with the emission of photons. An electron capture also causes a vacancy in the inner shell, which leads to the emission of a characteristic x-ray or Auger electron. Absence of high-energy electrons (beta particles) in EC causes a low dose of absorbed radiation to the tissue.

Isomeric Transition and Internal Conversion

Radionuclides at a metastable state emit only γ rays. The element remains the same with no change in A (isomeric transition). The atomic mass number of the isomer is, therefore, denoted by Am. For example, ^{99}mTc (technetium 99m) decays to ^{99}Tc. However, there is a definite probability that instead of a photon coming out, the energy may be transferred to an inner orbital electron. This is known as internal conversion, and the internally converted electrons are close to monoenergetic beta particles. For example, barium 135m (decaying by IT) emits about 84 percent IC electrons. They also create a vacancy in the shell, consequently leading to the emission of characteristic x-rays and Auger electrons.

Nuclear Fission

Usually a heavy nuclide will break up into two nuclides (more or less equal fragments). This may happen spontaneously, but is more likely with the capture of a neutron. The uranium fission products mostly range between atomic numbers 42 and 56. A number of medically useful radionuclides are produced as fission products, such as Xenon 133, which may be extracted by appropriate radiochemical procedures.

15.2.4 Radioactive Decay

The reduction of the number of atoms through disintegration of their nuclei is known as radioactive decay and is characteristic of all radioactive materials. Unaffected by changes in temperature, pressure, or chemical combination, the rate of the decay process remains constant with the same number of disintegrations occurring during each interval of time. Furthermore, this decay process is a random event. Consequently, every atom in a radioactive element has the same probability of disintegrating.

As the decay process continues, it is clear that fewer atoms will be available to disintegrate. This fraction of the remaining number of atoms that decay per unit of time is called the decay constant (λ).

One of the most common terms encountered in any discussion of radioactive materials is *half-life*, $T_{1/2}$. This is because all radioactive substances follow the same general decay pattern. After a certain interval of time, half of the original number of nuclei in a sample will have decayed, then in a second time interval equal to the first, half of those nuclei remaining will have decayed, and so on.

The half-life is the time required for half of a given number of radioactive nuclei to decay. The half-life and decay constant are obviously related, since the larger the value of the decay constant (λ), the faster the process of decay and consequently the shorter the half-life. In any event, the decay constant is an unchanging value throughout the decay process. Each radionuclide exhibits a distinctive disintegration process because of its inherent properties—that is, its decay constant (λ) and half-life ($T_{1/2}$). All nuclides decay in the same manner but not at the same rate, since this is a parameter determined by the unique nature of the particular radioactive element in question.

Considering a number of atoms (N) of a specific type of radionuclide present at a time t, the transformation rate can be defined by $-dN/dt$ (the minus sign denotes the decay/decrease), which would be proportional to the number of atoms, or

$$dN/dt = \lambda N \tag{15.1}$$

where λ is the transformation constant.

Taking the initial number of atoms as N_0 ($N = N_0$ when $t = t_0$), we get by integration

$$N = N_0 e^{-\lambda t} \tag{15.2}$$

where N is the number of radioactive nuclei present at time t, N_0 is the number present at time $t = 0$, and $e = 2.718\ldots$ is the base of the natural logarithm (see Figure 15.3).

The unit of activity is the *curie* (Ci), defined as

$$1\ \text{Ci} \equiv 3.7 \times 10^{10}\ \text{decays/s}$$

This number of decay events per second was selected as the original activity unit because it is the approximate activity of 1 g of radium. The SI unit of activity is the *becquerel* (Bq):

$$1\ \text{Bq} = 1\ \text{decay/s}$$

Therefore, $1\ \text{Ci} = 3.7 \times 10^{10}$ Bq. The most commonly used units of activity are the millicurie (10^{-3} Ci) and the microcurie (10^{-6} Ci).

The half-life, or $T_{1/2}$ (the time corresponding to transformation of 50 percent of the nuclides, when $N = N_0/2$), therefore may be obtained by solving for $\lambda T_{1/2}$ in Eq. (15.2). Therefore,

$$N = N_0 e^{-\lambda t}$$

and

$$T_{1/2} = ln\ 2/\lambda = 0.693/\lambda$$

EXAMPLE PROBLEM 15.6

Derive the equation for $T_{1/2}$.

Solution

Start with $N = N_0 e^{-\lambda t}$. After a time interval equal to one half-life, $t = T_{1/2}$, the number of radioactive nuclei remaining is $N = N_0/2$. Therefore,

$$N_0/2 = N_0 e^{-\lambda T_{1/2}}$$

Dividing both sides by N_0 gives

$$1/2 = e^{-\lambda T_{1/2}}$$

Taking the natural logarithm of both sides of this equation eliminates the exponential factor on the right, since $\ln e = 1$:

$$\ln 1/2 = -\lambda T_{1/2}$$

Since $\ln 1/2 = -0.693$ (use your calculator to check this), we have

$$-0.693 = -\lambda T_{1/2}$$

$$T_{1/2} = \frac{0.693}{\lambda}$$

Let us assume that the number of radioactive nuclei present at $t = 0$ is N_0. The number left after one half-life, $T_{1/2}$, passes is $N_0/2$. After the second half-life, the number remaining is again reduced by one-half. Hence, after a time $2T_{1/2}$, the number remaining is $N_0/4$, and so forth. Half-lives range from about 10^{-22} s to 10^{21} years.

EXAMPLE PROBLEM 15.7

The half-life of the radioactive nucleus $^{226}_{88}$Ra is 1.6×10^3 years. If a sample contains 3×10^{16} such nuclei, determine the activity.

Solution

First, calculate the decay constant, λ, using the fact that

$$T_{1/2} = 1.6 \times 10^3 \text{ years} = (1.6 \times 10^3 \text{ years})(3.15 \times 10^7 \text{ s/year}) = 5.0 \times 10^{10} \text{ s}$$

Therefore,

$$\lambda = \frac{0.693}{T_{1/2}} = \frac{0.693}{5.0 \times 10^{10}\text{s}} = 1.4 \times 10^{-11} \text{ s}^{-1}$$

The activity, or decay rate, of the sample at $t = 0$ is calculated using the form $-dN_0 = dt$, R_0 is the decay rate at $t = 0$, and N_0 is the number of radioactive nuclei present at $t = 0$. Since $N_0 = 3 \times 10^{16}$, we have

$$R_0 = \lambda N_0 = (1.4 \times 10^{-11}\text{s}^{-1})(3 \times 10^{16}) = 4.2 \times 10^5 \text{decays/s}$$

Since 1 Ci $= 3.7 \times 10^{10}$ decays/s, the activity, or decay rate is $R_0 = 11.3$ μCi.

Progress in emission scanning and nuclear medicine has been linked to the availability of radionuclides that could be used in human subjects in appropriate chemical forms. The choice of a radionuclide depends on its physical characteristics in relation to diagnostic and therapeutic applications and the possibility of incorporating it into an appropriate chemical compound suitable for biomedical investigation.

A radiopharmaceutical contains a specific chemical compound labeled with a radionuclide. This may be a simple inorganic salt or a complex organic molecule. It is well recognized that the chemical properties of an element reflect some of the possible biological behaviors, and the behavior of several elements within the same group of the periodic table appears similar. For example, strontium 85 has been used to represent calcium metabolism in bone. Table 15.1 summarizes some of the important radionuclides that are currently used in nuclear medicine.

15.2.5 Measurement of Radiation: Units

It should now be apparent that all radioactive substances decay and in the process emit various types of radiation (alpha, beta, and/or gamma) during this process. However, since they usually occur in various combinations, it is difficult to measure each type of radiation separately. As a result, measuring radioactivity is usually accomplished by (1) counting the number of disintegrations that occur per second in a radioactive material, (2) noting how effective this radiation is in producing atoms (ions) possessing a net positive or negative

TABLE 15.1 Commonly Used Radionuclides in Nuclear Medicine

Radionuclide	Half-life	Transition	Production	Chemical Form
Carbon 11	20.38 min	β^+	Cyclotron	3-*N*-methylspiperone
Flourine 18	109.77 min	β^+	Cyclotron	Flourodeoxyglucose
Phosphorus 31	14.29 days	β^-	Reactor	Phosphates
Chromium 51	27.704 days	EC	Reactor	Sodium chromate
Cobalt 57	270.9 days	EC	Cyclotron	Cyanocobalamin
Gallium 67	78.26 h	EC	Cyclotron	Citrate complex
Molybdenum 99	66.0 h	β^-	Reactor	Molybdate in column
Technetium 99m	6.02 h	IT	Generator	TcO_4 and complexes
Indium 111	2.83 days	EC	Cyclotron	DTPA and oxine
Iodine 123	13.2 h	EC	Cyclotron	Mainly iodide
Iodine 123	60.14 days	EC	Reactor	Diverse proteins
Iodine 131	8.04 days	β^-	Reactor	Diverse compounds
Xenon 133	5.245 days	β^-	Reactor	Gas
Thallium 201	3.044 days	EC	Cyclotron	Thallous chloride

charge, or (3) measuring the energy absorbed by matter from the radiation penetrating it. Using these fundamental concepts, three kinds of radiation units have been established: the curie (Ci), the roentgen (R), and the radiation-absorbed dose (rad).

It is important to remember that the curie defines the number of disintegrations per unit time and not the *nature* of the radiation, which may be alpha, beta, or gamma rays. The curie is simply a measure of the activity of a radioactive source.

The roentgen and the rad, on the other hand, are units based on the effect of the radiation on an irradiated object. Thus, while the curie defines a source, the roentgen and rad define the effect of the source on an object. One of the major effects of x-ray or gamma radiation is the ionization of atoms—that is, the creation of atoms possessing a net positive or net negative charge (ion pair). The roentgen is determined by observing the total number of ion pairs produced by x-ray or gamma radiation in 1 cc of air at standard conditions (at 760 mm Hg and 0°C). Since each ion pair has an electrical charge, this can be related to electrical effects that can be detected by various instruments. Thus, one roentgen is defined as that amount of x-ray or gamma radiation that produces enough ion pairs to establish an electrical charge separation of 2.58×10^{-4} coulombs per kilogram of air at atmospheric pressure. Thus, the roentgen is a measure of radiation quantity, not intensity. The rad, on the other hand, is based on the total energy absorbed by the irradiated material. One rad means that 0.01 joule (the unit of energy in the metric system) of energy is absorbed per kilogram of material.

Since human tissue is exposed to various types of radioactive materials in nuclear medicine, another unit of measure, the rem (roentgen equivalent man), is often used to specify the biological effect of radiation. Consequently, the rem is a unit of human biological dose resulting from exposure of the biological preparation to one or many types of ionizing radiation.

These radiation units are used in various situations in nuclear medicine. The roentgen, or more commonly its submultiple the milliroentgen (mr), is used as a value for most survey meter readings. The rad is used as a unit to describe the amount of exposure received, for example, by an organ of interest on injection of a radiopharmaceutical. The rem is the unit used to express exposure values of some personnel-monitoring devices such as film badges.

The classical method of dosimetry, however, has been replaced by a modern method of radiation dose calculation through the effort of the Medical Internal Radiation Dose (MIRD) Committee. The internationally accepted unit for radiation dose to tissue is the gray (Gy); 1 Gy is equivalent to the absorption of energy of one joule per kilogram of the tissue under consideration. Previously the radiation-absorbed dose was expressed as radiation absorbed dose (rad); 1 rad is equivalent to the absorption of energy of 100 ergs per gram or 0.01 J per kilogram, which is equal to 0.01 Gy. Thus, an expression of 1 rad per millicurie (mCi) is equivalent 0.27027 milligray per megabecquerel (mGy/MBq), or 1 mGy per megabecquerel is equivalent to 3.7 rad per millicurie. Most of the computations on radiation dosimetry that are available for nuclear medicine procedures, however, have been expressed in terms of rads.

Exposure of living organisms, mammals in particular, to a high level of radiation induces pathological conditions and even death. This feature is utilized in therapy for malignant diseases where the intention is to deliver a localized high-radiation dose to destroy the undesirable tissue (Figure 15.4). Radiation has in recent years become a common treatment

FIGURE 15.4 A patient undergoing radiation therapy. *Compliments of http://malignant-mesothelioma-cancer.com/mesothelioma/wp-content/uploads/2008/02/radiation-therapy.jpg.*

step in the fight against cancer. Approximately 50 percent of all diagnosed cancer patients undergo some form of radiation therapy. Various forms of radiation are implemented, depending on the depth and size of the targeted area. Unfortunately, radiation will kill both malignant and healthy tissue, and as such exposure to even lower levels of radiation, while not showing any apparent effect, will increase the risk of cancer as a long-term effect. It also increases the probability of genetic defects if the gonads are exposed. This imposes the need for minimization of radiation exposure to the worker, as well as specific guidelines for medical applications.

The United States Nuclear Regulatory Commission (NRC) has adopted standards that limit maximum exposure for the general public to 0.5 sievert (Sv) (Rem) per year. Limits for occupational exposure are 1.25 Sv/3 mo for the whole body and 18.75 Sv/3 mo for the extremities. Thus, one of the most important radiation safety procedures is to monitor both the personnel and the work area. Routine personnel monitoring is usually done with film badges and ring-type finger badges. Radiation survey and wipe tests are carried out at certain intervals and at times of incidental/accidental contamination. The common shielding material is lead, the thickness depending on the energy of gamma rays. Further, radioactive wastes are stored at assigned areas with appropriate shielding. The procedures used for radioactive waste disposal depend on the nature and half-life of the waste. All require a temporary storage facility, and they are finally disposed of locally according to a prescribed set of procedures.

15.3 INSTRUMENTATION AND IMAGING DEVICES

From the discussion of radiation measurement it should be clear that radiation energy can be measured only indirectly—that is, by measuring some effect caused by the radiation. The following are some of the various indirect techniques used to measure radioactivity:

1. *Photography*. The blackening of film when it is exposed to a specific type of radiation such as x-rays (which are the equivalent of gamma rays and will be discussed in the section on externally produced radiation).
2. *Ionization*. The passage of radiation through a volume of gas established in the probe of a gas detector produces ion pairs. The function of this type of detector depends on the collection of these ion pairs in such a way that they may be counted. This technique has been most effective in measuring alpha radiation and least effective in measuring gamma radiation.
3. *Luminescence*. The emission of light not due to incandescence. Since the flash of light produced by the bombardment of a certain type of material with penetrating-type radiation can be detected and processed, this technique is extremely useful. As a matter of fact, the fluorescent effect produced by ionizing radiation is the basis of the scintillation detector discussed following. This type of indirect detection scheme is excellent for observing the presence of all three types of radiation.

15.3.1 Scintillation Detectors

Since the majority of modern detector systems in nuclear medicine utilize probes based on the scintillation principle, it will be described in greater detail. Scintillation is defined as a flash of light emitted when a substance is struck by radioactive material. Detectors may be used for all types of radiation, depending on the particular type of scintillator used and its configuration. Regardless of the application, however, the general technique is the same for all scintillation probes. Certain materials—for example, zinc sulfide and sodium iodide—have the property of emitting a flash of light or scintillation when struck by ionizing radiation. The amount of light emitted is, over a wide range, proportional to the energy expended by the particle in this material. When the scintillator material is placed next to the sensitive surface of an electronic device called a *photomultiplier*, the light from the scintillator is then converted into a series of small electrical pulses whose height is directly proportional to the energy of the incident gamma ray. These electrical pulses can then be amplified and processed in such a way as to provide the operator with information regarding the amount and nature of the radioactivity striking the scintillation detector. Thus, scintillators may be used for diagnostic purposes to determine the amount and/or distribution of radionuclides in one or more organs of a patient.

Figure 15.3 illustrates the basic scintillation detection system. It consists of (1) a detector, which usually includes the scintillation crystal, photomultiplier tubes, and preamplifier; (2) signal processing equipment such as the linear amplifier and the single-channel pulse analyzer; and (3) data display units such as the scaler, scanner, and oscilloscope. Once the radioactive event is detected by the crystal and an appropriate pulse is generated by the photomultiplier circuitry, the resulting voltage pulses are still very small. To avoid any serious loss of information caused by distortion from unwanted signals (such as noise) and to provide a strong enough signal to be processed and displayed, the amplifier is used to increase the amplitude of the pulses by a constant factor. This process is called linear amplification.

In such a system, it should be apparent that because of the wide variation in energies of gamma rays striking the scintillation crystal, the linear amplifier receives pulses having a

wide variation in "pulse height." This is fine if one is interested in detecting all the radiant energy under the probe. If, on the other hand, the operator is interested only in the activity of a specific radionuclide, additional processing is necessary. This is accomplished by the pulse height analyzer. Using this device, the operator can discriminate against all radiant energy other than the one of interest. In this way, the activity of a specific radionuclide is analyzed by allowing only those pulses related to it to be processed. Consequently, single-channel and multichannel pulse height analyzers are commonly found in nuclear medicine laboratories. These processed data must then be displayed in such a way as to provide information regarding the amount of radioactivity present and its location within the body. This information is quite important in determining the status of the organs under investigation. Studies of the amount of radioactivity present as a function of time enable physicians to ascertain whether the organ is functioning properly, whereas studies of the location of the radionuclide enable physicians to display the organ or abnormal tissue. Both types of studies are valuable.

In order to actually provide a measure of the scintillation events occurring during many nonimaging applications, it is often necessary to count these events and provide some means to display this information. The scaler is the most common type of electronic device found in nuclear medicine that accomplishes this task. A scaler is used to count the pulses produced by the detector system and processed by a variety of electronic pieces of equipment. Thus, the scaler is an electronic device that accepts signal pulses representing a range of energy levels (energy of the incident radiation) and counts them. The scaler is usually designed so the operator has a choice between accepting a certain number of counts (preset count) or a predetermined period of time over which the counts can be accumulated (preset time). With preset time the scaling device will count the number of events that occur during a set period of time and will then shut off automatically; thus, the count is the variable. With a preset count, the predetermined number of counts is accumulated, after which the scaler is automatically shut off. In this case, time is the variable.

Both types of data can be observed by the operator by viewing the front panel of the scaler itself. However, this information can also be supplied to other devices for analysis. Consider the case of operating the scaler in the present time mode. This information can be presented in a variety of ways. For example, it can be (1) displayed as counts per unit time (second, minute, etc.) for continual observation by the operator, (2) supplied to a digital printer to provide a running tabulation of the counts as a function of time, or (3) directed to a computer that can retain the counts in memory and perform a variety of calculations on the incoming data as it actually is being collected (if the process is slow enough) and at the same time provide the operator with a visual display (on the screen of an oscilloscope) in the form of a plot of radioactive decay. Thus, once the radiation is detected by the scintillation detector and subsequently measured by the scaler, the operator can obtain information in a variety of formats that may be of clinical value.

Scintillation data output may also be processed through an imaging device. In general these devices take the pulse output from a detector and the electronic processing devices already described and place the pulse in some representation spatially, according to its point of origin in the radioactive source. Obviously, there is some distortion in this representation, since the point of origin lies in a three-dimensional plane, whereas most common instruments are capable of only displaying that point of origin in two dimensions.

However, in spite of this obvious shortcoming, this technique has proved to be of tremendous clinical value.

The most commonly used stationary detector system is the scintillation camera or gamma camera. This device views all parts of the radiation field continuously and is therefore capable of operating almost like a camera—building up an image quickly. Initially developed by Hal Anger in the late 1950s, the first Anger camera reflected the convergence of the disciplines of nuclear physics, electronics, optics, and data processing in a clinical setting. Its ultimate acceptance and further development have had a profound impact not only on the practice of clinical nuclear medicine but on the entire diagnostic process as well. The initial concepts introduced by Anger became basic to the art of imaging specific physiological processes.

15.3.2 The Gamma Camera

The operation of the basic gamma camera is shown in Figure 15.5. The detector of the gamma camera is placed over the organ to be scanned. In order to localize the radiation from a given point in the organ and send it to an equivalent point on the detector, a collimator is placed over the base of the scintillation crystal. Since gamma rays cannot be "bent," another technique must be used to selectively block those gamma rays that, if allowed to continue on their straight-line path, would strike the detector at sites completely unrelated to their points of origin in the subject. This process of selective interference is accomplished by the collimator. To prevent unwanted off-axis gamma rays from striking the crystal, collimators usually contain a large number of narrow parallel apertures made of heavy-metal absorbers.

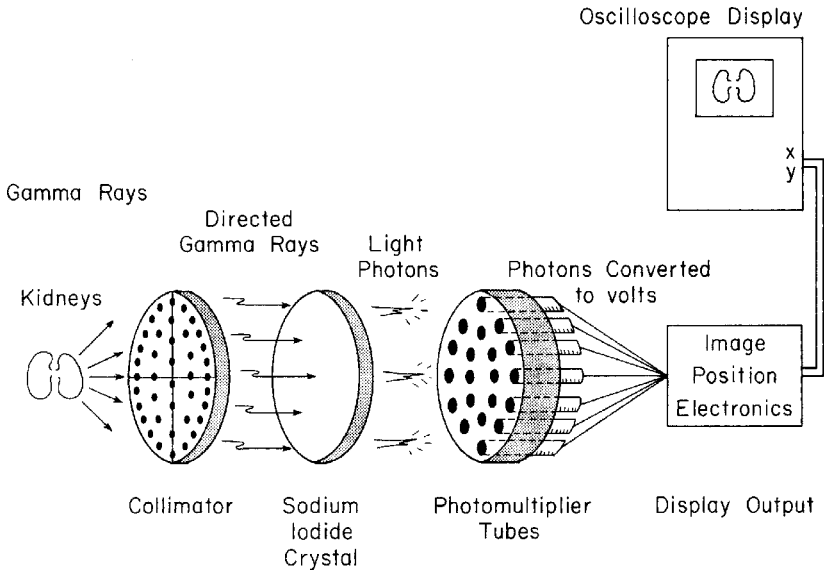


FIGURE 15.5 Basic elements of a gamma curve.

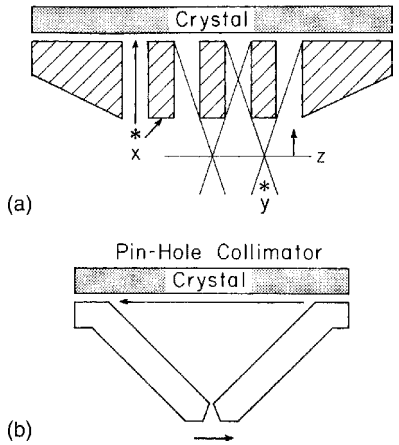


FIGURE 15.6 (a) Multihole collimator used in conjunction with scintillation detector. Event x is represented at only one location in the crystal, whereas event y has multiple sites associated with its occurrence. (b) The pinhole collimator allows magnification and is used for viewing small organs at short range.

Consider the multihole collimator shown in Figure 15.6a. In this case, the collimator consists of a flat lead plate through which narrow holes are drilled. As can be seen, only a gamma event occurring directly under each hole will penetrate the collimator, and it will be represented at only one location on the face of the crystal. If this gamma event occurred much farther away from the collimator, such as in region *Y*, then it would be represented by more than one location on the face of the crystal. When this occurs, the resolution, which may be defined as the ability of the detector to distinguish between two sources at various distances from the collimator, is greatly decreased. For the multihole collimator, then, the best resolution occurs when the area of interest is close to the collimator. Thus, as the subject is moved from point *Z* toward the detector, the resolution is improved. Obviously, when viewing an organ that lies beneath the surface of the skin, it becomes quite important to closely approximate its distance from the probe relative to the degree of resolution required.

There are also other types of collimators—for example, the pinhole collimator shown in Figure 15.6b. The pinhole collimator permits entry of only those rays aimed at its aperture. These gamma rays enter the collimator and proceed in a straight line to the crystal, where they are detected in inverted spatial correspondence to their source. When the source is located at a distance from the collimator equal to that of the pinhole to the crystal, then the source is represented on the crystal in exactly the same size as it exists. However, by proper positioning of the subject, it is possible to actually magnify or decrease the field of view of the detector. That is, as the source is moved closer to the aperture of the collimator, magnification occurs. A pinhole collimator, therefore, enlarges and inverts the image of the source located beneath it.

From this brief discussion, it can be seen that collimators in conjunction with a scintillation detector essentially "focus" the radioactivity occurring at a particular point in space within the organ to a particular point on the surface of the crystal. The radiation passing through the collimator then impinges on or interacts with the scintillation crystal.

A simple gamma camera system, at the present time, is the most important, as well as the most basic, instrument for diagnostic nuclear medicine. Interfacing with a digital computer makes the system even more versatile. Digital image-processing systems perform data collection, storage, and analyses. Data acquisition requires digitization of the image by an analog-to-digital converter (ADC), which divides a rectangular image area into small elements, or pixels, usually a 64 by 64, 128 by 128, or 256 by 256 matrix. As a result, one can select a particular region of interest to obtain certain quantitative information. Digital computation also improves dynamic studies, since the regional rate of uptake and clearance pattern can easily be obtained from the serial images for any particular region, making it possible, for example, to study cardiac wall motion.

15.3.3 Positron Imaging

The positrons emitted through transformation of a radionuclide can travel only a short distance in a tissue (a few millimeters) and then are annihilated. A pair of 511-keV photons that travel at 180° to one another is created. A pair of scintillation detectors can sense positron emission by measuring the two photons in coincidence. Annihilation coincidence detection provides a well-defined cylindrical path between the two detectors. Multihole collimators are unnecessary to define the position in a positron camera because electronic collimation accomplishes the task. Positron cameras have limited use; however, they have attained great importance in a highly modified form in positron-emitting transaxial tomography or positron emission tomography (PET scanning). A large number of small NaI(TI) detectors are arranged in an annular form so annihilation photons in coincidence at 180° permit detection of position. Tomographic images are, however, obtained by computer-assisted reconstruction techniques. Several large centers use positron emission tomography in conjunction with cyclotron-produced short-lived radiopharmaceuticals (mainly carbon-11, nitrogen-13, and fluorine-18) to provide structural as well as metabolic information.

15.4 RADIOGRAPHIC IMAGING SYSTEMS

This section discusses externally produced radiation that passes through the patient and is detected by radiation-sensitive devices behind the patient. These radiographic imaging systems rely upon the differential attenuation of x-rays to produce an image. Initial systems required sizable amounts of radiation to produce distinct images of tissues with good contrast (that is, the ability to differentiate between body parts having minimally different density, such as fat and muscle). High-contrast differences such as between soft tissue and air (as in the lung) or bone and muscle can be differentiated with a much smaller radiation dose. However, with the development of better film and other types of detectors, the radiation dosage has decreased for both high- and low-contrast resolutions. Even more significant was the development of the computerized axial tomography (CAT) scanner in the 1970s. This device produces a cross-sectional view of a patient instead of the traditional shadowgraph recorded by conventional x-ray systems by using computers to reconstruct the x-ray attenuation data. In the process, it provides the clinician with high-contrast resolved images of virtually any body part and can be reconstructed in any body plane. To better understand the significance of this imaging modality, let us review some of the basic concepts underlying routine x-ray imaging.

15.4.1 Basic Concepts

All x-ray imaging systems consist of an x-ray source, a collimator, and an x-ray detector. Diagnostic medical x-ray systems utilize externally generated x-rays with energies of 20–150 keV. Since the turn of the twentieth century, conventional x-ray images have been obtained in the same fashion: by using a broad-spectrum x-ray beam and photographic film. In general, x-rays are produced by a cathode ray tube that generates a beam of x-rays when excited by a high-voltage power supply. This beam is shaped by a collimator and passes through the patient, creating a latent image in the image plane. Depending on the type of radiographic system employed, this image is detected by x-ray film, an image intensifier, or a set of x-ray detectors. Using the standard film screen technique, x-rays pass through the body, projecting an image of bones, organs, air spaces, and foreign bodies onto a sheet of film (Figure 15.7). The "shadow graph" images obtained in this manner are the results of the variations in the intensity of the transmitted x-ray beam after it has passed through tissues and body fluids of different densities.

This technique has the advantages of offering high-resolution, high-contrast images with relatively small patient exposure and a permanent record of the image. On the other hand, its disadvantages include significant geometric distortion, inability to discern depth information, and incapability of providing real-time imagery. As a result, conventional radiography is the imaging method of choice for such tasks as dental, chest, and bone imagery. Since bone strongly absorbs x-rays, fractures are readily discernible by the standard radiographic technique. When this procedure is used to project three-dimensional objects into a two-dimensional plane, however, difficulties are encountered. Structures represented on the film overlap, and it becomes difficult to distinguish between tissues that are similar in density. For this reason, conventional x-ray techniques are unable to obtain distinguishable/interpretable images of the brain, which consists primarily of soft tissue. In an effort to overcome this deficiency, attempts have been made to obtain shadow graphs from a number of different angles in which the internal organs appear in different relationships to one another and to introduce a medium (such as air or iodine solutions) that is either translucent or opaque to x-rays. However, these efforts are usually time-consuming, sometimes difficult, sometimes dangerous, and often just not accurate enough.

In the early 1920s, another x-ray technique was developed for visualizing three-dimensional structures. With this technique, known as plane tomography, the imaging of specific planes or cross sections within the body became possible. In plane tomography, the x-ray source is moved in one direction, while the photographic film (which is placed on the other side of the body and picks up the x-rays) is simultaneously moved in the other direction (Figure 15.7b). The result of this procedure is that while the x-rays travel continuously, changing paths through the body, each ray passes through the same point on the plane or cross section of interest throughout the exposure. Consequently, structures in the desired plane are brought sharply into focus and are displayed on film, whereas structures in all the other planes are obscured and show up only as a blur. Such an approach is clearly better than conventional methods in revealing the position and details of various structures and in providing three-dimensional information by such a two-dimensional presentation. There are, however, limitations in its use. First, it does not really localize a single plane, since there is some error in the depth perception obtained. Second, large contrasts in radiodensity are usually required in order to obtain high-quality images that are easy to interpret. In addition, x-ray doses for

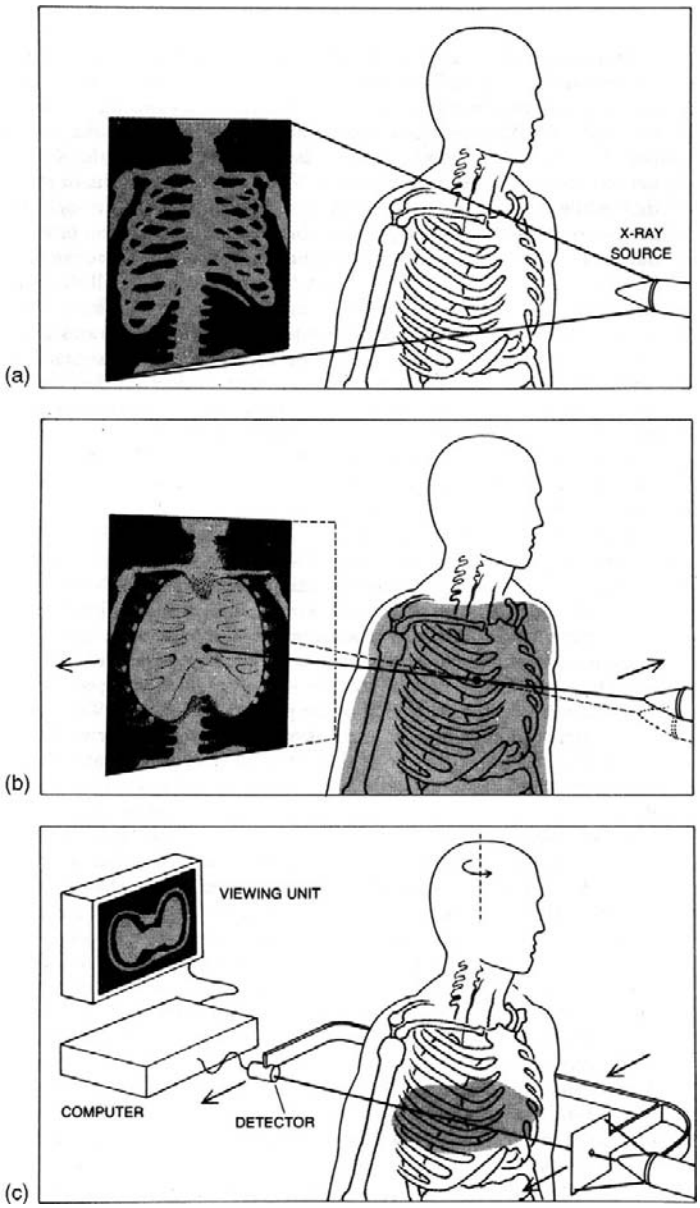


FIGURE 15.7 Overview of the three x-ray techniques used to obtain "medical images." (a) A conventional x-ray picture is made by having the x-rays diverge from a source, pass through the body, and then fall on a sheet of photographic film. (b) A tomogram is made by having the x-ray source move in one direction during the exposure and the film in the other direction. In the projected image, only one plane in the body remains stationary with respect to the moving film. In the picture, all other planes in the body are blurred. (c) In computerized tomography, a reconstruction from projections is made by mounting the x-ray source and an x-ray detector on a yoke and moving them past the body. The yoke is also rotated through a series of angles around the body. Data recorded by the detector are processed by a special computer algorithm or program. The computer generates a picture on a cathode-ray screen.

tomography are higher than routine radiographs, and because the exposures are longer, patient motion may degrade the image content.

Computerized tomography represents a completely different approach. Consisting primarily of a scanning and detection system, a computer, and a display medium, it combines image-reconstruction techniques with x-ray absorption measurements in such a way as to facilitate the display of any internal organ in two-dimensional axial slices or by reconstruction in the Z axis in three dimensions. The starting point is quite similar to that used in conventional radiography. A collimated beam of x-rays is directed through the section of body being scanned to a detector that is located on the other side of the patient (Figure 15.7c). With a narrowly collimated source and detector system, it is possible to send a narrow beam of x-rays to a specific detection site. Some of the energy of the x-rays is absorbed, while the remainder continues to the detector and is measured. In computerized tomography, the detector system usually consists of a crystal (such as cesium iodide or cadmium tungstate) that has the ability to scintillate or emit light photons when bombarded with x-rays. The intensity of these light photons or "bundles of energy" is in turn measured by photodetectors and provides a measure of the energy absorbed (or transmitted) by the medium that is penetrated by the x-ray beam.

Since the x-ray source and detector system are usually mounted on a frame or "scanning gantry," they can be moved together across and around the object being visualized. In early designs, for example, x-ray absorption measurements were made and recorded at each rotational position traversed by the source and detector system. The result was the generation of an absorption profile for that angular position. To obtain another absorption profile, the scanning gantry holding the x-ray source and detector was then rotated through a small angle and an additional set of absorption or transmission measurements was recorded. Each x-ray profile or projection obtained in this fashion is basically one-dimensional. It is as wide as the body but only as thick as the cross section.

The exact number of these equally spaced positions determines the dimensions to be represented by the picture elements that constitute the display. For example, in order to generate a 160 by 160 picture matrix, absorption measurements from 160 equally spaced positions in each translation are required. It will be recalled, however, that each one-dimensional array constitutes one x-ray profile or projection. To obtain the next profile, the scanning unit is rotated a certain number of degrees around the patient, and 160 more linear readings are taken at this new position. This process is repeated again and again until the unit has been rotated a full 180°. When all the projections have been collected, 160×180, or 28,800, individual x-ray intensity measurements are available to form a reconstruction of a cross section of the patient's head or body.

At this point the advantages of the computer become evident. Each of the measurements obtained by the preceding procedure enters the resident computer and is stored in memory. Once all the absorption data have been obtained and located in the computer's memory, the software packages developed to analyze the data by means of image-reconstruction algorithms are called into action. These image reconstruction techniques, which are based on known mathematical constructs developed for astronomy, were not used routinely until the advent of the computer because of the number of computations required for each reconstruction. Modern computer technology has made it possible to fully exploit these reconstruction techniques.

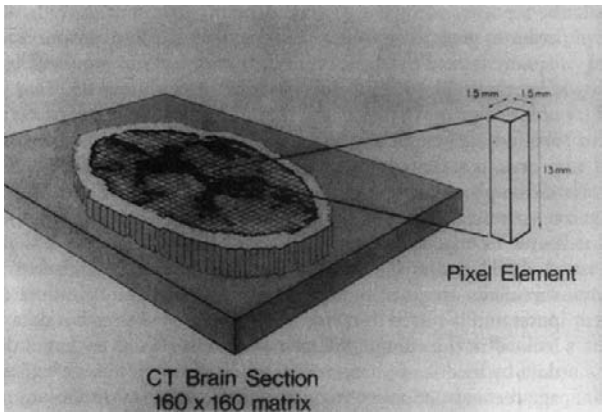


FIGURE 15.8 Sketch of a cross-sectional image available using computerized tomography. Each image is divided into discrete three-dimensional sections of tissue referred to as voxels (or volume elements). On the computer monitor, the slice is viewed in two dimensions as pixels (or picture elements)—in this case 1.5 × 1.5 mm.

To develop an image from the stored values of x-ray absorption, the computer initially establishes a grid consisting of a number of small squares for the cross section of interest, depending on the size of the desired display. The result of this process is something like the mesh of strings in a tennis racket. Since the cross section of the body has thickness, each of these squares represents a volume of tissue, a rectangular solid whose length is determined by the slice thickness and whose width is determined by the size of the matrix. Such a three-dimensional block of tissue is referred to as a "voxel" (or "volume element") and is on the display in two dimensions as a "pixel" (or "picture element") (Figure 15.8).

During the scanning process, each voxel is irradiated by a narrow beam of x-rays up to 180 times. Thus, the absorption caused by that voxel contributes to up to 180 absorption measurements, each measurement part of a different projection. Since each voxel affects a unique set of absorption measurements to which it has contributed, the computer calculates the total absorption due to that voxel. Using the total absorption and the dimension of the voxel, the average absorption coefficient of the tissues in that voxel is determined precisely and displayed in a corresponding pixel as a shade of grey.

EXAMPLE PROBLEM 15.8

Demonstrate how x-rays are attenuated as they pass through a cross-sectional area of the patient's body in multiple rays.

Solution

The cross section of interest can be considered to be made up of a set of blocks of material (Figure 15.9). Each block has an attenuating effect upon the passage of the x-ray energy or photons, absorbing some of the incident energy passing through it. In the line of blocks illustrated

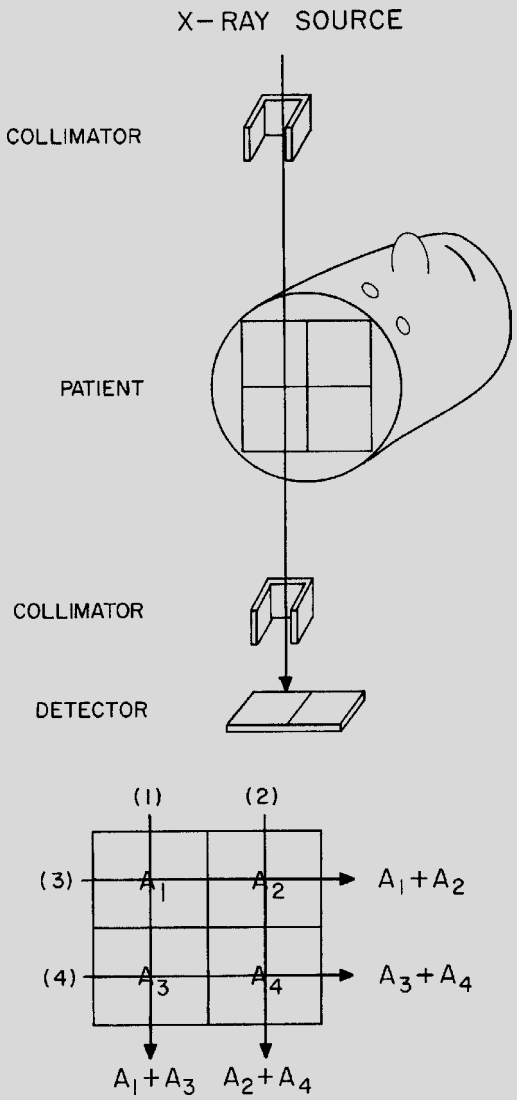


FIGURE 15.9 The method by which the absorption coefficient values for each picture element are obtained.

in Figure 15.9, the first block absorbs a fraction A_1 of the incident photons, the second a fraction A_2, and so on, so that the nth block absorbs a fraction A_n. The total fraction "A" absorbed through all the blocks is the product of all the fractions, while the logarithm of this total absorption fraction is defined as the measured absorption. Consider Figure 15.9 once again. The set of absorption measurements for an object that comprises only four blocks would be as follows:

Continued

$$\text{in position 1: } A(1) = A_1^* A_3 \tag{1}$$

$$\text{in position 2: } A(2) = A_2^* A_4 \tag{2}$$

$$\text{in position 3: } A(3) = A_1^* A_2 \tag{3}$$

$$\text{in position 4: } A(4) = A_3^* A_4 \tag{4}$$

In practice, only the measured absorption factors A(1), A(2), A(3), and A(4) would be known. The problem is, therefore, to compute A_1, A_2, A_3, and A_4 from the measured absorption values. The fact that the computation is possible can be seen from Eqs. (1) to (4). There are four simultaneous equations and four unknowns, so a solution can be found. To reconstruct a cross section containing n rows of blocks and n columns, it is necessary to make at least n individual absorption measurements from at least n directions. For example, a display consisting of 320 $\times$ 320 picture elements requires a minimum of 320 $\times$ 320, or 102,400, independent absorption measurements.

It was mentioned earlier that these absorption measurements are taken in the form of profiles. Imagine a plane parallel to the x-ray beam as defining the required slice. An absorption profile is created if the absorption of the emergent beam along a line perpendicular to the x-ray beam is plotted. This profile represents the total absorption along each of the x-ray beams. In general, the more profiles that are obtained, the better the contrast resolution of the resulting image.

From these individual measurements, a single two-dimensional plane can be reconstructed, and by simply (i.e., for the computer) stacking the appropriate sequence of such planes, it is possible to reconstruct a full three-dimensional picture. Image reconstruction of a three-dimensional object is therefore based primarily on a process of obtaining a cross section or two-dimensional image from many one-dimensional projections. The earliest method used to accomplish this was defined simply as "back projection," which means that each of the measurement profiles was projected back over the area from which it was taken. Unfortunately, this rather simple approach was not totally successful because of blurring. To overcome this, iterative methods were introduced that successively modified the profile being back-projected until a satisfactory picture was obtained. Iteration is a good method, but it is rather slow, requiring several steps to modify the original profiles into a set of profiles that can be projected back to provide an unblurred picture of the original image. To speed up the process, mathematic techniques involving convolution or filtering were introduced that permit the original profile to be modified directly into the final one.

Whichever method of reconstruction is used, the final result of the computation is the same. In each case, a file, usually known as the picture file, is created in the computer memory. The picture file contains an absorption coefficient or density reading for each element of the final picture (e.g., 25,600 for 160 $\times$ 160 pixels or over 100,000 for 320 $\times$ 320 pixels). The resultant absorption coefficients for each element of the image calculated in this manner can then be displayed as gray tones or color scales on a visual display. Most CT systems project an "image" onto the screen of a computer display terminal. Each element or "pixel" of the picture file has a value that represents the density (or more precisely the relative absorption coefficient) of a volume in the cross section of the body being examined.

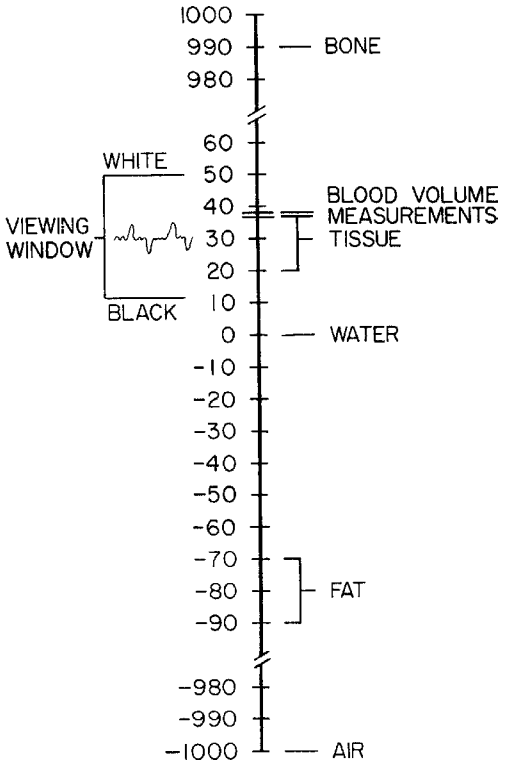


FIGURE 15.10 Hounsfield scale for absorption coefficients.

The scale developed by Hounsfield that is shown in Figure 15.10 demonstrates the values of absorption coefficients that range from air (–1,000) at the bottom of the scale to bone at the top. This original scale covers some 1,000 levels of absorption on either side of water, which is indicated as zero at the center. This is chosen for convenience, since the absorption of water is close to that of tissue due to the high percentage of water that is found in tissue. In using this scale, the number 1,000 is added to the absorption coefficient of each picture element. In the process, the absorption value of air becomes zero, while that of water becomes 1,000. Having assigned each pixel to a particular value in this scale, the system is ready to display each element and create the reconstructed image.

EXAMPLE PROBLEM 15.9

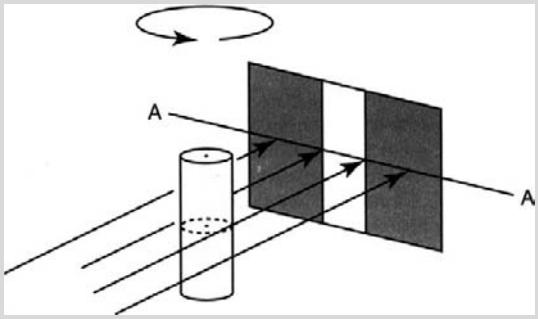
Illustrate the process of back propagation.

Solution

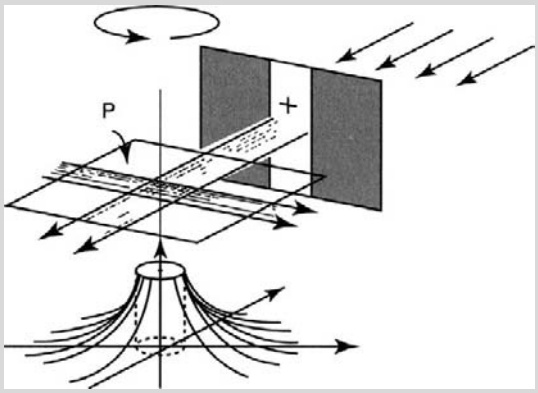
The process can be illustrated by means of a simple analog. In the figure below, a parallel beam of x-rays is directed past and through a cylinder-absorbing substance. As a result, a shadow of the

Continued

cylinder is cast on the x-ray film. The density of the exposed and developed film along the line AA can be regarded as a projection of the object. If a series of such radiographs are taken at equally spaced angles around the cylinder, these radiographs then constitute the set of projections from which the cross section has to be reconstructed.



An approximate reconstruction can be reproduced by directing parallel beams of light through all the radiographs in turn from the position in which they were taken, as shown below.



The correct cross section can be reconstructed by back-projecting the original shadow, as shown above, and subtracting the result of back-projecting two beams placed on either side of the original shadow as shown in the following. Mathematically, this is the equivalent of taking each transmission value in the projection and subtracting from it a quantity proportional to adjacent values. This process is called convolution and is actually used to modify projections.

In the display provided by any medical imaging device, factors affecting the resolution or accuracy of the image must be recognized and modified. Since the information provided by a CT scanner in essence reflects the distribution of x-ray absorption in a cross section of a patient, the smallest change in x-ray absorption that can be detected and projected in the image is defined as the density resolution. In addition, the smallest distance apart that two objects can be placed and still be seen as separate entities is defined as the spatial

resolution. The density resolution reflects the sensitivity of the method to tissue change, while spatial resolution indicates the fineness of detail possible in the image produced by the CT scanner. These two characteristics of resolution are related to each other and to the radiation dose received by the patient. The clarity of the picture and thus the accuracy with which one can measure absorption values are often impaired by a low signal-to-noise ratio that is caused by a reduction in the number of photons arriving at the detectors after penetrating the body. Fewer available photons, in turn, result in a reduction in the sharpness of the image, since the resultant differences between the absorption coefficients of adjacent pixels are reduced. This is a situation that must be accepted as long as low-radiation doses are used. Despite this limitation, the reconstructed image can be enhanced by "display" units that constitute another major component of the CT system.

Display units usually have the capacity to display a given number (16, 32, etc.) of different colors or gray levels at one time. These different colors (or gray levels) can be used to represent a specific range of absorption coefficients. In this way, the entire range of absorption values in the picture can be represented by the entire range of available gray tones or colors (or both). To sharpen the resolution or detail of the display, the range of gray tones between black and white can be restricted to a very small part of the scale. This process of establishing a display "window" that can be raised or lowered actually enhances the imaging of specific organs. For example, if an image of the tissue of the heart is desired, this window is raised above water density, but if an image of the detail in the lung is desired, it is lowered. The overall sensitivity can be further increased by reducing the window width. This permits ever-greater differentiation between absorption coefficients of specific organs and the surrounding tissue.

Various other techniques are available to make the picture more easily interpretable. For instance, a portion of the picture can be enlarged, or certain absorption values can be highlighted or made brighter. These methods do not add to the original information but enhance the image made possible by the data available in the computer's memory. Since a human observer can see an object only if there is sufficient contrast between the object and its surroundings, enhancement modifies the subjective features of an image to increase its impact on the observer, making it easier to locate and precisely measure obscure details.

There are a number of distinct advantages of CT scanning over conventional x-ray techniques. For example, computerized tomography provides three-dimensional information concerning the internal structure of the body by presenting it in the form of a series of slices (Figure 15.11). Because of its contrast sensitivity, it can show small differences in soft tissue clearly, which conventional radiographs cannot do. It accurately measures the x-ray absorption of various tissues, enabling the nature of these tissues to be studied, and yet the absorbed dose of x-rays given to the patient by CT scanning is often less than the absorbed dose by a conventional x-ray technique such as tomography.

In comparing the CT approach with conventional tomography, which also images a slice through the body by blurring the image from the material on either side of the slice, it becomes apparent that in the tomographic approach, only a thin plane of the beam produces useful information as the slice to be viewed. The remainder of the beam, which is by far the larger component, passes through material on either side of the slice, collecting unwanted information that produces artifacts in the picture. On the other hand, the x-ray beam in computerized tomography passes along the full length of the plane of the slice, thereby permitting measurements to be taken that are 100 percent relevant to that slice

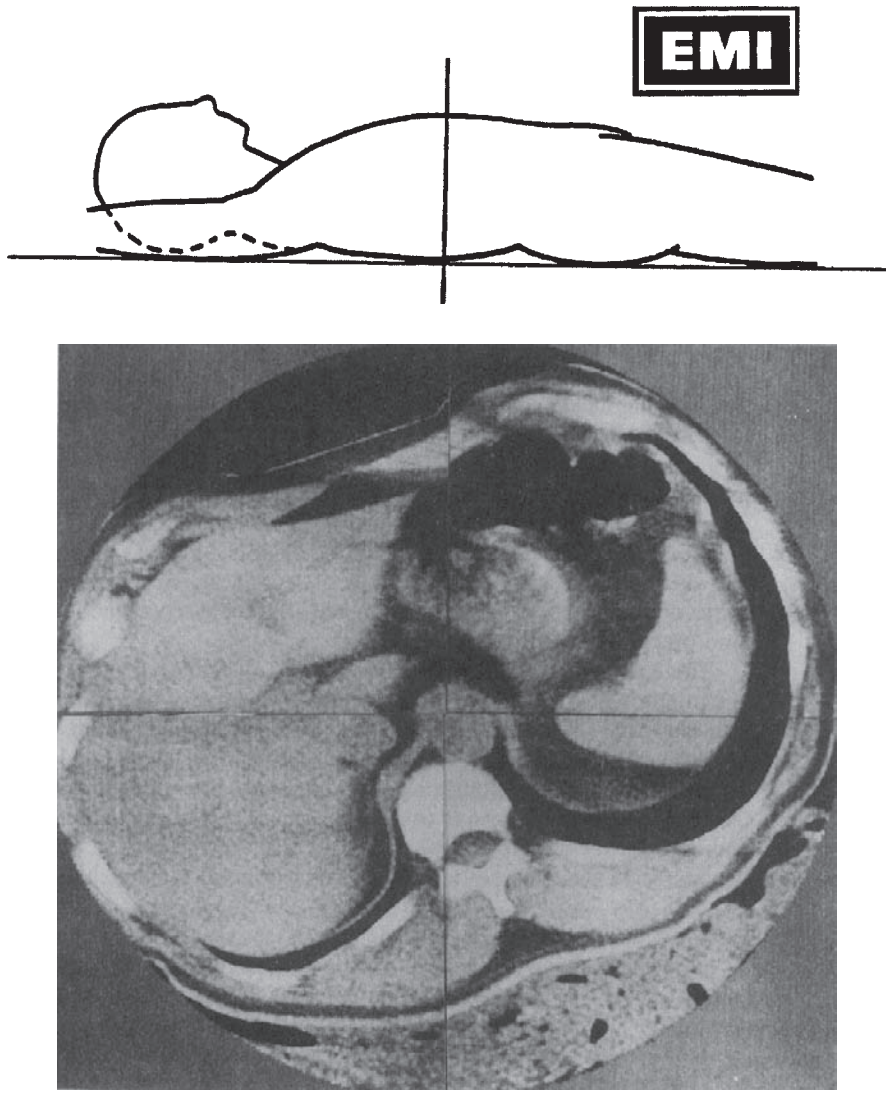


FIGURE 15.11 Photograph of a CT slice of the body.

and that slice alone. These measurements are not affected by the materials lying on either side of the section. As a result, the entire information potential of the x-ray beam is used to the fullest, and the image is more clearly defined.

The medical applications of such detailed pictures are considerable. Hard and soft tumors, cysts, blood clots, injured or dead tissue, and other abnormal morphological conditions can be easily observed. The ability to distinguish between various types of soft tissues means that many diagnostic procedures can be simplified, thus eliminating the need for a number of methods that require hospitalization, are costly, and involve a potential risk to

the patient due to their invasive nature. Since computerized tomography can accurately locate the areas of the body to be radiated and can monitor the actual progress of the treatment, it is effective not only in diagnosis but in the field of radiotherapy as well.

15.4.2 CT Technology

Since its introduction, computerized tomography has undergone continuous refinement and development. In the process, CT scanning has developed into a technology of significant importance throughout the world. The manufacture of quality CT systems requires a strong base in x-ray technology, physics, crystallography, electronics, and data processing.

The evolution of CT scanners has resulted from improvements in x-ray tube performance; scan times; image-reconstruction time; reduction and, in some cases, elimination of image artifacts; and reduction in the amount of radiation exposure to patients. Achievements in these areas have focused attention on two key issues. The first is image quality and its relationship to diagnostic value. The second is scanner efficiency and its relationship to patient care and the economic impact it entails. In order to attain a better understanding of the operation and utility of present CT systems, it is important to understand some of the evolutionary steps.

CT scanners are usually integrated units consisting of three major elements:

1. The *scanning gantry*, which takes the readings in a suitable form and quantity for a picture to be reconstructed.
2. The *data-handling unit*, which converts these readings into intelligible picture information, displays this picture information in a visual format, and provides various manipulative aids to enhance the image and thereby assist the physician in forming a diagnosis.
3. A *storage facility*, which enables the information to be examined or reexamined at any time after the actual scan.

The Scanning Gantry

The objective of the scanning system is to obtain enough information to reconstruct an image of the cross section of interest. All the information obtained must be relevant and accurate, and there must be enough independent readings to reconstruct a picture with sufficient spatial resolution and density discrimination to permit an accurate diagnosis. The operation of the scanning system is therefore extremely important.

CT scanners have undergone several major gantry design changes (Figure 15.12). The earliest generation of gantries used a system known as "traverse and index." In this system, the tube and detector were mounted on a frame, and a single beam of x-rays traversed the slice linearly, providing absorption measurements along one profile. At the end of the traverse, the frame indexed through 1° and the traverse was repeated. This procedure continued until 180 single traverses were made and 180 profiles were measured. Using the "traverse and index" approach, the entire scanning procedure took approximately 4–5 minutes. The images it produced had excellent picture quality and thus high diagnostic utility. However, this approach had several disadvantages. First, it was relatively slow, resulting in relatively low patient output. Second, streak artifacts were common, and although they normally were not

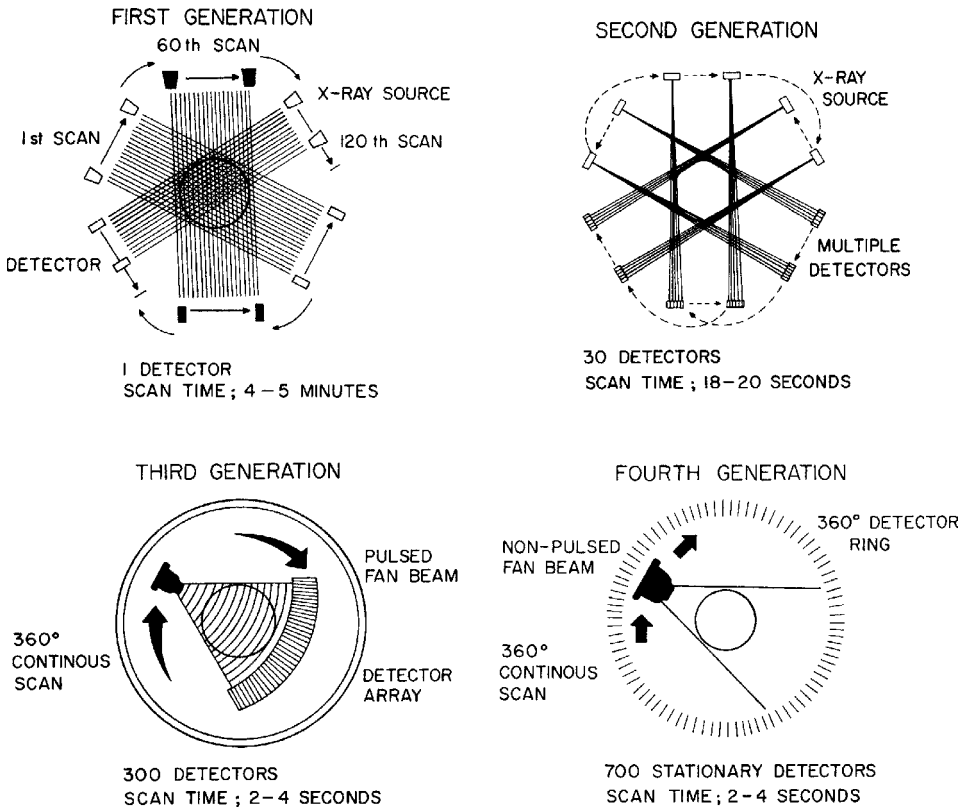


FIGURE 15.12 Four generations of scanning gantry designs. With modern slip ring technology, third- or fourth-generation geometry allows spiral volumetric scanning using slice widths from 1 to 10 mm and pixel matrixes to 10,242. Typically, a 50-cm volume can be imaged with a single breath hold.

diagnostically confusing, the image quality did suffer. Finally, and perhaps the biggest drawback, was that since the patient had to remain absolutely immobile during the scan, its use was restricted solely to brain scans.

In an attempt to solve these problems, a second generation of gantries evolved. The scanning procedures in this approach included a translation step as in the first generation, but it incremented the scanning gantry 180° around the patient in 10° steps. In addition, in an effort to obtain more profiles with each traverse, a larger fan beam was used. By using a 10° fan beam, it was possible to take ten profiles—at 1° intervals—with each traverse. By indexing through 10° before taking the next set of profiles, it was possible to obtain a full set of 180 profiles by making only 18 traverses. At the rate of approximately one second for each traverse, scanning systems of this type could operate in the range of 18–20 seconds. Even in this extremely reduced scan time, the scanning and detection system was able to obtain over 300,000 precise absorption measurements during the complete (i.e., over 180°) scan. These values were then used to construct a picture of the slice under investigation. With this new generation of CT scanners, it became possible to obtain cross sections of

any part of the body. However, it still was not fast enough to eliminate the streak artifacts, nor did it seem possible to further reduce scan time with this approach. In essence, the first and second generations had reached their limits. Scanning times could not be reduced further and still achieve an acceptable image quality. Nevertheless, this geometry had the potential to achieve the highest spatial and contrast resolutions of all the gantry designs.

A major redesign ensued, and many of the scanning system components were improved. In the process, a third generation of gantries evolved that consisted of a pure rotational system. In these systems, the source-detector unit had a pulsing, highly collimated wide-angle (typically, 20°–50°) fan beam, and a multiple detector array was rotated 360° around the patient. This single 360° smoothly rotating movement produced scan times as low as 500 ms, increased appreciably the reliability of the data because they were taken twice and increased the quality of the reconstructed image. A by-product of this design was that the pulsing x-ray source could be synchronized with physiologic parameters, enabling rhythmic structures such as the heart to be more accurately imaged.

The main advantages of this type of system were its simplicity and its speed, but it had two major disadvantages. First, its fixed geometric system, with a fan beam usually established for the largest patient, was inefficient for smaller objects and in particular for head scanning. Second, it was particularly prone to circular artifacts. These "ring" artifacts were most obvious in early pictures from this type of machine. The magnitude of the difficulty in achieving stability can be appreciated when it is realized that an error of only one part in 10,000 can lead to an artifact that is diagnostically confusing.

Consequently, a fourth-generation gantry system consisting of a continuous ring (360°) of fixed detectors (usually numbering between 600 and 4,800) was developed. In this system, the x-ray source rotated as before, and the transmitted x-rays were detected by the stationary detector ring. The fan beam is increased slightly so the detectors on the leading and trailing edge of the fan can be continuously monitored and the data adjusted in case of shifts in detector performance. As a result, the image obtained using this approach is more reliable, and the ring artifact is usually eliminated. Fourth-generation scan times of approximately 500 ms or less with reconstruction times of 30 seconds or less are now commonplace.

The development of high-voltage slip rings has enabled third- and fourth-generation scanners to continuously rotate about the patient. By moving the patient couch at a uniform speed in or out of the scan plan, the x-ray beam will describe a helix or spiral path through the patient. Slice widths of 1–10 mm and pixel matrixes up to $1{,}024^2$ may be employed, and typically a 50-cm-long volume can be imaged in a single breath hold. Because of improvements in x-ray tube technology, spatial and contrast resolution are maintained. New display techniques allow for real-time 3D manipulations and virtual reality fly-through of tubular structures such as the aorta or large intestine.

Another gantry design, which might be called a fifth generation, utilizes a fixed detector array arranged in a 210° arc positioned above the horizontal plane of the patient and four target rings of 210° arc positioned below the horizontal plane and opposite from the detector array. There is no mechanical motion in the gantry except the movement of the couch into and out of the imaging plane (Figure 15.13).

X-rays are produced when a magnetic field causes a high-voltage focused election beam to sweep along the path of the anode target. Up to four 1.5-mm simultaneous slices are created in 50 or 100 ms and displayed in up to a 512^2 pixel matrix. This ultrafast CT can

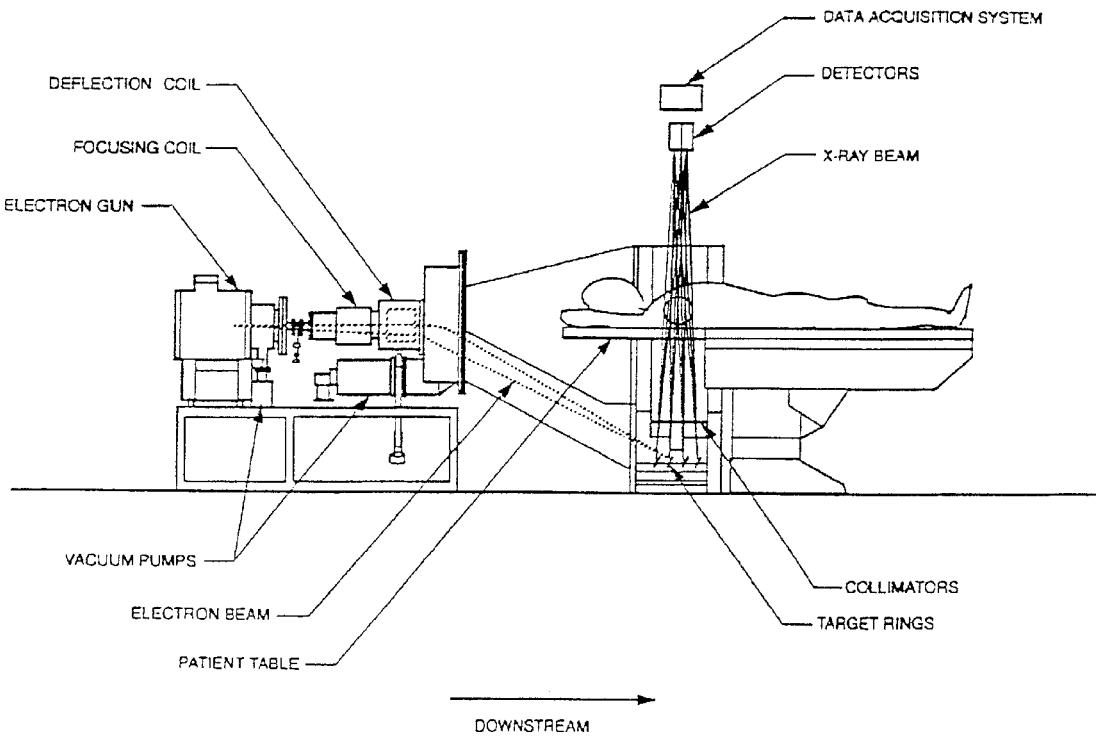


FIGURE 15.13 An ultrafast CT scanner. In this device, magnetic coils focus and steer the electron beam through the evacuated drift tube, where they strike the target rings. The presence of four target rings and multiple x-ray collimators allows four unique x-ray beams to pass through the patient. These beams are then interrupted by two contiguous detector arrays and form eight unique image slices. *Courtesy of Imatrol.*

volumetrically examine the heart and evaluate the presence and severity of coronary calcific atherosclerosis. Although the ultrafast scanner can almost stop biologic motion, its contrast and spatial resolution are presently less than third- or fourth-generation mechanical scanners, so its general applications are somewhat less.

X-Ray Tubes and Detectors

It must be emphasized that the whole purpose of the scanning procedure is to take thousands of accurate absorption measurements through the body. Taken at all angles through the cross section of interest, these measurements provide an enormous amount of information about the composition of the section of the body being scanned. Since these readings are taken by counting photons, the more photons counted, the better will be the quality of the information. With these factors in mind, it is essential that the radiation dose be sufficient to obtain good-quality pictures. Since the total radiation dose is a function of the maximum power of the tube and its operating time, faster scanners have to use a significantly higher-powered tube. These aspects of dose efficiency are very important in tube and detector design.

Currently, CT systems that operate in the region of 20 s use fixed-anode, oil-cooled tubes that run continuously during the scan time, while those that operate in much less than 20 s have a rotating-anode, air- or water-cooled tube that is often used in a pulse mode. The fixed-anode tubes, being oil cooled, can be run continuously, and the sequence of measurements is obtained by electronic gating of the detectors. Faster scanners need tubes of substantially higher power, thereby requiring the rotating-anode type. There is a tendency to use these in a pulse mode largely for convenience, since it simplifies the gating of the detectors. One of the problems encountered with fast machines, however, is the difficulty in providing enough doses within the scan time. For high-quality pictures, many of these scanners run at much slower rates than their maximum possible speed.

Of equal importance in the accurate detection and measurement of x-ray absorption is the detector itself. Theoretically, detectors should have the highest possible x-ray photon capture and conversion efficiency to minimize the radiation dose to the patient. Clearly, a photon that is not detected does not contribute to the generation of an image. The detection system of a CT scanner that is able to capture and convert a larger number of photons with very little noise will produce an accurate, high-quality image when aligned with suitable reconstruction algorithms. Consequently, the selection of a detector always represents a trade-off between image quality and low-radiation dose.

Three types of detectors have been used: scintillation crystals (such as sodium iodide, calcium fluoride, bismuth germinate, or cadmium tungstate) combined with photomultiplier tubes; gas ionization detectors containing Xenon; and scintillation crystals with integral photodiodes. It will be recalled that the principal factors in judging detector quality are how well the detector captures photons and its subsequent conversion efficiency, which results in optimal dose utilization. Those systems employing scintillation crystals combined with photomultiplier tubes do not lend themselves to the dense packing necessary for maximal photon capture, even though they offer a high conversion efficiency. The less than 50 percent dose utilization inherent in these detectors requires a higher dose to the patient to offset the resultant inferior-quality image. Although gas detectors are relatively inexpensive and compact and lend themselves for use in large arrays, they have a low conversion efficiency and are subject to drift. Consequently, use of Xenon detectors necessitates increasing the irradiation dose to the patient.

The advantage of scintillation crystals with integral photodiodes extends beyond their 98 percent conversion efficiency, since they lend themselves quite readily to a densely packed configuration. The excellent image quality combined with high-dose utilization (i.e., approximately 80 percent of the applied photon energy) and detector stability ensure that solid-state detectors will continue to gain favor in future CT systems.

Data Handling Systems

In conjunction with the operation of the scanning and detection system, the data obtained must be processed rapidly to permit viewing a scan as quickly as possible. Data handling systems incorporating the latest advances in computer technology have also been developed to coincide with changes in the design of the scanning and detector components of CT scanners. For example, with the evolution of the third-generation scanners, the data handling unit had to digitize and store increasing amounts of data consisting of the following:

- Positional information, such as how far the scanning frame was along its scan
- Reference information from the reference detector, which monitored the x-ray output
- Calibration information, which was obtained at the end of each scan
- The bulk of the readings—the actual absorption information obtained from the detectors

Electronic interfaces are now available that convert this information to digital form and systematically store it in the memory of the resident computer. The next step is the actual process of image reconstruction (Figure 15.14).

It will be recalled that the first stage of this computation process is to analyze all of the raw data and convert them into a set of profiles, normally 180 or more, and then convert these profiles into information that can be displayed as a picture and used for diagnosis. This is the heart of the operation of a CT scanner—that element that makes CT totally different from conventional x-ray techniques and most other imaging techniques. The algorithms or computer programs for reconstructing tissue x-ray absorption properties from a series of x-ray absorption profiles fall into the categories of simple back-projection (sometimes called summation), Fourier transforms, integral equations, or series expansions. The choice of an algorithm for a CT scanner depends on the algorithm's speed and accuracy.

Mathematical algorithms for taking the attenuation projection data and reconstructing an image can be classified into two categories: iterative and analytic. The iterative techniques (also known as the algebraic reconstruction technique [ART]), such as the one used by Hounsfield in the first-generation scanner, require an initial guess of the two-dimensional pattern of x-ray absorption. The attenuation projection data predicted by this guess are then calculated and the results compared with the measured data. The difference between the measured data and predicted values is used in an iterative manner so the initial guess is modified and that difference goes to zero. In general, a large number of iterations are required for convergence, with the process usually halted when the difference between the

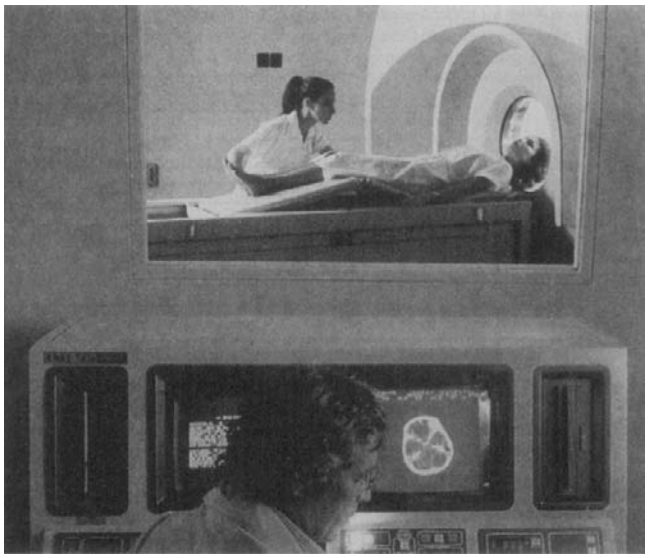


FIGURE 15.14 A modern display terminal for a CT scanner.

calculated and the measured data is below a specified error limit. A number of different versions of the ART were developed and used with first- and second-generation CAT scanners. Later-generation scanners used analytic reconstruction techniques, since the iterative methods were computationally slow and had convergence problems in the presence of noise.

Analytic techniques include the Fourier transform, back-projection, filtered back-projection, and convolution back-projection approaches. All of the analytic methods differ from the iterative methods in that the image is reconstructed directly from the attenuation projection data. Analytic techniques use the central section theorem and the two-dimensional Fourier transform, which is shown in Figure 15.15. Given an image $f(x, y)$, a single projection is taken along the x direction, forming a projection $g(y)$ described by

$$g(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

This projection represents an array of line integrals as shown in Figure 15.15. The two-dimensional Fourier transform of $f(x, y)$ is given by

$$F(u,v) = \int\int_{-\infty}^{\infty} f(x,y)\exp[-j2\pi(ux+vy)]dxdy$$

In the Fourier domain, along the line $u = 0$, this transform becomes

$$F(0,v) = \int\int_{-\infty}^{\infty} f(x,y)\exp(-j2\pi vy)dxdy$$

which can be rewritten as

$$F(0,v) = \int\int_{-\infty}^{\infty} f(x,y)dx/\exp(-j2\pi vy)dy$$

or

$$F(0,v) = F_1[g(y)]_1$$

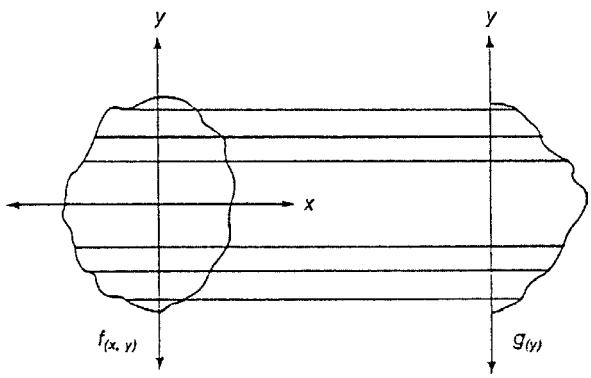


FIGURE 15.15 A projection via the central-sectioning theorem.

where F_1 [ ] represents a one-dimensional Fourier transform. It can be shown that the transform of each projection forms a radial line in $F(u, v)$, and therefore $F(u, v)$ can be determined by taking projections at many angles and taking these transforms. When $F(u, v)$ is completely described, the reconstructed image can be found by taking the inverse Fourier transform to obtain $f(x, y)$.

The image reconstruction method used by most modern CAT scanners is the filtered back-projection reconstruction method. In this method, attenuation projection data for a given ray or scan angle are convolved with a spatial filter function either in the Fourier domain or by direct spatial convolution. The filtered data are then back-projected along the same line, using the measured value for each point along the line, as shown in Figure 15.16. The total back-projected image is made by summing the contributions from all scan angles. Depending on the filter technique, the image is obtained directly after back-projection summation or via the inverse Fourier transform of the back-projected image.

It should be obvious by now that compared to early head and body scanners, current models process information faster, use energy more efficiently, are more sensitive to differences in tissue density, have finer spatial resolution, and are less susceptible to artifacts. CT scanners may be compared with one another by considering the following ten factors:

1. *Gantry design*, which affects scan speed, patient processing time, and cost-effectiveness
2. *Aperture size*, which determines the maximum size of the patient along with the weight-carrying capacity of the couch
3. *The type of x-ray source*, which affects the patient radiation dose and the overall life of the scanning device
4. *X-ray fan beam angle and scan field*, which affects resolution
5. *The slice thickness*, as well as the number of pulses and the angular rotation of the source, which are important in determining resolution
6. *The number and types of detectors*, which are critical parameters in image quality
7. *The type of minicomputer employed*, which is important in assessing system capability and flexibility
8. *The type of data-handling routines* available with the system, which are important user and reliability considerations
9. *The storage capacity* of the system, which is important in ascertaining the accessibility of the stored data
10. *Upgradeability and connectivity*—that is, they should be capable of modular upgradeability and should communicate to any available network

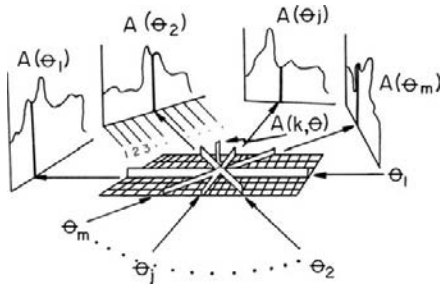


FIGURE 15.16 The results obtained with a simple back-projection algorithm.

In reviewing the available CT systems, it is important to compare them not only in technical terms but against what one considers to be an ideal scanner. The ultimate objective of computerized tomography is to provide accurate diagnostic information that significantly improves patient care. In the ideal case, a CT scanner would provide an unambiguous diagnosis, thereby eliminating the necessity for further tests. Since financial considerations are also important and CT scanners require a high capital outlay and have considerable maintenance costs, the ideal system should process a large patient load quickly, providing high "patient throughput." A good scanner, therefore, must satisfy two major criteria: it must provide good diagnostic information and it must permit high patient throughput.

15.5 EXERCISES

1. Represent the decay process discussed in Example Problem 14.2 in symbolic form.
2. Compute the energy liberated when $^{238}_{92}$U (amu = 238.050786) decays to $^{234}_{90}$Th (amu = 232.038054) via α emission.
3. Carbon 14, $^{14}_{6}$C is a radioactive isotope of carbon that has a half-life of 5,730 years. If an initial sample contained 1,000 ^{14}C nuclei, how many would still be around after 22,920 years?
4. A 50-g sample of carbon is taken from the pelvis bone of a skeleton and is found to have a ^{14}C decay rate of 200 decays/min. It is known that carbon from a living organism has a decay rate of 15 decays/min X g and that ^{14}C has a half-life of 5,730 years = 3.01×10^9 min. Find the age of the skeleton.
5. The half-life of a radioactive sample is 30 min. If you start with a sample containing 3×10^{16} nuclei, how many of these nuclei remain after 10 min?
6. Find the energy liberated in the beta decay of $^{14}_{6}$C to $^{14}_{7}$N.
7. How long will it take for a sample of polonium of half-life 140 days to decay to one-tenth its original strength?
8. Suppose that you start with $10^{\sim 3}$ g of a pure radioactive substance and 2 h later determine that only $0.25 \times 10^{\sim 3}$ g of the substance remains. What is the half-life of this substance?
9. The half-life of an isotope of phosphorus is 14 days. If a sample contains 3×10^{16} such nuclei, determine its activity.
10. How many radioactive atoms are present in a sample that has an activity of 0.2 μCi and a half-life of 8.1 days?
11. A freshly prepared sample of a certain radioactive isotope has an activity of 10 mCi. After 4 h, the activity is 8 mCi.
    (a) Find the decay constant and half-life of the isotope.
    (b) How many atoms of the isotope were contained in the freshly prepared sample?
    (c) What is the sample's activity 30 h after it is prepared?
12. Tritium has a half-life of 12.33 years. What percentage of the nuclei in a tritium sample will decay in 5 years?
13. For the following process $^{23}_{10}\mathrm{Ne}_{13} \rightarrow {}^{23}_{11}\mathrm{Ne}_{12} + e + v$, what is the maximum kinetic energy of the emitted electrons?
14. The half-life of ^{235}U is 7.04×10^8 years. A sample of rock that solidified with the earth 4.55×10^9 years ago contains N atoms of ^{235}U. How many ^{235}U atoms did the same rock have when it solidified?

Continued

15. Of the three basic types of radioaction—alpha, beta, and gamma—which have the greatest penetration into tissue? Explain the rationale behind your answer.
16. Discuss the principle of scintillation. How is it detected? How can it be used to generate an image?
17. One method of treating cancer of the thyroid is to insert a small radioactive source directly into the tumor. The radiation emitted by the source can destroy cancerous cells. Why do you suppose $^{131}_{53}$I is used for this treatment?
18. Provide a clinical example of the use of instrumentation employed to detect the rate of "radioactive workout."
19. In a photomultiplier tube, assume that there are seven diodes with potentials 100, 200, 300, 400, 500, 600, and 700 V. The average energy required to free an electron from the dynode surface is 10 eV. For each incident electron released from rest, how many electrons are freed at the first dynode? At the last dynode?
20. Discuss the process of producing a tomograph. Provide figures to illustrate your answer.
21. Describe the operation of the iterative method to produce an image.
22. Provide an illustration of how the net absorption coefficient is obtained for each volume element.

Suggested Readings

E.L. Alpen, Radiation Biophysics, Prentice Hall, New Jersey, 1990.
J.D. Bronzino, Computer applications in patient care, Addison-Wesley, Reading, Mass, 1982.
Z.H. Cho, J.P. Jones, M. Singh, Foundations of Medical Imaging, John Wiley and Sons, Inc., New York, 1993.
B.F. Croft, B.M.N. Tsui, Nuclear Medicine, in: The Biomedical Engineering Handbook, CRC Press, 1995, pp. 1046–1076.
I.A. Cunningham, P.F. Juoy, Computerized Tomography, in: The Biomedical Engineering Handbook, CRC Press, 1995, pp. 990–1005.
G.N. Hounsfield, Computerized transverse axial scanning (tomography) part 1, description of system, Br. J. Radiol. 46 (1973) 1016–1022.
C.H. McCollough, R.L. Morin, The technical design and performance of ultrafast computed tomography, Cardiac Imagery 32 (1994) 521–536.
C.H. McCollough, Acceptance testing of a fifth generation scanning-electron-beam computed tomography scanner, Med. Phys. 19 (1992) 846.
R.A. Serway, J.S. Faughn, College Physics, second ed., Saunders College Publishing, Philadelphia, 1989.
R.E. Shroy Jr., M.S. Van Lipel, M.J. Yaffe, X-rays, in: The Biomedical Engineering Handbook, CRC Press, 1995, pp. 953–989.

CHAPTER

16

Medical Imaging

*Thomas Szabo, PhD**

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Distinguish between the principles of pulse-echo ranging and ultrasound imaging.
- Describe how ultrasound images are formed.
- Explain the fundamentals of acoustic wave propagation, reflection, and refraction.
- Describe the operation and characteristics of a piezoelectric transducer.
- Explain the principle of an acoustic matching layer.
- Discuss the basic types of acoustic scattering.
- Calculate the effects of acoustic absorption.
- Describe the fundamentals of beam formation and focusing.
- Discuss the block diagram of an ultrasound imaging system.
- Explain basic ultrasound Doppler.
- Explain four basic principles of the interaction of magnetic fields and charges.
- Discuss spin states and precession of nuclear dipoles in a magnetic field.

*With contributions from Kirk K. Shung and Steven Wright.

- Explain the Larmor frequency and nuclear magnetic resonance.
- Explain how flip angles affect recovery and relaxation time constants.
- Distinguish between free induction decay and spin echo signals.
- Explain how detected resonance signals are spatially localized.
- List the steps involved in creating a magnetic resonance image.
- Explain what is displayed in a magnetic resonance image.
- Draw and explain the block diagram of an MRI system.
- Explain the significance of k-space in MRI.
- Discuss applications of MRI including fMRI.
- Discuss the principles of magneto-encephalography.
- Compare contrast agents for different imaging modalities.
- Explain image fusion.
- Compare major imaging modalities.

16.1 INTRODUCTION

Of the major diagnostic imaging modalities, ultrasound is the most frequently used, second only to standard plane-view x-rays. Over the years, the cumulative number of ultrasound exams completed is estimated to be in the billions. Unlike x-rays and computed tomography (CT) scanning, ultrasound imaging involves no ionizing radiation and therefore is considered to be noninvasive. Furthermore, it is portable, easy to apply, low in cost, and provides real-time diagnostic information about the mechanical nature and motion of soft tissue and blood flow. The basic principle of ultrasound imaging is the display of pulse-echoes backscattered from tissues.

Magnetic resonance imaging (MRI) also obtains detailed anatomic information without using ionizing radiation. MRI differentiates among types of organs by sensing the spin of their atoms when a person is placed in a large, static, magnetic field. Static cross-sectional images of the body include both bone and soft tissue, and the appearance of the images can vary considerably by the selection of specific parameters. Of the major imaging modalities, MRI is the most abstract and complicated technically. In addition to its precise anatomical capability, it is often used for presurgery planning and for cancer detection. Functional MRI (fMRI) provides images of brain activity in response to various stimuli.

In this chapter, the operation and principles of both ultrasound imaging and MRI will be explained. At the end of the chapter, the main features of all major imaging modalities, including CT and x-ray, will be compared. An understanding of the basic principles of the Fourier transform, introduced in Chapter 10, will enhance comprehension of the imaging concepts introduced in this chapter. More information about the Fourier transform is provided in the next section.

16.1.1 Review of Fourier Transforms

Because Fourier transforms simplify the understanding of the imaging principles of ultrasound imaging and MRI, their relevant properties are reviewed here. One important Fourier transform concept from Chapter 6 is the equivalence of convolution in the time

domain and multiplication in the frequency domain. Another is the use of an impulse response function and a transfer function for describing system and filter responses in the time and frequency, respectively.

An impulse function is a generalized function that has the unusual property that it samples the integrand:

$$\int_{-\infty}^{\infty} \delta(t - t_0) g(t) dt = g(t_0) \tag{16.1}$$

When the transform of the impulse function is taken, the result is an exponential

$$H(\omega) = \int_{-\infty}^{\infty} \delta(t - t_0) e^{-i\omega t} dt = e^{-i\omega t_0} \tag{16.2}$$

which shows that a delay in time is equivalent to a multiplicative exponential delay factor in the frequency domain. When the impulse has no delay or $t_0 = 0$, $H(\omega) = 1.0$, a constant.

The preceding relation can be generalized to the form where I represents the Fourier transform operation,

$$\Im[g(t - b)] = e^{-i2\pi bf} G(\omega) \tag{16.3}$$

A scaling factor can be added to this time shifting/delay theorem to make it even more useful,

$$\Im[g(a(t - b))] = \frac{e^{-i\omega b}}{|a|} G(\omega/a) \tag{16.4a}$$

Note that a similar relation exists for the inverse transform denoted by $\Im^{-1}$,

$$\Im^{-1}[G(a(\omega - b))] = \frac{e^{-ibt}}{|a|} g(t/a) \tag{16.4b}$$

Finally, an important unique property of the impulse function is

$$\delta(at) = \frac{1}{|a|} \delta(t) \tag{16.5}$$

EXAMPLE PROBLEM 16.1

Find the inverse Fourier transform of $R(\omega) = \sin[3(\omega - \omega_0)t_1]$.

Solution

Recognize the basic function G in $R(\omega)$ as

$$G(\omega) = \sin \omega t_1 = \frac{e^{i\omega t_1} - e^{-i\omega t_1}}{2i}$$

Continued

Then from the Fourier transform pair expressed in Eq. (16.2),

$$\Im^{-1}[G(\omega)] = g(t) = (i/2)[\delta(t - t_1) - \delta(t + t_1)]$$

From Eq. (16.4b) with $a = 3$ and $b = \omega_0$, find $r(t)$ as

$$\Im^{-1}[R(\omega)] = (e^{i\omega_0 t}/3)(i/2)\left[\delta\left(\frac{t - t_1}{3}\right) - \delta\left(\frac{t + t_1}{3}\right)\right]$$

and from scaling in the impulse function in Eq. (16.5),

$$r(t) = (e^{i\omega_0 t}) = (i/2)[\delta(t - t_1) - \delta(t + t_1)]$$

16.2 DIAGNOSTIC ULTRASOUND IMAGING

Diagnostic ultrasound imaging is the most widely used form of medical imaging. Its ability to reveal body structures and dynamics and blood flow in real time at low cost in a safe manner (using non-ionizing radiation) has expanded its growth and utility for over fifty years. Improvements in technology and electronic miniaturization have decreased the cost and size of ultrasound systems and made highly portable systems possible. In this section, the history and principles of diagnostic ultrasound are discussed.

16.2.1 Origins of Ultrasound Imaging

Though it has long been known that bats use sound for echolocation, the intentional use of ultrasound (sound with frequencies above our range of hearing) for this purpose began, surprisingly, with the sinking of the *Titanic* in 1913. L. F. Richardson, a British scientist, filed patents within months of the *Titanic* disaster for echolocation of icebergs (and other objects) using sound in either water or air. By the end of World War I, C. Chilowsky and P. Langevin in France invented practical implementations of echolocation with high-powered electronic transmitters and piezoelectric transducers for locating submarines and echo ranging. These principles of echo ranging were applied much later to electromagnetic waves to create RADAR (RAdio Detection And Ranging). The circular sweep of a RADAR echo ranging line is displayed on PPI (plan position indicator) monitors, an early example of pulse-echo imaging. This type of display and technology was in turn adopted by underwater investigators to develop SONAR (SOund Navigation And Ranging).

After experiencing these technologies during World War II, several doctors wanted to apply echolocation principles to the interior of the human body. Fortunately, a device—the supersonic reflectoscope—that was developed for finding defects in solid objects by ultrasound echolocation became available in addition to other surplus wartime equipment. This type of equipment sent out short pulses a few microseconds in length, repeated at longer intervals of a millisecond, as shown in Figure 16.1.

Short pulses were needed to determine the location of tissue boundaries. A typical setup is shown in Figure 16.2, along with a record of echoes displayed as a function of time. Here,

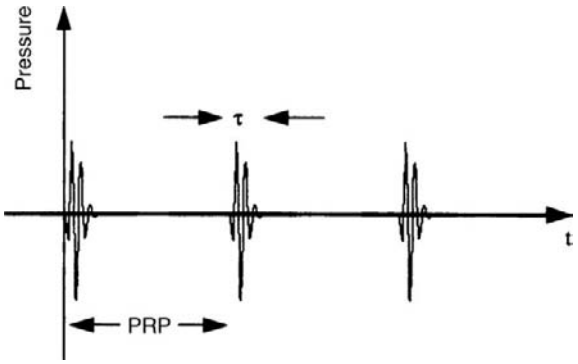


FIGURE 16.1 A plot of a series of transmitted acoustic pulses repeated at intervals called a pulse repetition period (PRP). Typically, the –3 dB width of a pulse is only a few microseconds in order to resolve different tissue interfaces, whereas the PRP is about a millisecond. For the purposes of illustration, the PRP has been compressed to enhance the presentation of pulses. The horizontal, or time, axis represents ambient pressure. Positive values on the vertical axis show compressional acoustic pressure, and negative values show rarefactional pressure.

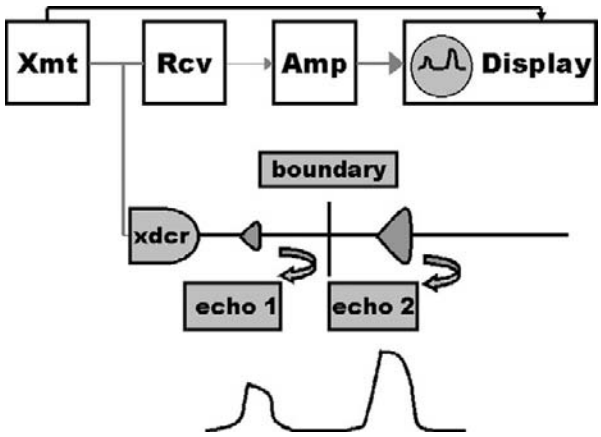


FIGURE 16.2 Basic echo ranging system consisting of a transmitter, a transducer, a receiver, an amplifier, and an oscilloscope display. Pulse echoes are shown below at delays corresponding to depths of reflecting objects above.

the delay, t, to each echo is the round-trip distance to the object, $2z$, divided by the speed of sound in the material, c, or $t = 2z/c$. The echo range instrument consisted of a piezoelectric transducer that converted the electrical pulses from the transmitter to acoustic pulses and reconverted received echoes from targets into electrical signals. These signals were then amplified and displayed as a time record on an oscilloscope. This type of display was known as an A-mode display, with A signifying amplitude.

In 1949, U.S. Naval doctor G. Ludwig reported his measurements of sound speed in parts of the body, and the results showed they had an average value of $c_0 = 1540$ m/s. More recent precise measurements showed that sound speed for most tissues varied by only a few percent, as listed in Table 16.1. This important finding meant that the location of organs

TABLE 16.1 Acoustic Properties of Tissue

Material	C (m/s)	α(dB/MHz^y-cm)	y	ρ (kg/m^3)	Z (MegaRayls)
Air	343			1.21	0.0004
Bone	3,360	3.54	0.9	1,789	6.00
Blood	1,550	0.14	1.21	1,039	1.61
Fat	1,450	0.6	1.0	952	1.38
Honey	2,030			1,420	2.89
Liver	1,570	0.45	1.05	1,051	1.65
Muscle	1,580	0.57	1.0	1,041	1.645
Water @ 20°C	1,482.3	2.17×10^{-3}	2.0	1,000	1.482

could be determined by the simple delay equation with a single value for sound speed, c_0, which is still used today in modern imaging systems. In other words, the geometric accuracy of the placement of organs in an ultrasound image largely depends on the uniformity of the sound speed in the field of view. Ludwig also measured the characteristic acoustic impedance of tissues, $Z = \rho c$, in which ρ and c represent the density and speed of sound of each tissue, respectively. He found that at the boundary of two different tissues, the reflection factor was related to the impedances of the individual tissues, Z_1 and Z_2:

$$RF = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{16.6a}$$

Independently, Dr. D. Howry, another U.S. doctor, found that because the reflection factors were small, ultrasound penetrated through multiple layers of soft tissue with ease (except for regions with gas or bone). In 1956, he was able to make detailed anatomical maps of the body with sound and showed that they corresponded with known locations and sizes of organs and tissues. A graph of reflection factors in dB (decibels),

$$RF_{\mathrm{dB}} = 20\ \log_{10}(RF) \tag{16.6b}$$

all with reference to Z_1 for blood is shown in Figure 16.3. These factors are different enough to provide adequate discrimination, an important factor in differentiating among tissues in imaging.

Dr. J. J. Wild, an English surgeon, and J. Reid, an electrical engineer (now professor emeritus) working in Minnesota with a surplus 15 MHz radar simulator in 1951, recognized the value of ultrasound for diagnosis and attempted to use it to detect cancer in the stomach. They also made calculations on pulse-echo waveforms to infer properties of healthy and diseased tissue and began a new field now called tissue characterization. In the process, Wild and Reid developed near real-time ultrasound imaging systems and ways of placing transducers (that transmit sound and receive pulse echoes and convert them to electrical signals) directly on the skin.

In the 1950s, a number of groups around the world became interested in the diagnostic possibilities of ultrasound. S. Satomura, Y. Nimura, and T. Yoshida in Japan detected blood

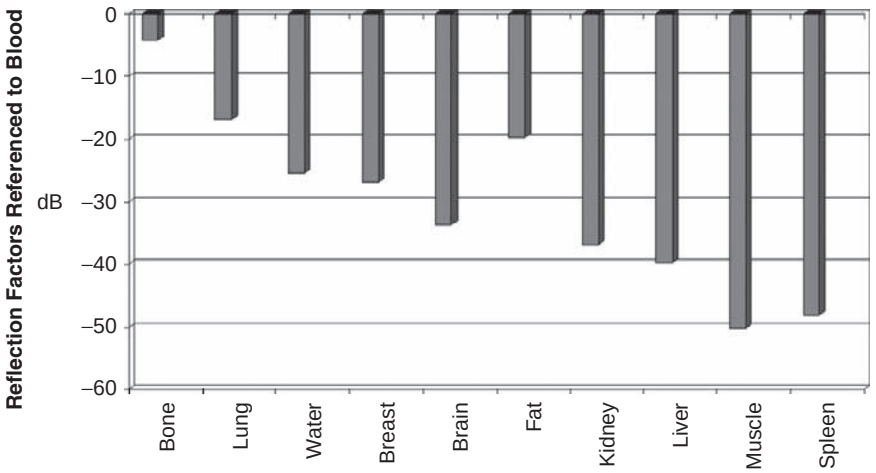


FIGURE 16.3 Amplitude reflection factor for tissues normalized to the impedance of blood and plotted on a decibel scale.

flow in the heart through Doppler-shifted motion, and I. Edler and C. H. Hertz in Sweden studied the motion of the heart and started echocardiography, the use of ultrasound to study the properties and dynamics of the heart.

Investigators around the world continued to make progress with the development of ultrasound imaging. Even though internal organs and the dynamics of heart motion could be detected with ultrasound, ultrasound remained a laboratory curiosity. Until the early 1960s, mothers and fetuses were imaged by x-rays. A 1956 report by Alice Stewart, an English epidemiologist, linked deaths from cancer in children to their mothers' exposure to x-rays during pregnancy. This tragic finding gave ultrasound imaging its first commercial opportunity to provide a safe alternative for fetal imaging.

During the early 1960s, several companies developed ultrasound imaging systems suitable for imaging fetuses and other internal organs. R. Soldner of Siemens designed the first real-time mechanical ultrasound imaging scanner in 1965 in Germany. Drs. I. McDonald and T. G. Brown developed the first commercially successful diagnostic ultrasound imaging system, the Diasonograph, in 1968.

These investigators eventually preferred a pulse-echo method of ultrasound imaging with equipment similar to that shown in Figure 16.4. The image is made up of a sequential arrangement of echo ranging lines, where each line would correspond to each pulse in a sequence, such as that depicted in Figure 16.1. The diagram shows an oval-shaped object with a sound beam piercing it. The sound is transmitted, and received echoes corresponding to the front and back boundaries of the oval at that beam location are then displayed as white dots on a vertical echo line with time increasing in the downward direction on a monitor. Next, the transducer is moved to another location, and the process is repeated until all the lines have been sent. The resulting series of lines, geometrically arranged on the display to correspond to the actual positions of the transducer, result in an ultrasound image of the object. These early images were viewed either on long persistence cathode ray tubes or captured with

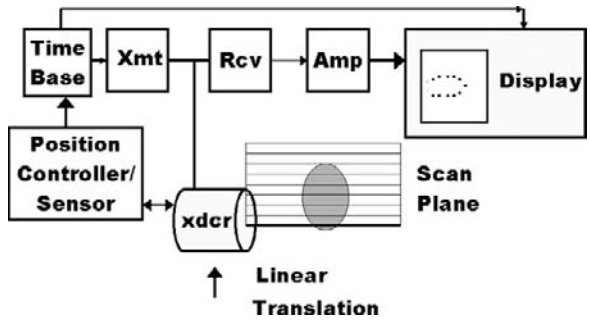


FIGURE 16.4 Basic elements of a pulse-echo imaging system shown with mechanical scanning. Note that a position controller/sensor is included along with a different form of information display compared to that of an echo ranging system.

cameras using long time exposures. The imaging presentation shown here was called B-mode, or brightness mode, in which the echo amplitudes were assigned white for maximum and black for minimum. Later, as electronics improved, a true gray scale for intermediate amplitudes could be implemented so more subtle differences among tissues could be seen.

Other investigators worked on arrays, an arrangement of small transducers that could be electronically switched and excited so mechanical movement of a single transducer could be replaced by rapid real-time array imaging. Two contributors were J. C. Somer, who showed that ultrasound phased arrays were possible in 1968, and N. Bom, who developed the first linear arrays in 1971 in the Netherlands. By the 1980s, almost all commercial imaging systems used arrays.

Many other important developments followed, some of which are highlighted in Sections 16.2.6 and 16.2.7. Some of the upcoming topics that make modern ultrasound imaging possible are scattering and absorption from tissues, transducers, focusing with arrays, imaging systems, and special signal processing methods.

16.2.2 Acoustic Wave Propagation, Reflection, and Refraction

Ultrasound imaging is based on waves that are sent into the body and received. A wave is a disturbance that moves through or along the surface of a medium. A small rock dropped on a smooth surface of water will create ripples in the form of expanding circles. In other words, the disturbances in time also have a spatial extent.

Two useful simple models for wave propagation are the plane wave and the spherical wave. A model for a spherical wave consists of a sphere expanding in radius with increasing time. A plane wave is a one-dimensional model that assumes that the lateral dimensions of the wave perpendicular to the direction of propagation are infinite in extent. A simple one-dimensional wave can be described by

$$p(z,t) = A\exp\left(i\omega(t - z/c_0)\right) + B\exp\left(i\omega(t + z/c_0)\right) \tag{16.7}$$

where t is time, and z is the propagation coordinate.

The two arguments in the exponents represent waves traveling along the positive z-axis and along the negative z-axis, respectively. For a constant frequency, for example, when the argument of the first term is equal to π, the term is real and has an amplitude of –1. The next time the amplitude of the first term has this value is for an argument of 3π. For this situation, both t and z must increase; therefore, these arguments can be recognized as wavefronts that propagate forward with a delay equal to the distance divided by the sound speed, $t - z/c_0$.

If k is a wavenumber defined as

$$k = 2\pi f/c_0 = \omega/c_0 \tag{16.8}$$

and f is frequency, a sinusoidal wave can be interpreted as a function of propagation distance at a fixed time with a wavelength, $\lambda = 2\pi/k$, or as a function of time at a specific location with a period, $T = 2\pi/\omega$, as shown in Figure 16.5.

The ratio of a traveling pressure wave, p, to the particle velocity, v, of the fluid is called the specific acoustic or characteristic impedance,

$$Z_L = p/v_L = \rho c_L \tag{16.9}$$

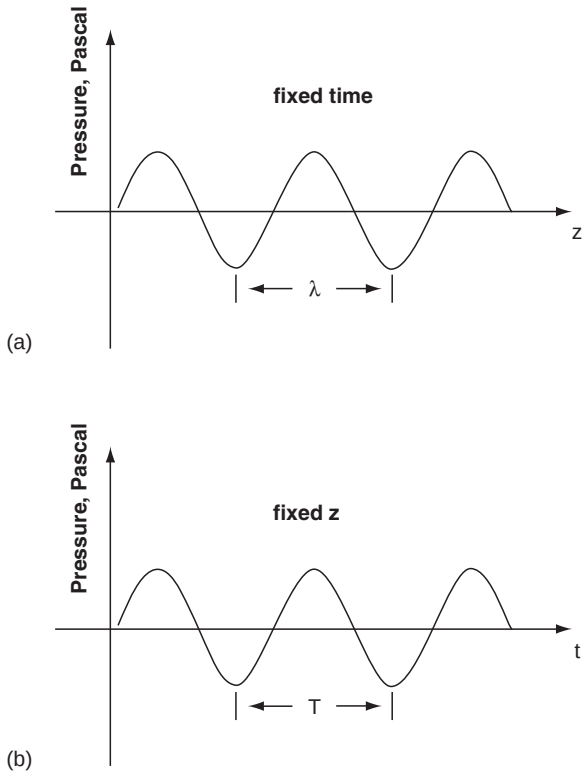


FIGURE 16.5 (a) Sinusoidal acoustic wave propagation as a function of space at a fixed time. (b) Sinusoidal acoustic wave propagation as a function of time at a fixed distance z; pressure is the acoustic parameter plotted here. The distance and time between two troughs or peaks are defined as the wavelength, (λ), and period, (T), of the wave, respectively.

Note that Z_L is positive for forward traveling waves and negative for backward traveling waves. For fresh water at 20°C, $c_L = 1{,}481$ m/s, $Z_L = 1.48$ MegaRayls (10^6 kg/m^2 sec), and $\rho = 998$ kg/m^3. The subscript L designates longitudinal waves, the type that is most important in imaging. In a longitudinal wave, the changes in wave amplitude are aligned along the direction of propagation with positive parts of sinusoidal waves, called compressional half cycles, and negative ones, called rarefactional half cycles, as depicted in Figure 16.5b.

The instantaneous intensity is

$$I_L = pv^* = pp^*/Z_L = vv^*Z_L \tag{16.10}$$

To determine the amplitude of a reflected wave, a solution similar to that of Eq. (16.7) can be applied. Consider the problem of a single frequency acoustic plane wave propagating in an ideal fluid medium a distance d to a boundary with a different medium, as shown in Figure 16.6. A plane wave is a sinusoidal-type wave with infinite extent in the lateral directions (x and y). For the example shown, the propagating medium has a wavenumber k_1 and an impedance Z_1, whereas the second medium has impedance Z_2 and a wavenumber k_2. As shown in Figure 16.6, the waves to the left of this second medium can be described by a combination of forward and backward traveling waves

$$p = p_0 \exp(i(\omega t - k_L z)) + RFp_0 \exp(i(\omega t + k_L z)) \tag{16.11}$$

which is a modified version of Eq. (16.7), $k_L = k_1$, and RF is a reflection factor for the amplitude of the negative going wave.

To determine the factor RF, it is useful to make analogies between acoustic variables and more well-known electrical parameters: pressure to voltage and particle velocity to electrical current. As shown in Figure 16.6, a source is situated at $z = d$, and the second medium is represented by a real load of impedance Z_2 located at $z = 0$, with a semi-infinite length. By analogy, the pressure at $z = 0$ is like a voltage drop across Z_2, so from Eq. (16.11), drop out common exp ($i\omega t$) terms to obtain

$$p_2 = p_0 (1 + \mathrm{RF}) \tag{16.12}$$

The particle velocity there is like the sum of currents flowing in opposite directions corresponding to the two wave components. From Eq. (16.9),

$$v_2 = (1 - RF)p_0/Z_1 \tag{16.13}$$

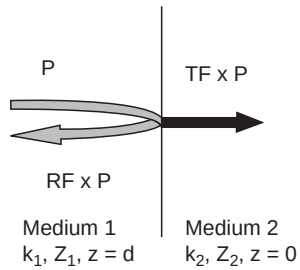


FIGURE 16.6 One-dimensional model of wave propagation at a boundary between media having impedances of Z_1 and Z_2.

The impedance, Z_2, can be found from

$$Z_2 = \frac{p_2}{v_2} = \frac{(1 + RF)Z_1}{1 - RF} \tag{16.14}$$

Finally, the right-hand side of Eq. (16.14) can be solved to obtain

$$RF = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{16.15a}$$

A transmission factor, TF, can be determined from

$$TF = 1 + RF \tag{16.15b}$$

or

$$TF = \frac{2Z_2}{Z_1 + Z_2} \tag{16.15c}$$

EXAMPLE PROBLEM 16.2

Find the reflection and transmission factors for the case of a free, a perfectly matched, and a rigid boundary. What is the acoustic pressure transferred under these conditions? Assume water as medium 1.

Solution

For a free or air-type boundary or open-circuit condition, $Z_2 = 0$, so from Eq. (16.15a), there will be a 180° inversion of the incident wave, or $RF = –1$. Here, the reflected wave cancels the incident, so $TF = 0$. For a matched condition, $Z_2 = Z_1$, $RF = 0$ or no reflection, and $TF = 1$ for perfect amplitude transfer. For a rigid boundary, $Z_2 = \infty$, corresponding to a short-circuit condition or a stress-free boundary, and the incident wave will be reflected back, or $RF = +1$, without phase inversion. In this case, $TF = 2$.

Oblique Waves at a Liquid–Liquid Boundary

What happens when the incident wave is no longer normal to the boundary? This situation is depicted in Figure 16.7, in which a single-frequency longitudinal wave traveling in a liquid medium 1 is incident at an angle to a boundary with a different liquid medium 2 in the plane x–z. At the boundary, pressure and particle velocity are continuous. The tangential components of wavenumbers also must match, so along the boundary,

$$k_{1x} = k_1 \sin \theta_I = k_2 \sin \theta_T = k_1 \sin \theta_R \tag{16.16a}$$

where k_1 and k_2 are the wavenumbers for mediums 1 and 2, respectively. The reflected angle, θ_R, is equal to the incident angle, θ_I, and an acoustic Snell's law results,

$$\frac{\sin \theta_I}{\sin \theta_T} = \frac{c_1}{c_2} \tag{16.16b}$$

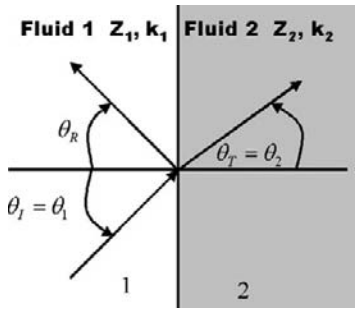


FIGURE 16.7 Oblique longitudinal waves at a liquid–liquid interface. The x-axis lies along the boundary, and the z-axis is normal to the boundary.

which can be used to find the angle θ_T. Equation (16.16a) enables the determination of θ_R. From the previous relations, the wavenumber components along z are

$$\text{incident } k_{Iz} = k_1 \cos\theta_I \tag{16.17a}$$

$$\text{reflected } k_{Rz} = k_1 \cos\theta_R \tag{16.17b}$$

$$\text{transmitted } k_{Tz} = k_2 \cos\theta_T \tag{16.17c}$$

which indicate that the effective impedances at different angles are

$$Z_{1\theta} = \frac{\rho_1 c_1}{\cos\theta_I} = \frac{Z_1}{\cos\theta_I} \tag{16.18a}$$

and

$$Z_{2\theta} = \frac{\rho_2 c_2}{\cos\theta_T} = \frac{Z_2}{\cos\theta_T} \tag{16.18b}$$

Here, the impedance is a function of angle that reduces to familiar values at normal incidence and otherwise becomes larger with angle. When the incident wave changes direction as it passes into medium 2, the bending of the wave is called refraction. For semi-infinite fluid media joined at a boundary, each medium is represented by its characteristic impedance given by Eqs. (16.18a) and (16.18b). Then, just before the boundary, the impedance looking toward medium 2 is given by Eq. (16.18b). The reflection coefficient there is given by Eq. (16.15a),

$$RF = \frac{Z_{2\theta} - Z_{1\theta}}{Z_{2\theta} + Z_{1\theta}} = \frac{Z_2 \cos\ \theta_I - Z_1 \cos\ \theta_T}{Z_2 \cos\ \theta_I + Z_1 \cos\ \theta_T} \tag{16.19a}$$

where the direction of the reflected wave is along θ_R, and the transmission factor along θ_T is

$$TF = \frac{2Z_{2\theta}}{Z_{1\theta} + Z_{2\theta}} = \frac{2Z_2 \cos\ \theta_I}{Z_2 \cos\ \theta_I + Z_1 \cos\ \theta_T} \tag{16.19b}$$

To solve these equations, θ_T is found first from Eq. (16.16b). This liquid–liquid interface is often used to model waves approximately at a tissue-to-tissue boundary.

EXAMPLE PROBLEM 16.3

For a wave incident at a water–honey interface, determine the reflection factor at 45° and at 50°. The sound speeds and impedances for water (medium 1) and honey (medium 2) are 1.48 km/s and 2.05 km/s and 1.48 MegaRayls and 2.89 MegaRayls, given in Table 16.1.

Solution

From Snell's law, Eq. (16.16b), $\theta_T = \arcsin[(2.05/1.48)\sin 45] = 78.4°$. Then from Eq. (16.19a),

$$RF = \frac{2.89 \times \cos 45 - 1.48 \times \cos 78.4}{2.89 \times \cos 45 + 1.48 \times \cos 78.4} = 0.746$$

Before finding the result for 50°, it is worth finding the critical angle at which the incident wave is directed along the boundary. This angle, found by setting $\theta_T = 90°$ in Snell's law, is known as the critical angle, which is $\theta_I = 46.22°$ in this case. For an incident angle of 50° that is past the critical angle, the wave is just reflected and does not get transmitted, or $RF = 1$.

16.2.3 Transducer Basics

The essential part of an ultrasound system is a means to generate and receive acoustic waves. This function is performed by the transducer, which can convert electrical signals to acoustic pressure waves and vice versa. Inside the transducer is a piezoelectric crystal or ceramic that deforms when a voltage is applied and transmits acoustic waves. A reciprocal piezoelectric effect, in which electrical charge is created by the mechanical deformation of a crystal, allows the transducer to convert returning acoustic waves back to electrical signals. The piezoelectric effect was discovered by the Curie brothers in 1880.

A simple model is presented that describes the basic acoustic and electrical characteristics of a piezoelectric transducer. Consider a transducer to consist of a rectangular piece of piezoelectric material with electrodes on the sides, as shown in Figure 16.8. Each side has a cross-sectional area, A, with a top and bottom that are much longer (>10X) than the thickness, d. Because piezoelectric material is dielectric, it is essentially a capacitor with a clamped capacitance

$$C_0 = \varepsilon^S A/d \tag{16.20}$$

in which ε^S is a clamped dielectric constant under the condition of zero deformation. Because the crystal is a solid rather than a liquid, elastic waves are created internally. The elastic counterpart of pressure is stress, or force per unit area, represented by the symbol T. Another important elastic variable is strain, S, the relative change in length of the crystal divided by its original length. Both stress and strain vary with direction in a piezoelectric material, but for the single direction of wave propagation, it is possible to relate the stress to the strain as

$$T = C^D S - hD \tag{16.21}$$

in which h is a piezoelectric constant. This equation is known as a modified Hooke's law, in which the first term relates the stress to the strain and the second term relates stress to the

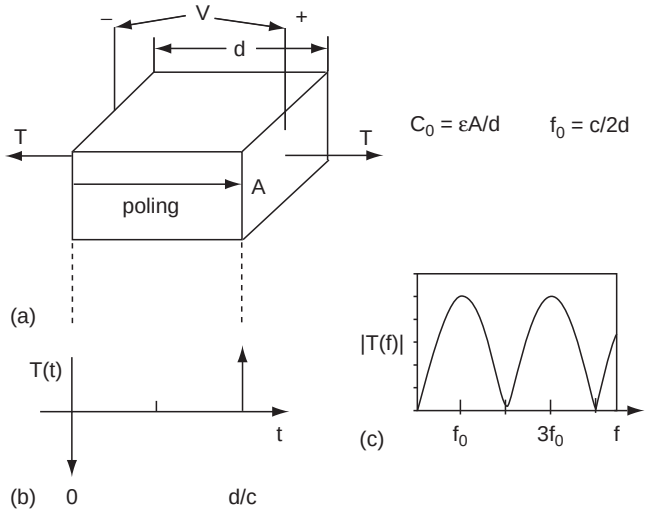


FIGURE 16.8 (a) Diagram for a piezoelectric crystal radiating into a medium matched to its impedance, (b) stress time response, and (c) stress frequency response.

applied electric field. The elastic stiffness constant, C^D, is obtained under a constant dielectric displacement, D, and if E is the electric field,

$$D = \varepsilon^S E = \frac{\varepsilon^S AV}{dA} = C_0 V/A \tag{16.22}$$

When a voltage impulse is applied across the electrodes, the piezoelectric effect creates impulsive forces at the sides of the transducer, given by

$$F(t) = TA = (hC_0V/2)[-\delta(t) + \delta(t - d/c)] \tag{16.23}$$

where the media above and below have the same acoustic impedance, Z_c, as the transducer; the speed of sound between the electrodes is given by $c = \sqrt{C^D/\rho}$, and δ represents an impulse (see Figure 16.8b).

Since it can create acoustic waves, the crystal could be regarded as a singing capacitor with its own unique voice or spectral characteristics and resonant frequency. To obtain the spectrum of this response, take the Fourier transform of Eq. (16.23),

$$F(f) = -i(hC_0V)e^{\frac{-i\pi fd}{c}} \sin\left[\pi(2n+1)f/2f_0\right] \tag{16.24}$$

which has maxima at odd harmonics (note $n = 0, 1, 2, 3, \ldots$) of the fundamental resonance $f_0 = c/2d$, as shown in Figure 16.8c.

Transducer Electrical Impedance

Because of the forces generated at the sides of the transducer, the electrical impedance looking through the voltage terminals is affected. Across the wires connected to the transducer (Figure 16.8a), a radiation impedance, Z_A, is seen in addition to the capacitive reactance, so the overall electrical impedance is

$$Z_T = Z_A - i(1/\omega C_0) = R_A(f) + i[X_A(f) - 1/\omega C_0] \tag{16.25}$$

From the transducer electrical equivalent circuit in Figure 16.9a, Z_A is radiation impedance of which R_A and Z_A are its real and imaginary parts. $R_A(f)$ can be found from the total real electrical power, W_E, flowing into the transducer for an applied voltage V, and current I

$$W_E = II^*R_A/2 = |I^2|R_A/2 \tag{16.26a}$$

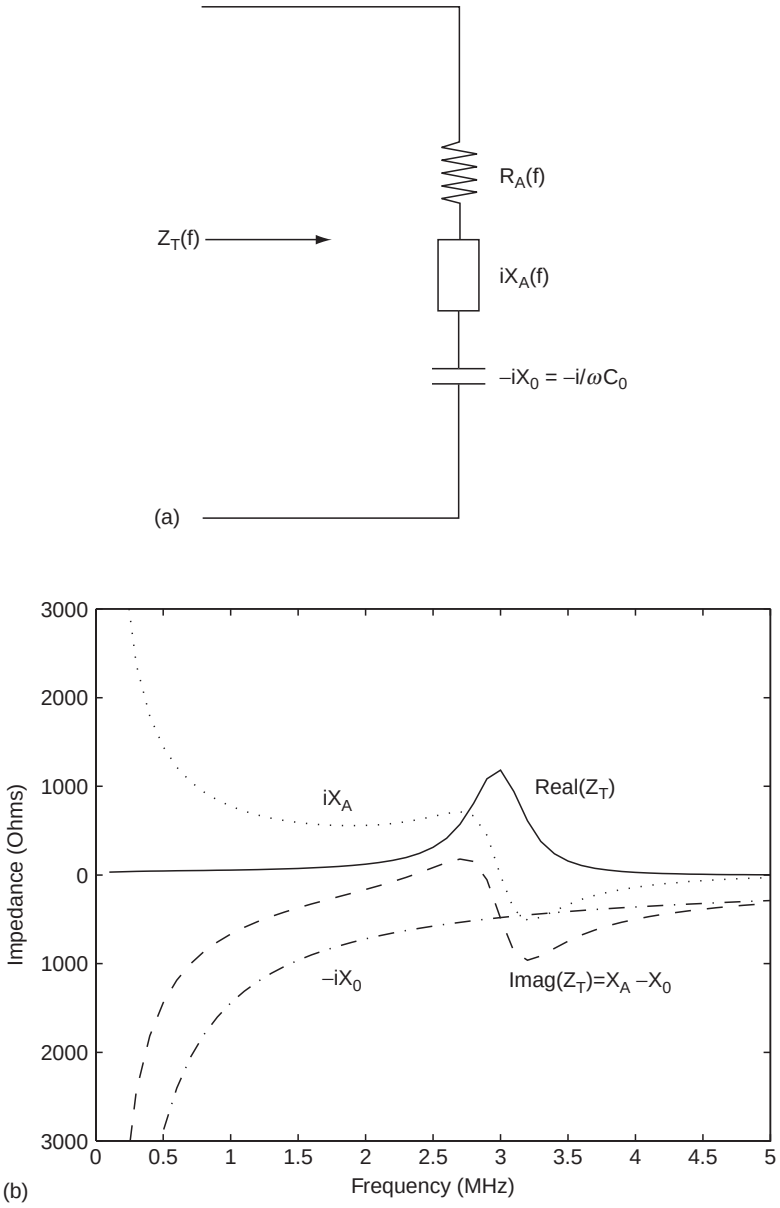


FIGURE 16.9 (a) Transducer equivalent circuit and (b) transducer impedance as a function of frequency.

where current is $I = i\omega Q = i\omega C_0 V$ and Q is charge. The total power radiated from both sides of the transducer into a surrounding medium of a modified specific acoustic impedance (including an area factor), $Z_C = \rho c A$, equal to that of the crystal, is

$$W_A = ATT^*/(2Z_C/A) = A^2|F(f)/A|^2/2Z_C = |hC_0V\sin(\pi f/2f_0)|^2/2Z_C. \tag{16.26b}$$

Setting the powers of Eqs. (16.26a) and (16.26b) equal leads to a solution for R_A

$$R_A(f) = R_{AC}\ \text{sin c}^2(f = 2f_0) \tag{16.27a}$$

where $\text{sin c}(x) = \sin(\pi x)/(\pi x)$ and

$$R_{AC} = \frac{k_T^2}{4f_0C_0} = \frac{d^2k_T^2}{2A\varepsilon^S} \tag{16.27b}$$

in which the electroacoustic coupling constant is k_T, and $k_T = h/\sqrt{C^D/\varepsilon^S}$. Note that R_{AC} is inversely proportional to the capacitance and area of the transducer and directly dependent on the square of the thickness, d. Also, note that at resonance,

$$R_A(f_0) = \frac{k_T^2}{\pi^2 f_0 C_0} \tag{16.27c}$$

Network theory requires that the imaginary part of an impedance be related to the real part through a Hilbert transform, so the radiation reactance can be found as

$$X_A(f) = \Im_{Hi}[R_A(f)] = R_{AC}\frac{[\sin(\pi f/f_0) - \pi f/f_0]}{2(\pi f/2f_0)^2} \tag{16.28}$$

The transducer impedance is plotted as a function of frequency in Figure 16.9b. Here R_A is maximum at the resonant frequency f_0 and X_A is zero.

This simple model describes the essential characteristics of a piezoelectric transducer. The electrical impedance has a maximum of real radiation resistance at the resonant frequency. The force spectrum is also peaked at the resonant frequency and has a certain shape.

EXAMPLE PROBLEM 16.4

Find the transducer impedance at resonance for a square transducer 2.5 mm on a side for a resonant frequency of 3 MHz for the ceramic PZT5A with $\varepsilon^s/\varepsilon_0 = 830$, $c = 4.35$ km/s, $k_T = 0.49$, and $\varepsilon_0 = 8.85$ pF/m.

Solution

First find the thickness needed to achieve a resonant frequency of 3 MHz from $f_0 = c/2d$, or $d = c/2f_0 = 4.35\ \text{mm}/\mu\text{s}/2 \times 3\ \text{MHz} = 0.725$ mm. Then the capacitance is

$$C_0 = \frac{830 \times 8.85 \times 10^{-12}F/\text{m} \times (2.5 \times 10^{-3}\text{m})^2}{7.25 \times 10^{-4}\text{m}} = 63.3\ \text{pF}$$

Find R_A from Eq. (16.27c):

$$R_A = \frac{0.49^2}{\pi^2 \times 3 \times 10^6 \text{ Hz} \times 63.3 \times 10^{-12} F} = 128.2 \text{ ohms}$$

Note that the units in the denominator are Hz – F = amps/volt = ohm^{-1}. Then, since at resonance $X_A = 0$, from Eq. (16.25),

$$Z_T = 128 - i/(2 \times \pi \times 3 \times 10^6 \text{ Hz} \times 63.3 \times 10^{-12} F) = 128 - i838 \text{ ohms}$$

Transducer Frequency Response

Transducer design is concerned with altering the shape of the spectrum to achieve a desired bandwidth and a short, well-behaved pulse shape, or impulse response. Another design goal is to improve the electroacoustic efficiency of the transducer. One measure of this efficiency is transducer loss, the ratio of time average acoustic power reaching the desired medium, usually tissue, W_R, divided by the maximum electrical power available from an electrical source, W_g,

$$TL(f) = \frac{W_R}{W_g} \tag{16.29a}$$

and defined in dB as

$$TL_{\text{dB}}(f) = 10 \log_{10} TL(f) \tag{16.29b}$$

The transducer loss can be broken down into an electrical loss factor, $EL(f)$, and an acoustic loss factor, $AL(f)$:

$$TL(f) = EL(f)AL(f) \tag{16.30}$$

The problem of optimizing the transfer of electrical power into the radiation resistance is that of maximizing $EL(f)$ over a desired bandwidth. Figure 16.10 shows the relevant parameters involved. The transducer electrical impedance is connected through a tuning inductor to a voltage source with impedance R_g. An expression for electrical loss for this situation is

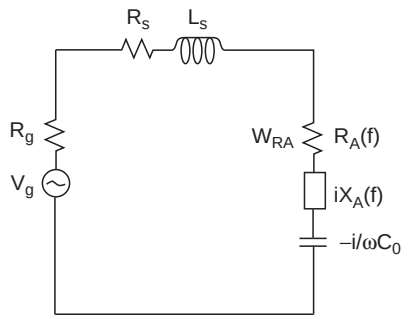


FIGURE 16.10 Electrical voltage source with tuning inductor connected to transducer equivalent circuit.

$$EL(f) = \frac{4R_A(f)R_g}{(R_A(f) + R_g + R_s)^2 + \left(X_A(f) - \dfrac{1}{\omega C_0} + \omega L_s\right)^2} \tag{16.31a}$$

If the capacitance is tuned out by a series inductor, $L_S = 1/(\omega_0^2 C_0)$ and $R_s = 0$, then at resonance,

$$EL(f_0) = \frac{4R_A(f_0)R_g}{[R_A(f_0) + R_g]^2} \tag{16.31b}$$

Notice that if $R_A = R_g$, and $R_s \ll R_g$, then $EL(f_0) \sim 1$.

The determination of acoustic loss is a more complicated function of frequency, but it is straightforward to determine its value at resonance. In the simple transducer model introduced in Section 16.2.3, a simplifying assumption was made that the acoustic loading on both sides of the piezoelectric crystal was the same acoustic impedance as that of the crystal; however, this is not the case in general, as illustrated in Figure 16.11. Note the arrangement in this figure is rotated 90° from that in Figure 16.8. On the top of this figure, the piezoelectric is loaded by a backing material and below the lens, by water or tissue. The bottom of the crystal shown here is usually regarded as the "business end" of the transducer, where the forward waves propagate along the positive z-axis that is directed downward in Figure 16.11. Waves in the backward direction are suppressed or absorbed by a backing material that aids in broadening the spectrum of the transducer. By the left-right convention of Figure 16.8, if the acoustic impedance looking out of the crystal to the left is Z_L, then that to the right along the positive z-axis is Z_R. In the case shown on the left

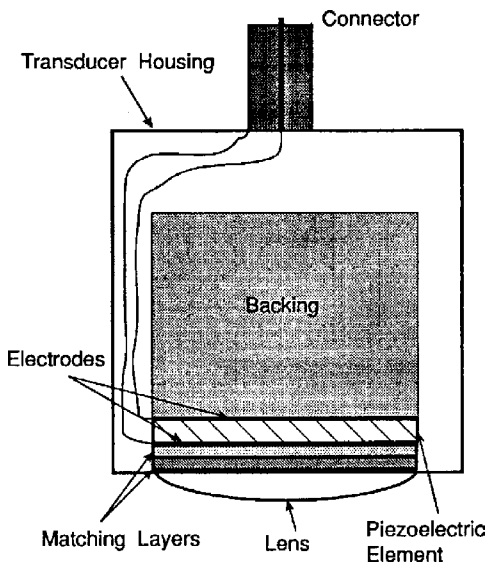


FIGURE 16.11 Construction of a single-crystal mechanical transducer. Forward propagation is directed downward.

of the figure, $Z_L = Z_B$ (backing impedance) and $Z_R = Z_W$ (impedance of water). The acoustic loss for this case at resonance can be found from (note all these impedances include the area factor A)

$$AL(f_0) = Z_R/(Z_L + Z_R) = Z_W/(Z_B + Z_W) \tag{16.32}$$

The physical meaning of this equation is that the electrical power reaching the radiation resistance is converted into acoustical power radiating from the two electroded faces of the piezoelectric; the proportion of power in the forward (right) direction is represented by the ratio in Eq. (16.32).

Construction of a single crystal transducer is shown in Figure 16.11. The key parts are the backing, the piezoelectric crystal coated with thin electrodes for electrical contact, matching layers, and a lens material, with an acoustic impedance close to water, for focusing (discussed in more detail in Section 16.2.6). Matching layers will be discussed shortly; in practice, more than one matching layer is used to broaden the bandwidth of the transducer.

In the first simple model, the crystal was loaded by acoustic impedances equal to that of the crystal, or $Z_C = Z_L = Z_R$ so that $AL(f_0) = 0.5$, as expected. Another example is that in which the left side is loaded by air or $Z_L = 0A$ MRayls, and the right by water, $Z_R = 1.5A$ MRayls. In this case, $AL(f_0) = 1.0$ but at the expense of extremely narrow bandwidth and a correspondingly long pulse. One case of interest is where the backing material is matched to that of the crystal so that $Z_L = Z_c = 30A$ MRayls and $Z_R = 1.5A$ MRayls, and $AL\ (f_0) = 0.048$, an inefficient transfer.

To improve the forward transfer efficiency, as described by the acoustic loss factor, matching layers are employed. At the resonant frequency, a matching layer is designed to be a quarter of a wavelength thick and to be the mean value of the two impedances to be matched, or

$$Z_{ml} = \sqrt{Z_1 Z_2} \tag{16.33}$$

EXAMPLE PROBLEM 16.5

Match a crystal of impedance $Z_c = 30A$ MRayls to water, $Z_w = 1.5A$ MRayls.

Solution

From Eq. (16.33), the matching layer impedance is

$$Z_{ml} = \sqrt{Z_W Z_C} = \sqrt{1.5A \text{ MRayls} \times 30A \text{ MRayls}} = 6.7A \text{ MRayls}$$

The acoustic input impedance at resonance is

$$Z_1 = Z_{ml}^2/Z_2$$

therefore, the impedance looking to the right is

$$Z_R = Z_{ml}^2/Z_W = (6.7A)^2/1.5A = 30A \text{ MRayls}$$

leading to $A(f_0) = 30/(30 + 30) = 0.5$ for a water load with an intervening matching layer or an improvement of over 10 dB.

Continued

The effect of adding a matching layer can be included in a more generalized equation for radiation resistance at resonance by adding the loading from both sides,

$$R_A(f_0) = \frac{2k_T^2}{\pi^2 f_0 C_0}\left(\frac{Z_c}{Z_L + Z_R}\right) \tag{16.34}$$

The acoustical and electrical losses can now be combined to estimate the overall transducer loss. To do this, a simplification is made for the purposes of illustration that the acoustical loss is constant over the transducer bandwidth. The actual losses can be computed accurately by a more complete equivalent circuit model. From Eqs. (16.30) and (16.31a),

$$TL(f) \approx \left[\frac{4R_A(f)R_g}{(R_A(f) + R_g + R_s)^2 + (X_A(f) - \frac{1}{\omega C_0} + \omega L_s)^2}\right]\left[\frac{Z_R}{Z_R + Z_L}\right] \tag{16.35}$$

EXAMPLE PROBLEM 16.6

Find the transducer loss at resonance for the transducer described in Example Problem 16.4 under the following two conditions for a source resistance $R_g = 50\Omega$ (ohms): (1) No tuning, $Z_L = Z_B = 6A$ MRayls and $Z_R = Z_W = 1.5A$ MRayls. (2) Add a tuning inductor and a matching layer.

Solution

(1) Here, $R_A = 128$ ohms, $\dfrac{1}{\omega C_0} = 838$ ohms,

$$\begin{aligned} TL(f_0) &= \frac{4 \times 128 \text{ ohms} \times 50 \text{ ohms}}{(128 \text{ ohms} + 50 \text{ ohms})^2 + (838 \text{ ohms})^2} \times \frac{1.5A}{6A + 1.5A} \\ &= 6.98 \times 10^{-3} \text{or } 10\log_{10}(TL) \\ &= -21.6 \text{ dB} \end{aligned}$$

(2) R_A is determined from Eq. (16.34) and Example Problem 16.4, ohms. For tuning, find $L_S = 1/(\omega_0^2\, C_0) = 838/(2\pi \times 3e6) = 44.5\ \mu H$. Then,

$$\begin{aligned} TL(f_0) &= \frac{4 \times 128 \text{ ohms} \times 50 \text{ ohms}}{(128 \text{ ohms} + 50 \text{ ohms})^2 + (838 \text{ ohms} - 838 \text{ ohms})^2} \times \frac{30A}{30A + 1.5A} \\ &= .673 \times \text{ or } 10\log_{10}(TL) \\ &= -1.72 \text{ dB} \end{aligned}$$

One way of describing the difference in shapes as a figure of merit is that a bandwidth is stated in dB. For example, a –6 dB bandwidth is defined relative to the peak spectral value. A center frequency is defined through the upper and lower –6 dB frequencies as

$$f_c = (f_{high} - f_{low})/2$$

An example of a –6 dB bandwidth with its associated frequencies is shown in Figure 16.12. Also included in this figure are similar widths for the pulse envelope (pulse-widths) in dB levels.

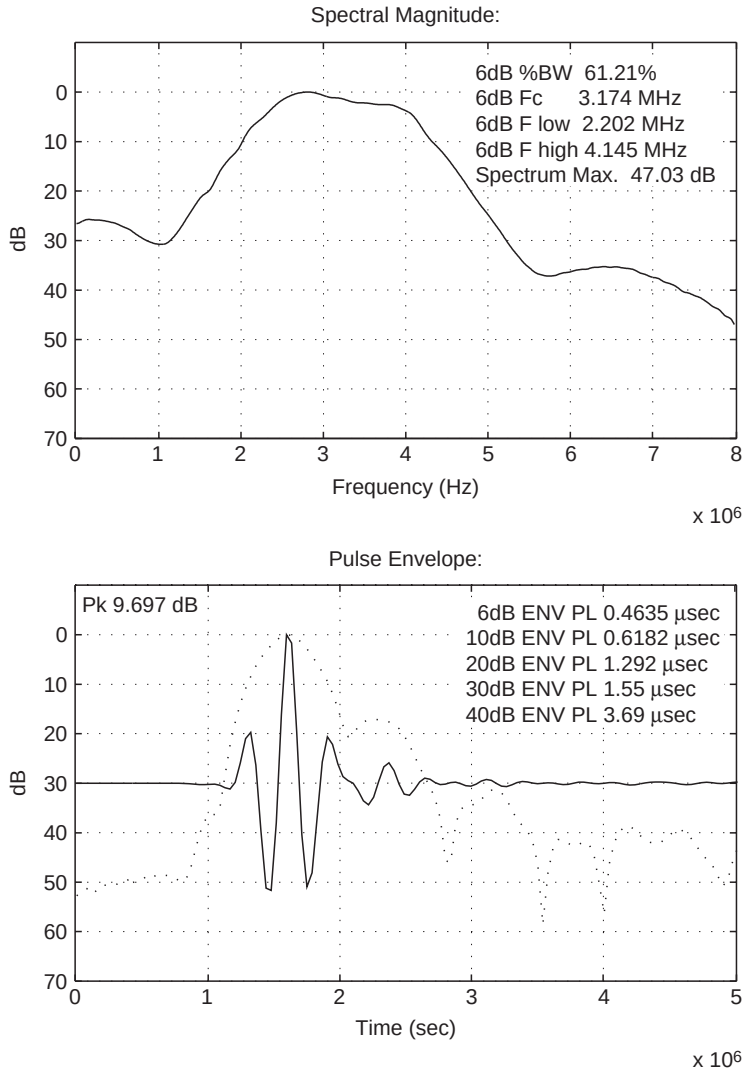


FIGURE 16.12 (Top) Transducer response spectral magnitude and –6 dB bandwidth dB. (Bottom) Transducer impulse response and its envelope measured in pulse-widths in dB.

Transducer Types and Arrays

Most imaging systems use arrays rather than single-crystal transducers. A one-dimensional array is an in-line arrangement of transducers called elements that are addressable individually or in small groups. The elements are spaced at regular intervals, typically one-half to two wavelengths apart in water (Figure 16.13). The key advantages of arrays are that they can be rapidly focused and steered electronically or electrically switched, whereas a strictly mechanical single transducer has a fixed focal length and can only be

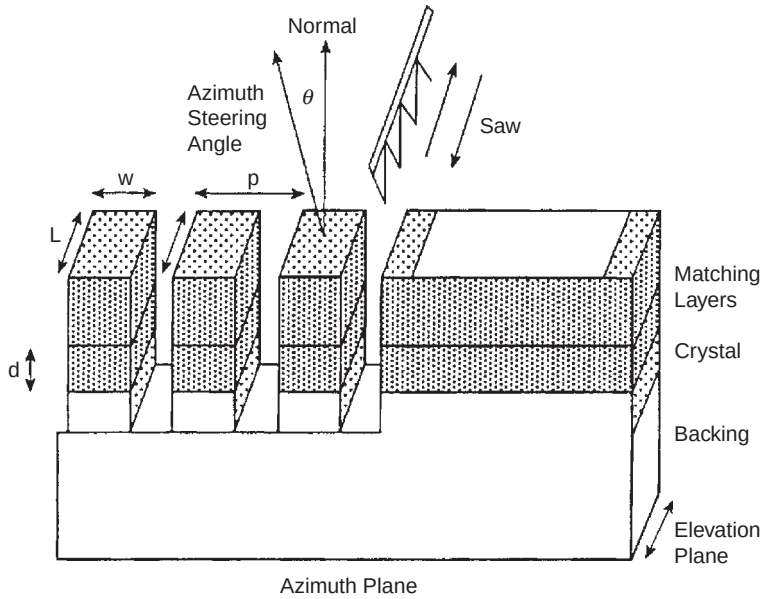


FIGURE 16.13 Construction of a one-dimensional array with a saw dicing up a multilayer sandwich structure into individual elements. Here, the propagation direction is along the normal axis, pointing upward.

steered or translated mechanically. Each element, in terms of basic design, is treated as an individual transducer. Arrays and focusing will be discussed in more detail in Section 16.2.6 on diffraction.

The two most common types of arrays are the linear array and the phased array. The linear array forms an image by translating the active aperture (a certain number of elements) one element at a time along the length of the array, as shown in Figure 16.14a. At each position of the active aperture, an acoustic line is created (i.e., a pulse-echo time record of a selected length or scan depth). A total number of lines, N, is formed in this

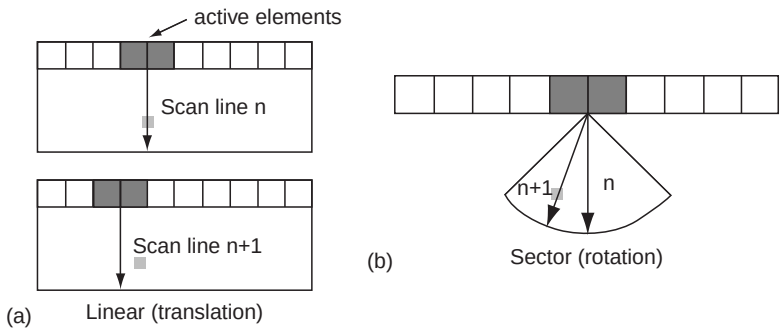


FIGURE 16.14 The time sequencing used for image formation for (a) linear array format and (b) sector array format.

way. In a phased array, the center of the active aperture is always the same, and the scanning of acoustic lines is accomplished through electronic angular steering. Each line is steered by a small incremental angle from the previous one, as shown in Figure 16.14b. When N lines have been received, these lines form the basis for an image frame. Examples of the image formats formed by these two array types can be seen in the B-mode images of Figures 16.15 and 16.16. A variant of the linear array is the curved linear array, which operates like a linear array but on a curved convex surface rather than on a flat surface. An example of an image from a curved linear array can be seen in Figure 16.17.

EXAMPLE PROBLEM 16.7

For the example shown in Figure 16.14a, determine the frame rate for a scan depth of $s_d = 150$ mm, 100 lines per frame, and $c_0 = 1.5$ mm/μs.

Solution

The round-trip time for one line is $2^*s_d/c_0 = 200$ μs. The time for a full frame is N lines/frame or, in this case, 100 lines/frame × 200 μs/line = 20 ms/frame or 50 frames per second.

Transducer arrays come in a variety of sizes, shapes, and center frequencies to suit different clinical applications, as shown in Figure 16.18. Access to the body is made externally through many possible "acoustic windows," where a transducer makes contact by coupling

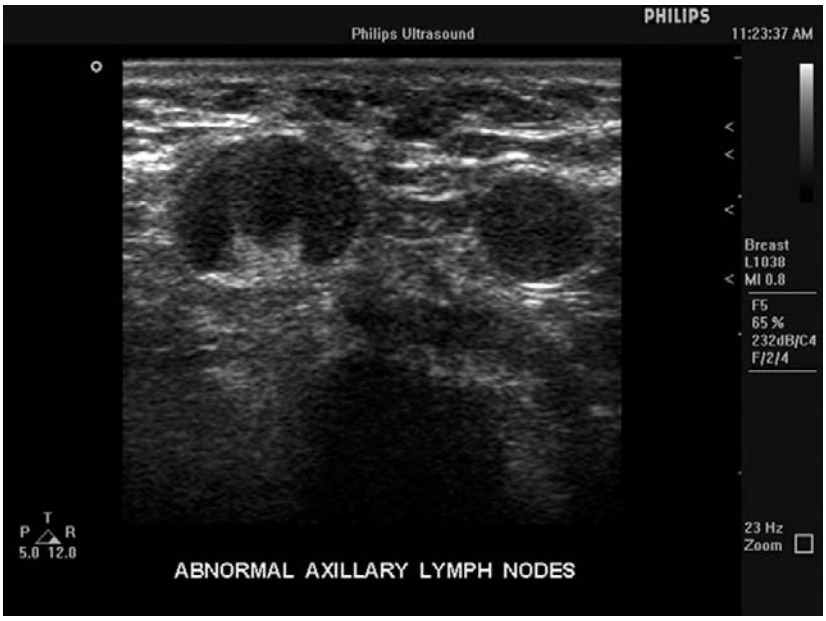


FIGURE 16.15 B-mode image of lymph nodes in the breast at 12 MHz, an example of a linear array format. *Courtesy of Philips Medical Systems.*

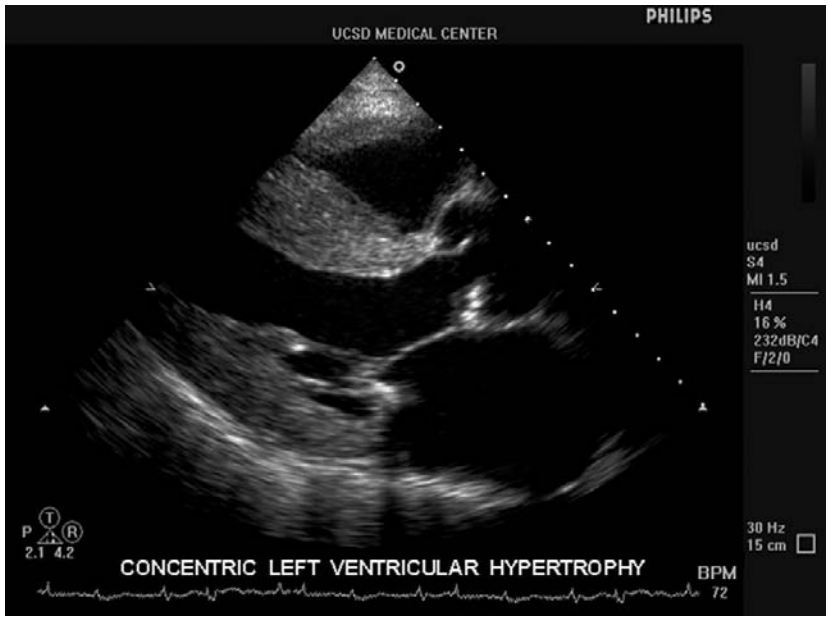


FIGURE 16.16 B-mode image of the heart at 4 MHz, an example of a sector array format. *Courtesy of Philips Medical Systems.*

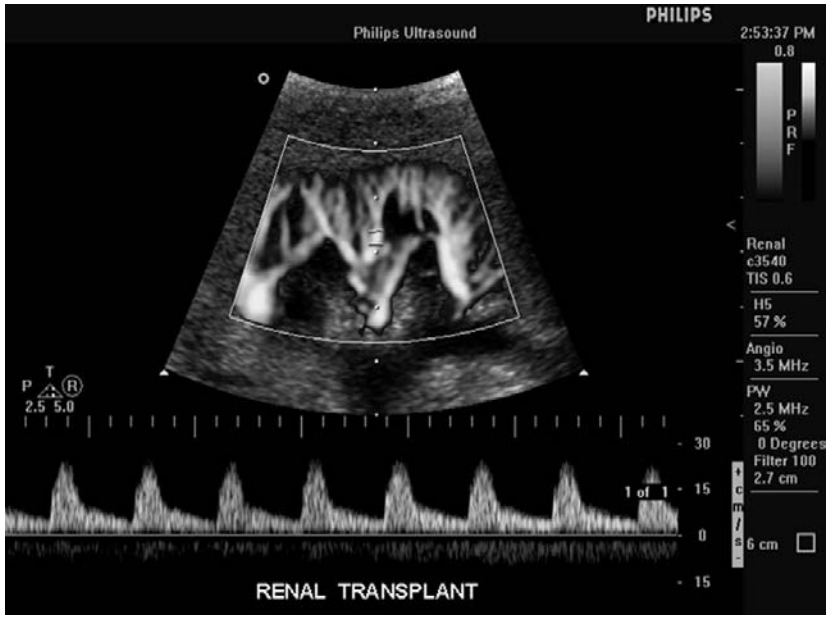


FIGURE 16.17 (Top) Power Doppler image of kidney with pulsed Doppler line down the center and Doppler gate in center of line. (Below) Corresponding pulsed wave Doppler spectrum. Image from a curved linear array. *Courtesy of Philips Medical Systems.*

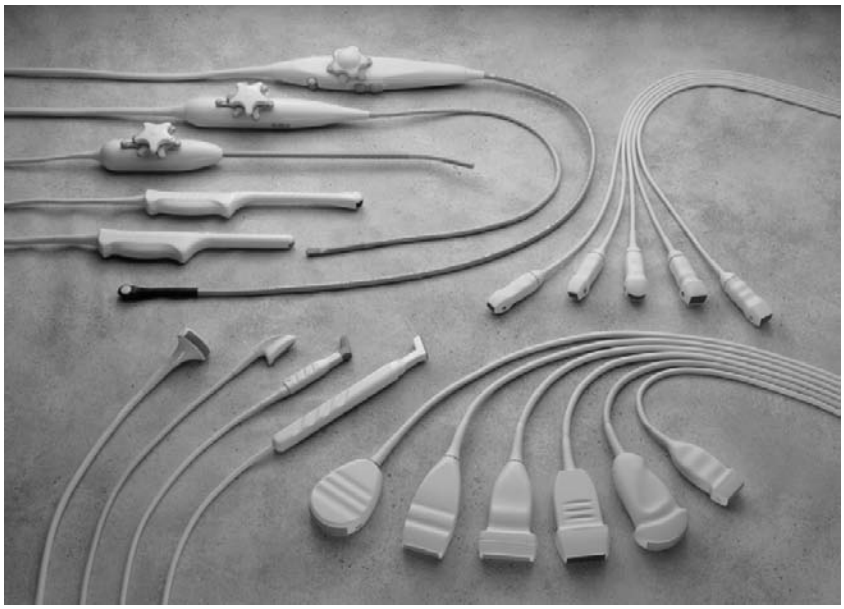


FIGURE 16.18 Transducers available on a modern imaging system are designed for a wide variety of clinical applications. Transducer groups: bottom right: linear and curved linear arrays; top right: phased arrays; left side: specialty probes including intraoperative, transesophageal, and transvaginal arrays. *Courtesy of Philips Medical Systems.*

to the body with a water-based gel. Except for regions containing bones, air, or gas, which are opaque to imaging transducers, even small windows can be enough to visualize large interior regions. The limitation of accessibility to viewing certain organs clearly is offset by specialized probes such as transesophageal (down the throat) and intracardiac (inside the heart) transducers that image from within the body.

16.2.4 Scattering

Interference of waves within the transducer plays an important role in the transduction process. The crystal resonates when its thickness is half a wavelength, and the matching layer is designed to be a quarter of a wavelength thick at resonance. The size of an object relative to a wavelength is also a useful way of looking at acoustic scattering from objects. Since ultrasound imaging is based on pulse-echoes returning from organs, it is necessary to determine how the reflected signals are affected by the dimensions and shape of an object relative to the insonifying wavelength.

Scattering falls within three ranges of effects: specular, diffractive, and diffusive (Figure 16.19). Specular scattering is already familiar. When the dimensions of an object are much greater than a wavelength, the reflected sound returns at the reflected angle equal to the incident angle relative to the surface, as described in Section 16.2.2 on refraction. In this case, the reflected amplitude is determined by the reflection

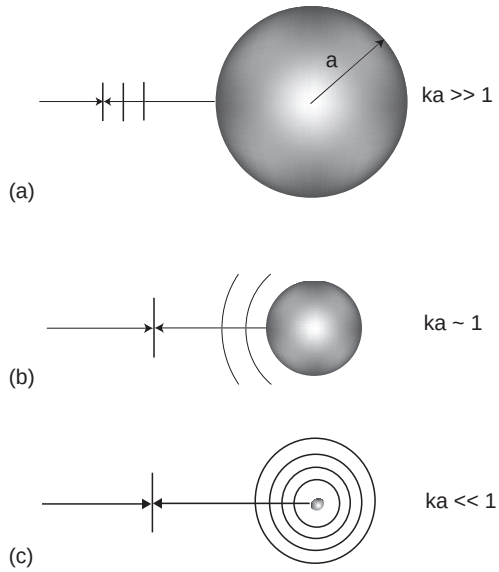


FIGURE 16.19 (a) Specular scattering from an object much greater than a wavelength. (b) Diffractive scattering from an object on the order of wavelengths. (c) Diffusive scattering from an object much smaller than a wavelength.

factor appropriate for the angle of incidence. At the other extreme, when an object is much smaller than a wavelength, its reflections are diffusive. Here, the object is so small relative to a wavelength that its features no longer contribute to the reflected wave. If the object is a sphere, then the reflected pressure is proportional to the frequency squared, inversely with the radius cubed and a term related to the differences in elastic constants and densities between the sphere and the surrounding medium.

The intermediate range between these two extremes, diffractive scattering, occurs when the dimensions of the object are on the order of wavelengths. If the surface of the object is divided into infinitesimal points and lines are drawn from these points to an observation point, then the differences in path lengths along these lines can be significant in terms of wavelengths. Contributions from these points could cause constructive and destructive interference effects at the observation point because of phase differences along the paths. Depending on the shape of the object and its orientation and distance to an observation point, complicated scattering patterns can result. Even for simply shaped objects, such as spheres or cylinders, the scattering pattern or directivity of these objects is highly frequency dependent as the object shape and dimensions change relative to a wavelength.

Human optical vision depends mainly on specular reflections (objects much greater than wavelengths of light). An acoustic image is formed from pulse-echoes along acoustic lines from an observation point to parts of the object. Unlike optics, the acoustic appearance of an object can change both with the orientation of the observer relative to the object and with the insonifying frequency.

16.2.5 Absorption

When waves propagate in real media, losses are involved. Just as forces encounter friction, pressure, and stress, waves lose energy to the medium of propagation and result in weak local heating. These small losses are called attenuation and can be described by an exponential law with distance. For a single frequency, f_c, plane wave, a multiplicative amplitude loss term can be added:

$$A(z,t) = A_0 \exp\,(i(\omega_c t - kz)) \exp\,(-\alpha z) \tag{16.36}$$

The attenuation factor, α, is usually expressed in terms of nepers/cm in this form. Another frequently used measure of amplitude is the deciBel (dB), which is most often given as the ratio of two amplitudes, A and A_0 on a logarithmic scale

$$\text{Ratio(dB)} = 20\ \log_{10}(A/A_0) \tag{16.37}$$

or in those cases where intensity is simply proportional to amplitude squared ($I_0 \propto A_0^2$),

$$\text{Ratio(dB)} = 10\ \log_{10}(I/I_0) = 10\ \log_{10}(A/A_0)^2 \tag{16.38}$$

Most often, α, is given in dB/cm:

$$\alpha_{dB} = 1/z\{20 \times\ \log_{10}[\exp\,(-\alpha_{\text{nepers}} z)]\} = 8.6886\ (\alpha_{\text{nepers}}) \tag{16.39}$$

Data indicate that the absorption is a function of frequency (see Table 16.1). Many of these losses obey a frequency power law defined as

$$\alpha(f) = \alpha_0 |f|^y \tag{16.40}$$

in which y is a power law exponent. The pressure amplitude can be written to first order as a function of frequency

$$A(z,f) = A_0 \exp\,(-i2\pi f z/c) \exp[-\alpha_0 |f|^y z] \tag{16.41}$$

where the first factor describes the propagation delay to z. The actual loss per wavenumber is very small, or $\alpha/k \ll 1$. Even though the loss per wavelength is small, absorption has a strong cumulative effect over many wavelengths. Absorption for a round-trip echo path usually determines the allowable tissue penetration for imaging.

EXAMPLE PROBLEM 16.8

Determine the absorption loss for 10 cm of propagation in muscle for a frequency of 5 MHz.

Solution

From Table 16.1, $\alpha_0 = 0.57$ dB/cm-MHz. Method 1: Convert the absorption coefficient to nepers from Eq. (16.39), α (nepers/MHz-cm) $= 0.57/8.6886 = 0.0656$. Use this value in Eq. (16.41) to obtain $A/A_0 =$ exp (–0.0656 nepers/MHz-cm $\times$ 5 MHz $\times$ 10 cm) $= 3.76 \times 10^{-2}$. In dB terms, 20 $\log_{10}$ (A/A_0) = –28 5 dB. Method 2: If the linear answer is not needed, the answer in dB can be obtained directly by using the absorption coefficient in dB as $\alpha_0 fz$ = –0.57dB/MHz-cm $\times$ 5 MHz $\times$ 10 cm = –28.5 dB.

16.2.6 Diffraction

Waves transmitted by a transducer are not plain but form a complicated pattern. The formation of these patterns, as shown in Figure 16.20, is caused by the radiation of sound waves from different locations on the aperture (transducer face) and the mutual interference of these radiated waves. This phenomenon, also called diffraction, is a consequence of the aperture dimensions that are on the order of wavelengths.

The two most common aperture shapes are the circle and rectangle, as shown in Figure 16.21. A slice of the three-dimensional beam in a plane is what is usually depicted in graphs. For the circle, because of symmetry, any plane through the beam axis (here the z-axis) will be identical. For the rectangular aperture, the beam formation differs in all planes through the beam axis, and the most important planes for imaging are the x–z and y–z planes. The beam amplitude described by Figure 16.20 corresponds to an x–z plane from a rectangular aperture.

Beams have recognizable landmarks. A method borrowed from maps is a contour plot of the acoustic pressure magnitude, often depicted in dB relative to maximum amplitude at each depth. Of particular interest is the –6 dB contour. A cross section of the beam, perpendicular to the beam axis, is called a beam plot. The width between points of this –6 dB

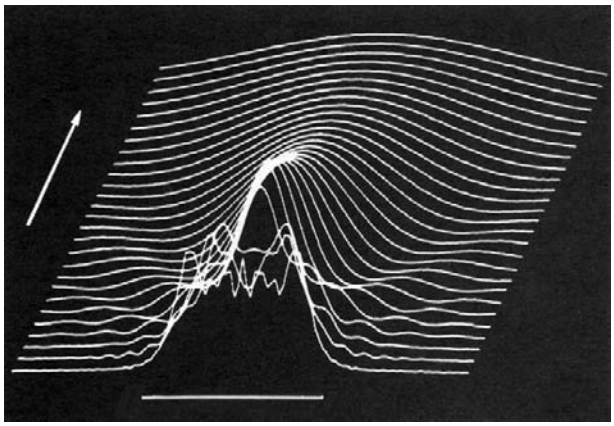


FIGURE 16.20 Diffracted field of a 40-wavelength-wide line aperture along the x-axis. The vertical axis intensity and beam profiles are shown at intervals of about 70 wavelengths along the beam axis that is compressed relative to the lateral dimension (1,920 wavelengths are shown along the z-axis).

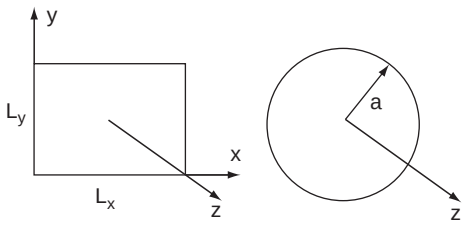


FIGURE 16.21 (Left) Circular aperture of radius a. (Right) Rectangular aperture with lengths of L_x parallel to the x-axis and L_y parallel to the y-axis. The z-axis is perpendicular to the x–y plane of the aperture.

contour on a beam plot is called the full-width-half-maximum (FWHM). A curious outcome of the radiation from these apertures is that there is a region in which the beam narrows. The depth where the last axial peak occurs is called the transition distance or natural focal length, F_N. This transition depth demarcates two regions, one with peaks and valleys, called the near field, and one with a beam with a single peak diminishing in amplitude width and broadening with distance, called the far field, as shown in Figure 16.20. The transition depth for a circular aperture of radius a is

$$z_t = a^2/\lambda \tag{16.42a}$$

For a rectangular aperture, the transition distance for an aperture L_x in the x–z plane is

$$z_t \approx L_x^2/(\pi\lambda) \tag{16.42b}$$

The natural focal length is the distance to the last axial peak and is approximately the transition distance.

The far-field beam pattern for a rectangular aperture is the Fourier transform of the amplitude across the aperture. In the case of uniform illumination

$$A(x_0, y_0, 0) = \prod(x_0/L_x)\prod(y_0/L_y) \tag{16.43a}$$

where

$$\prod(x/L) = \begin{Bmatrix} 0 & |x| > L/2 \\ 1/2 & |x| = L/0 \\ 1 & |x| < L/2 \end{Bmatrix} \tag{16.43b}$$

the far-field pattern in the x–z plane is a sinc function

$$p(x, z, \omega) = \frac{L_x\sqrt{p_0}}{\sqrt{\lambda z}} e^{i\pi/4} \frac{\sin(\pi_x x/\lambda z)}{(\pi L_x x/\lambda z)} = \frac{L_x\sqrt{p_0}}{\sqrt{\lambda z}} e^{i\pi/4} \mathrm{sinc}\left(\frac{L_x x}{\lambda z}\right) \tag{16.44a}$$

A plot of this pattern is shown in Figure 16.22.

In the case of a uniform amplitude u_0 on a circular aperture, the far-field pattern is the two-dimensional Fourier transform of the circularly symmetric aperture function

$$p(\overline{\rho}, z, \lambda) \approx \frac{ip_0\pi a^2 2J_1(2\pi\overline{\rho}a/(\lambda z))}{2\pi\overline{\rho}a/(\lambda z)} = ip_0\left(\frac{\pi a^2}{\lambda z}\right) jinc\left(\frac{\overline{\rho}a}{\lambda z}\right) \tag{16.44b}$$

where J_1 is the Bessel function of the first kind, $jinc\ (x) = 2J_1\ (2\pi x)/(2\pi x)$, and $\overline{\rho}$ is the radial distance to an observation point at $(\overline{\rho}, z)$. A plot of this pattern is shown in Figure 16.22. Note that the shapes of the far-field patterns are maintained with distance as their amplitudes fall and beams broaden with distance.

From these far-field patterns, it is easy to determine the FWHM beam widths. For the rectangular aperture in the x–z or y–z plane

$$\mathrm{FWHM} = 1.206\lambda z/L \tag{16.45a}$$

where L is the appropriate aperture for that plane. Similarly, for a circular aperture

$$\mathrm{FWHM} = 0.7047\lambda z/a \tag{16.45b}$$

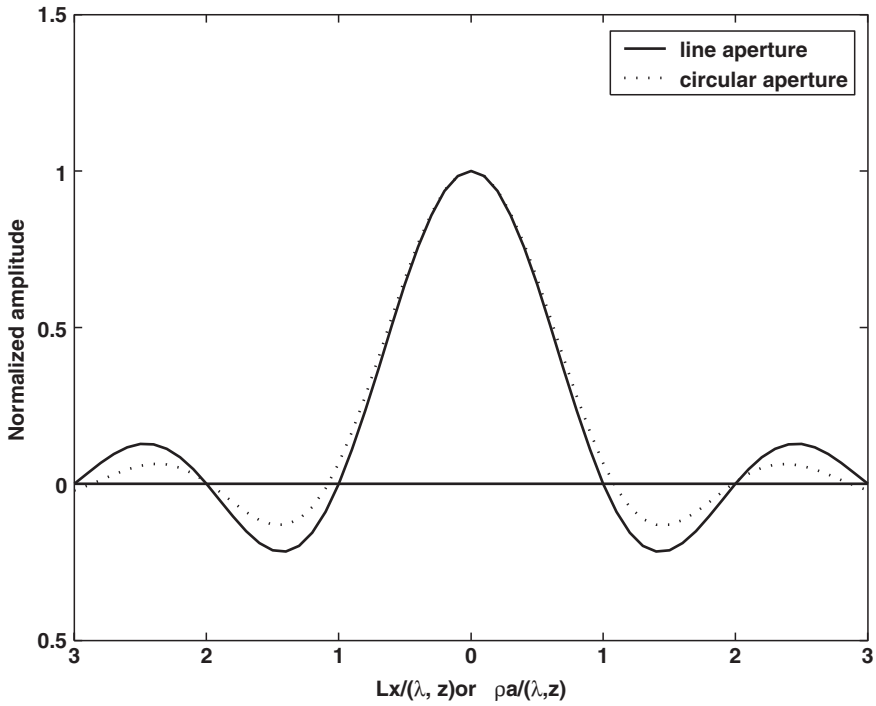


FIGURE 16.22 (a) Far-field jinc beam shape from a circular aperture normalized to (b) far-field sinc function from a square aperture with the same aperture area.

EXAMPLE PROBLEM 16.9

To understand the normalization in Figure 16.22, find the FWHM for a circular aperture having the same area as a square aperture and compare it to the FWHM of the square aperture.

Solution

Set $\pi a^2 = L^2$, or a(m) = L(m)/$\sqrt{\pi}$. Substitute in Eq. (16.45b), FWHM = $0.7047\lambda(m)z(m)$ ($\sqrt{\pi}/L(m)$ = $1.249\lambda(m)z(m)/L(m)$. This is close to the value for the square aperture from Eq. (16.45a), FWHM = $1.206\lambda z/L(m)$.

To narrow the beams even more and at different depths, geometric focusing is applied. Like optics, acoustic focusing can be implemented with a type of lens. Unlike optics, both concave and convex converging lenses can be made because materials exist such that their sound speeds are either greater or less than that of water (tissue). Under the principles of ray optics, the rays converge at the geometric focal point, F. From the reciprocal law of lenses, the overall total focal length is the combined effect of the natural focal length and the geometrical focal length

$$1/F_{\text{total}} = 1/F_N + 1/F \tag{16.46}$$

This relationship shows that the location of the axial peak for a focusing aperture is now moved in from the geometrical focal length. For example, if the natural focal length is 100 mm and the geometric focal length is 50 mm, the overall effective focal length is 33.3 mm. The shape of beams in the focal plane is the same as the far-field pattern of an unfocused beam; consequently, Eqs. (16.44)–(16.45) can be applied with $z = F$.

So far, solid apertures have been described. Arrays can also be considered to be spatially sampled apertures. To first order, the beams for both a solid and an adequately sampled aperture are similar. The one-dimensional array actually has two types of focusing, as shown in Figure 16.23. The azimuth or scan plane, here the x–z plane, is focused electronically, whereas the elevation or yz plane is focused by a fixed mechanical lens.

16.2.7 Ultrasound Imaging Systems

The formation of an image can be understood through the operation of an imaging system. In Section 16.2.1, an imaging system was introduced as having a transducer that could be scanned either mechanically or electronically to produce a pattern of sequential lines to form an image. With reference to Figure 16.24, the management of an array imaging system is accomplished by a computer or central processing unit (CPU). Once the scan depth and mode are selected, transmit pulses, each repeating at the time interval for a line, are sent in synchronism with a master timing clock like those in Figure 16.1. Each of these pulses initiates a set of transmit pulses from the transmit beamformer that are sent to each element, each one of which is delayed as necessary to form an electronic lens for focusing and steering the acoustic beam for the selected line direction in the azimuth plane.

Sound is scattered from tissue interfaces and inhomogeneities and is picked up as a series of pulse echoes by the array acting as a receiver. Depth-dependent time gain compensation (TGC) amplifiers can be controlled by the user to improve the image uniformity at different depths. The receive beamformer applies electronic dynamic focusing (nearly

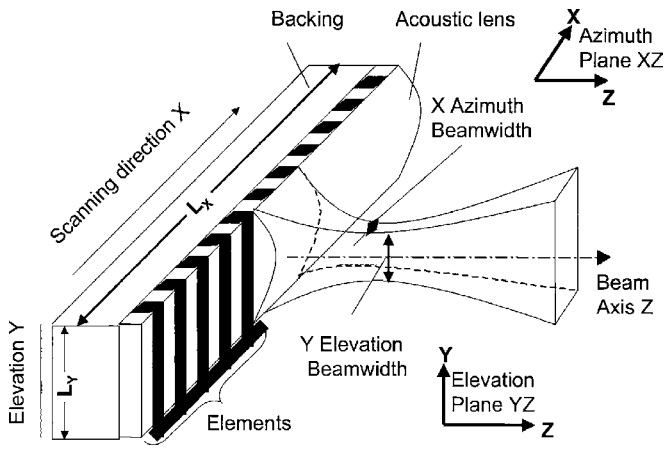


FIGURE 16.23 Focusing of a one-dimensional phased array in both the azimuth and elevation planes. *Adapted from [6].*

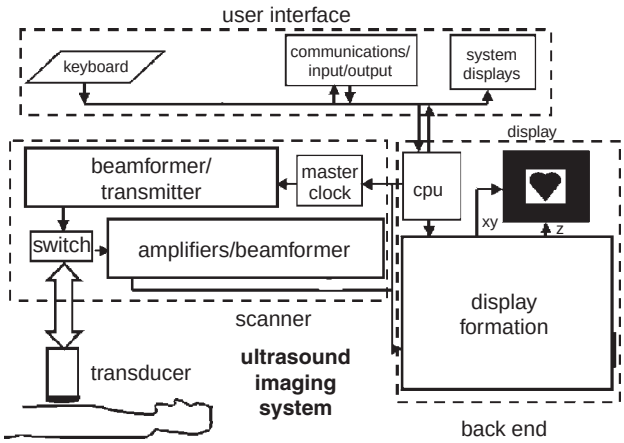


FIGURE 16.24 Block diagram of a digital ultrasound imaging system.

perfect focusing at each depth point), and signal processing (often in terms of filtering) provides the signals needed for the particular selected mode. For ordinary B-mode imaging, the envelopes of the pulse echoes are extracted, and lines containing the echoes are scan converted, the process of interpolating the lines containing pulse-echo envelopes into a filled-out, grayscale image that can be displayed on conventional PC or TV screens such as the images in Figures 16.15 through 16.17. Additional steps such as log compression are used to improve the range of pulse-echo amplitudes visible on the screen. The images in Figures 16.15 and 16.16 are of the most common type: B-mode or brightness mode.

16.2.8 Imaging and Other Modes

Other modes can supply additional information, especially about blood flow. The Doppler effect takes advantage of the apparent change in the ultrasound frequency caused by the velocity v of the blood and the angle θ the transducer makes with the vessel, as seen in Figure 16.25.

The shift in frequency from the transmitted one, f_0, or the classic Doppler shift frequency, f_D, can be expressed as

$$f_D = [2(v/c_0)\cos\ \theta]f_0 \tag{16.47}$$

where c_0 is the sound speed of the intervening medium. Scattering from blood is mainly from groups of red blood cells. In a vessel there is a distribution of velocities, so what is displayed is a Doppler spectrum containing frequencies corresponding to the range of sound speeds insonified. This scattering is usually not visible by ordinary B-mode imaging but is detectable by sensitive ultrasound Doppler instrumentation. Doppler shifts are either detected along the length of an acoustic line selected to intersect a vessel of interest as continuous wave (CW) Doppler, or a small time interval is placed over the region of interest by pulsed wave (PW) Doppler as shown in the image of Figure 16.17.

A more global view of blood flow, especially for the chambers of the heart and larger blood vessels, can be obtained by color flow imaging (CFI) mode. While lacking the

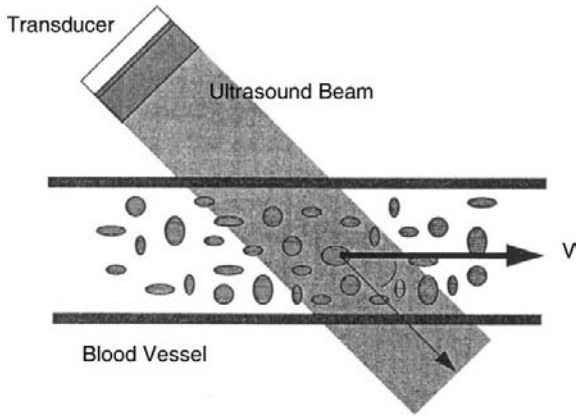


FIGURE 16.25 Sound beam intersecting blood moving at velocity v in a vessel tilted at angle θ.

precision of the Doppler modes, color flow imaging provides a real-time view of the approximate velocity and direction of blood flow. This information is represented by mapping of colors assigned to the velocity magnitude and direction (toward or away from the transducer). In this mode, several acoustic lines are sent in the same direction to obtain flow information over time, and then signal processing extracts the velocity information for display. A second type of Doppler image is called power Doppler, in which the amplitude of the flow is presented but not the direction, as illustrated by the top of Figure 16.17.

Most ultrasound imaging is two-dimensional in that a picture is created in an imaging or scan plane. A three-dimensional (3D) image can be created by scanning a volume rather than a plane. By mechanically moving the array in a direction perpendicular to the imaging planes, pausing long enough to acquire each image plane, and interpolating between planes, it is possible to fill in the overall volume with image data. The 3D images can be viewed in selected cut planes through the volumes or through surface or volume rendering. For example, a common application is 3D imaging of the fetus through surface rendering, in which the boundary between the fetal skin and amniotic fluid is used to create a 3D opaque surface (Figure 16.26). Volume rendering allows internal details of organs to be visualized through transparent boundaries between organs and layers. Images can also be formed through other ways of creating a set of volume data by different mechanical scanning techniques or by complete electronic scanning by a stationary two-dimensional array.

16.3 MAGNETIC RESONANCE IMAGING

16.3.1 Introduction

Magnetic resonance has been applied successfully to medical imaging of the body because of its high water content. The hydrogen atoms in water (H_2O) and fat make up approximately 60 percent of the body by weight. Because there is a proton in the nucleus of each hydrogen atom, as the nucleus spins, a small magnetic field or moment is created.

FIGURE 16.26 3D ultrasound image of a surface-rendered fetal head. *Courtesy of Philips Medical Systems.*

When hydrogen is placed in a large static magnetic field, the magnetic moment of the atom spins around it like a tiny gyroscope at the Larmor frequency, which is a unique property of the material. For imaging, a radiofrequency rotating field in a plane perpendicular to the static field is needed. The frequency of this field is identical to the Larmor frequency, and once the atom is excited, the applied field is shut off and the original magnetic moment decays to equilibrium and emits a signal. This voltage signal, detected by the same coils used for the applied field, and two relaxation constants are sensed. The longitudinal magnetization constant, T_1, is more sensitive to the thermal properties of tissue. The transversal magnetization relaxation constant, T_2, is affected by the local field inhomogeneities. These constants are used to discriminate among different types of tissue and for image formation. T_1 weighted images are used most often.

Today, MRI finds widespread application in the detection of disease and surgical planning. MR images are highly detailed representations of internal anatomy. These may be called parameterized images because considerable skill is involved in adjusting the instrument to obtain images that emphasize different types of tissue contrast, the discrimination among different organ types and between healthy and pathological tissues. MRI is used to examine most of the body, including the brain, abdomen, heart, large vessels, breast, bones, as well as soft tissue, joints, cartilage, muscle, and the head and neck. It is used for both children and adults and for detecting cancer pathologies, tumors, and hemorrhaging.

An early precedent to MRI was nuclear magnetic resonance (NMR), first observed as a phenomenon by Felix Bloch and Edward Purcell and their coworkers. They discovered that

not only did precessing nuclei emit a radiofrequency (rf) signal, but a radiofrequency could also be used to control precession at the Larmor resonant frequency, and, once stopped, the nuclei would emit a detectable RF signal at the same frequency. They won a Nobel Prize in 1951 for their work.

Interest shifted to determining composition of materials through unique frequency shifts associated with different chemical compounds. Eventually, biological NMR experiments were under way, and soon detailed spectral information from phosphorus, carbon, and hydrogen nuclei were obtained. Specialized magnets were designed to accommodate parts of the body for study.

Paul Lauterbur was one of the first to realize that images could be made using NMR principles. He published an image of a heterogeneous object in 1973 [4]. Using the rf signals from NMR, he was able to localize them in space by changing the magnetic field gradient. By the mid- and late 1970s, early MR images were produced of animals and the human body. At first, because the signals were so weak, these results were regarded as a laboratory curiosity. In 1971, Raymond Damadian demonstrated that the relaxation constants, T_1 and T_2, differed for malignant tumors and normal tissue. Peter Mansfield developed a mathematical model to analyze signals from within the human body in response to a strong magnetic field, as well as a very fast imaging method. Continuous research spurred the evolution of modern MRI instruments with high signal to noise and generated a considerable knowledge base of how to apply MRI to diagnostic imaging. Lauterbur and Mansfield shared the 2003 Nobel Prize for medicine for their MRI discoveries.

16.3.2 Magnetic Fields and Charges

To understand how MRI works, several relevant characteristics of magnetic fields are reviewed here, in particular, the interactions between electrical charges and magnetic fields. Einstein pointed out that it is useful to consider electricity and electromagnetic fields as aspects of the same energy.

Four cases will be covered, each one useful for providing insights into aspects of MRI processes. In the first case, a magnetic field is generated when an alternating current travels along a wire. For an infinitely long wire, the Biot–Savart law reveals that a circular or circumferential magnetic field flux is generated by the current, as shown in Figure 16.27,

$$B_\phi = \frac{\mu_0 I}{2\pi r} \tag{16.48}$$

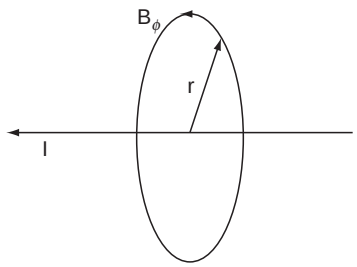


FIGURE 16.27 Circular magnetic field generated by an electrical current flowing down an infinitely long wire.

where the magnetic flux B is in units of Webers (1 Tesla(T) $= 10^4$ Webers/m^2), $\mu_0 = 4\pi 10^{-7}$ Henry/m [Weber/(amp-meter)] is the permeability of free space, I is current in Amperes, and r is the radial distance from the wire in meters. The right-hand rule applies: As the fingers curl about the direction of the B field, the thumb points in the direction of the current.

EXAMPLE PROBLEM 16.10

Calculate the magnetic field in T at 5 mm from a wire carrying 20 A of current.

Solution

Using Eq. (16.48),

$$B_\phi = \frac{4\pi e - 7 \times 20}{2\pi \times 5e - 3}\ (\text{Weber/m}^2) \times 1\ \text{T}/10^4\ (\text{Weber/m}^2) = 8 \times 10^{-6}\ \text{T}$$

If the wire is coiled into a circular loop, a magnetic dipole with north and south poles is created. As a second case, an equivalent situation is created by a rotating charge, as shown in Figure 16.28. Current I, flowing along an increment of wire dl, in a loop is equivalent to a charge q, of mass m, orbiting at a frequency, v. The magnetic dipole moment is the product of the equivalent current and area at a large distance r,

$$\hat{\mu} = (qvA)\hat{z} = qvpr^2\ \hat{z} = \frac{1}{2}qr^2\omega\hat{z} \tag{16.49}$$

where the direction of $\hat{\mu}$ is along unit vector $\hat{z}$ according to the right-hand rule. A vector is a quantity that has a magnitude and a direction; in this case, a unit vector has a magnitude of one and is directed along the z-axis. If the mass of the charge is m, the orbital angular momentum is

$$\mathrm{L} = mr^2\omega \tag{16.50a}$$

The classic gyromagnetic ratio is defined as

$$\gamma_c = \frac{\mu}{L} = \frac{q}{2m} \tag{16.50b}$$

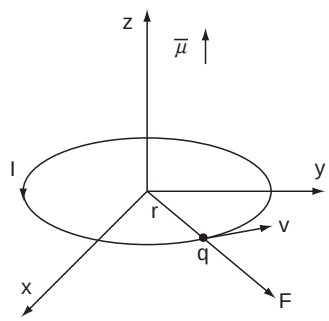


FIGURE 16.28 Magnetic dipole moment of a charge in a circular orbit.

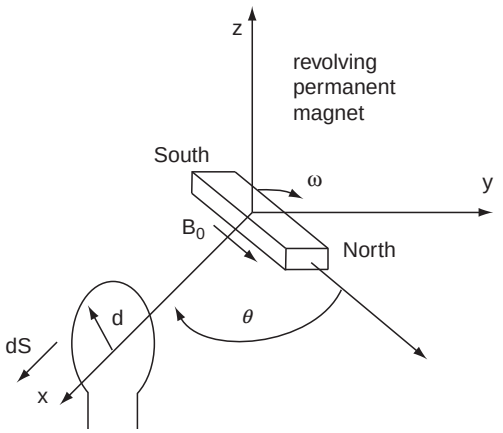


FIGURE 16.29 Voltage induced in wire loop of radius d by a revolving permanent magnet of strength B_0.

This overall result implies that an orbiting charge can act like a small magnet with its own north and south pole.

Can a moving magnetic field create a current or voltage in a wire? For the third case, consider the arrangement in Figure 16.29, where a wire loop of radius d is perpendicular to the x-axis and where a permanent magnet of strength B_0 is whirling about the z-axis at a constant angular frequency ω. The angle between the x-axis and the magnet axis can be described as $\theta = \omega t$. Then, Faraday's law specifies that the voltage created in the loop by the spinning magnet can be written in terms of the electric field E around the loop. The area of the loop is $S = \pi d^2$, and its vector is perpendicular to the loop. If the field rotating relative to the x-axis is $B_0 \cos \omega t$ and has a direction along vector B, then according to Faraday's law,

$$V = -\frac{d}{dt}(B_0 \pi d^2 \cos \omega t) = -\omega B_0 \pi d^2 \sin \omega t \tag{16.51}$$

The voltage picked up from the rotating magnet is sinusoidal and is maximum when the axis of the magnet is perpendicular to the plane of the loop and is zero when the axis is parallel.

For the fourth case, a whirling charge is placed in a strong static magnetic field B_0, as shown in Figure 16.30. Here, the action of the field on the charge exerts a force on the charge described by the Lorentz force equation

$$\hat{F} = q\hat{v} \times \hat{B}, \; or |F| = q|v||B| \sin \theta \tag{16.52a}$$

where this vector cross-product notation means the velocity vector $\hat{v}$ is tangential to the orbit at the position of the charge and the force is exerted outward perpendicular to both $\hat{v}$ and the applied field direction, $\hat{B}_0$ and the angle between $\hat{v}$ and $\hat{B}_0$, $\theta = 90°$. The magnitude of this force can be rewritten for this case as

$$|F| = q|v||B| \sin\theta = mv^2/r \tag{16.52b}$$

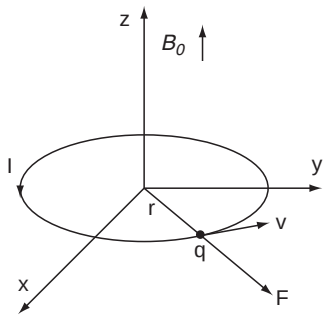


FIGURE 16.30 Lorentz force on a charge revolving about a static magnetic field. Vectors for velocity, v, force, F, and magnetic field, B_0, are shown at the position of the charge.

The classical angular frequency of the charge can be expressed in terms of the velocity of the charge and the radius, which can be rewritten from the previous equation to give

$$\omega_c = v/r = v/(mv/qB_0) = (q/m)B_0 \tag{16.53a}$$

From the definition of the classic gyromagnetic ratio, Eq. (16.50), comes an important equation in MRI for the classical frequency,

$$v_c = \frac{\omega}{2\pi} = \frac{(2\gamma_c)B_0}{2\pi} = \frac{\gamma_c B_0}{\pi} \tag{16.53b}$$

which shows that the orbital frequency of the charge is proportional to the applied magnetic field. Unfortunately, this is not exactly what is needed for MRI because classical electromagnetic theory is for a charge that does not revolve on its own axis. The charge of interest in MRI is for an electron, which has its own individual spin. This situation is analogous to the revolution of the earth around the sun in combination with the revolution of the earth about its own axis. To obtain this important equation, an explanation of spin states is necessary from quantum mechanics.

16.3.3 Spin States

Based on the previous discussion, one could expect that the electron spinning on its own axis would create a miniature magnetic field; consequently, it would behave like a magnetic dipole with its own north and south poles.

Permanent bar magnets are dipoles that have a strong polarization in the form of north and south poles. If two equal permanent bar magnets are placed in the north to south setup shown in Figure 16.31a, they are strongly attracted and are said to have a strong attractive force between them. If they are placed close to each other in a north–north (or a south–south) configuration as in Figure 16.31b, the magnets are forced apart by a strong repulsive force. These two arrangements of magnets are two positions in which the strongest forces are stabilized in equilibrium.

If a number of the tiny magnetic dipoles are placed in a strong static magnetic field, B_0, then they will align either with the direction of the field (parallel) or lock into a position

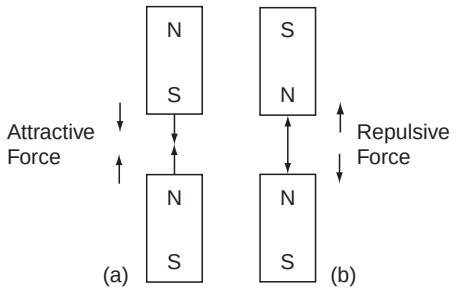


FIGURE 16.31 Attractive (a) and repulsive (b) magnetic forces for two arrangements of permanent magnets with north (N) and south (S) poles.

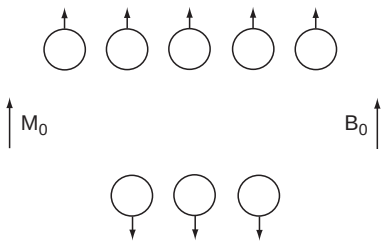


FIGURE 16.32 Alignment of magnetic dipoles in a static magnetic field B_0 and the net magnetization.

opposing the field (antiparallel), as shown in Figure 16.32. Most of them will line up along the applied field because less energy is required to maintain that orientation. The overall vector summation of the individual dipole moments results in an overall net magnetization

$$\hat{M}_0 = \sum_n \hat{\mathrm{u}}_n \tag{16.54}$$

as shown in the figure.

The energy required for the transition from one spin state to another is governed by

$$\Delta E = \hbar v \tag{16.55}$$

in which Plank's constant is $\hbar = 6.626 \times 10^{-34}$ *J-s*. The physical meaning of this relation is that a photon of frequency v is either absorbed to send an electron to a higher energy state or it is radiated for the electron to fall to a lower state. Eventually, excited electrons return to their original equilibrium state. Unlike x-ray imaging that emits ionizing radiation, only harmless photons are emitted during this mechanism that is used in MRI.

What is the frequency necessary for this transition? Returning to the last case of a charge revolving about the nucleus in a static magnetic field, the spin gyromagnetic ratio, γ, from quantum mechanics for a spinning electron is q/m. The orientation of the spin results in a left-hand rule (clockwise about B_0). Like Eq. (16.53b), substitution for frequency equation yields

$$v_L = \frac{\omega}{2\pi} = \frac{\gamma}{2\pi} B_0 = \gamma' B_0 \tag{16.56}$$

TABLE 16.2 Characteristics of Isotopes Relevant to MRI

Isotopes	Spin State	γ' (MHz/T)	Natural Isotropic Abundance %	Sensitivity Relative to ^{1}H
^{1}H	1/2	42.58	99.99	100
^{13}C	1/2	10.71	1.11	1.6
^{19}F	1/2	40.05	100	3.4
^{23}Na	3/2	11.26	100	9.3
^{31}P	1/2	17.24	100	6.6

which is known as the Larmor frequency that can be used to calculate the transition energy. Note that the gyromagnetic constant, γ, is associated with angular frequency, and its normalized version, γ', with frequency. This is the important equation that connects frequency to an applied magnetic field. For the hydrogen isotope most often used in MRI, $\gamma' = 42.58\,\text{MHz/T}$. Fortunately, the values of parameters needed for MRI fall within reasonable ranges: typical frequencies are under 100 MHz and fields are 0.1 to 4 T.

The Larmor frequency is a resonance excitation frequency used for externally producing transitions in spin states. For MRI, only isotopes that have unpaired nuclear spins that are multiples of ½ are of interest. For example, the spin state for ^{1}H aligned with the applied field is assigned the name "½," and has a dipole moment of $\mu_z = \gamma\hbar \times (+½)$. The antiparallel state for hydrogen has a spin designation of "–½" and has a moment of $\mu_z = \gamma\hbar \times (–½)$. In Table 16.2, values of γ' are given for different isotopes. In general, $\mu_z = \gamma\hbar m$, in which m can include any of the following values, $m = 0, \ldots \pm (I - 1), I$, where I is the spin-state number in Table 16.2.

How many of these spin states are there? If the number of spins in the low-energy state is n_- and the number of those in the higher state is n_+, then Boltzmann statistics predict the ratio of spins in either of these two states at any given time

$$n_-/n_+ = \exp\left(-\Delta E/\underline{K}T\right) \tag{16.57}$$

where the Boltzmann constant is $\underline{K} = 1.3805 \times 10^{-23}\ J/K$, ΔE is from Eq. (16.55), and T is temperature in °Kelvin (K).

EXAMPLE PROBLEM 16.11

Find the fraction of excess population in the spin-up (n_+) state for a frequency of 20 MHz and a temperature of 300 K.

Solution

Calculate the change in energy for a given frequency of 20 MHz from Eq. (16.55):

$$\Delta E = \hbar v = 6.626 \times 10^{-34} J - s \times 20 \times 10^6/s = 1.325 \times 10^{-26} J$$

From Eq. (16.57), the excess population can be written as

$$(n_+ - n_-)/n_+ = 1 - n_-/n_+ = 1 - \exp(-\Delta E/\underline{K}T) \tag{16.58}$$

Substituting the necessary values,

$$\begin{aligned}(n_+ - n_-)/n_+ &= 1 - \exp[-1.325 \times 10^{-26} J/(1.3805 \times 10^{-23} J/\mathrm{K})300\ \mathrm{K}] \\ &= 3.2 \times 10^{-6}\end{aligned}$$

This is a small number based on the material under examination being entirely of one type. In reality, certain types of isotopes occur more commonly than others. The ratio of one type of isotope to the total number available in percent is called the natural isotropic abundance and is listed in Table 16.2. A third relevant question is how many of these isotopes occur in the human body? The third factor is sensitivity relative to the hydrogen isotope or the equivalent number of nuclei in a field, also listed in Table 16.2. Fortunately, hydrogen is plentiful in the body, especially in fat and water.

EXAMPLE PROBLEM 16.12

Compare the isotopes ^{1}H, ^{13}C, and ^{31}P for imaging. Use Table 16.2.

Solution

Note that ^{1}H is 99.98 percent naturally abundant. ^{13}C is not, and occurs at only 1.11 percent, so this isotope is not suitable. On the other hand, ^{31}P is abundant but is difficult to detect due to its low sensitivity. Because the body is 60 percent water by weight, it is not surprising that ^{1}H, with its high sensitivity and abundance, is usually used for MR imaging.

16.3.4 Precession

To excite hydrogen dipoles into a number of spin states for imaging, an external magnetic field can be applied. The natural inclination for the spin magnetic dipoles to align along the z-axis makes detection by a coil difficult if the coil is placed perpendicular to the x-axis as in the fourth case from Section 16.3.2. It is necessary to find a means to bring the dipoles down into the x–y plane so they can be detected by a coil in a manner analogous to case 4. A force is required to push the dipoles into a precession, a downward spiraling orbit. This mechanism is similar to the action of gravity on a spinning top, which is initially vertical and eventually is tilted by the force of gravity into a precessing orbit and finally into a final horizontal position (Figure 16.33).

At first, the net magnetization vector, $\hat{M}_0$, is aligned with the static magnetic field along the z-axis as shown in Figure 16.33a. The application of a time-varying magnetic field, B_1, along the x-axis at the Larmor frequency causes the magnetization to precess at an angle ϕ about the z-axis at this frequency, as shown in Figure 16.33b.

To clarify what happens to $\hat{M}_0$, next, it is worth introducing a reference frame notation to simplify the description. This frame will have coordinates described by a prime notation,

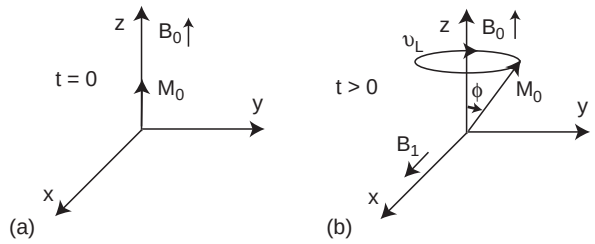


FIGURE 16.33 (a) Initial alignment of magnetization with the static magnetic field B_0. (b) Tipping of the net magnetization from the z-axis through an angle ϕ by the application of the rf B_1 field aligned along the x-axis causes M_0 to spin about the z-axis at the Larmor frequency.

such as $\hat{x}'$. One reason for using this frame is that the trajectories of these spinning, precessing magnetic vectors are complicated, and simplified methods are helpful in visualizing them. Another reason is that a bundle of frequencies is usually involved in these measurements, and it is easier to track their movements relative to the reference frequency v_L.

Rotating Frame

The approach is called a rotating frame, a coordinate frame of reference that rotates with the magnetization vector whirling at a Larmor frequency, v_L, in contrast to a reference point in a fixed Cartesian coordinate system. This methodology is similar to a stroboscopic viewpoint: To make a rotating object appear still, a stroboscopic light flashes in synchronism with each revolution of the object. This light frequency serves as a reference so if the revolving object rotates faster or slower, its deviation from this reference frequency is easily seen. Similarly, a frequency rotating clockwise relative to v_L in a rotating frame is called a positive frequency, and a frequency rotating counterclockwise is called a negative frequency.

Recall case 3 from Section 16.3.2, a description of a configuration where a signal was detected from a permanent magnet spinning around the z-axis. The resulting detected signal was a sinusoidal signal with a frequency ω. Instead of a permanent magnet, consider a time-varying magnetization vector spinning at a Larmor frequency about the z-axis in the x–y plane at a position initially aligned with the x-axis at time t:

$$\hat{B}_x(t) = \hat{x}B_1 \cos \omega t \tag{16.59}$$

To find out what this magnetic field vector would look like in a rotating frame, a coordinate transformation is applied to obtain the rotated components of the field in the new frame, which are part of a new vector B'_1 with components

$$B_{x'} = \cos(\omega t)B_x - \sin(\omega t)B_y \tag{16.60a}$$

$$B_{y'} = \sin(\omega t)B_x + \cos(\omega t)B_y \tag{16.60b}$$

which with the substitution of Eq. (16.58) becomes

$$B_{x'} = B_1 \cos^2 \omega t \tag{16.60c}$$

$$B_{y'} = B_1 \cos \omega t \sin \omega t \tag{16.60d}$$

From trigonometric identities for half-angles, the preceding result simplifies to

$$\hat{B}_1'(t) = \hat{x}'B_{x'} + \hat{y}'B_{y'} = \hat{x}'\left(\frac{B_1}{2} - \frac{B_1}{2}\cos 2\omega t\right) + \hat{y}'\frac{B_1}{2}\sin 2\omega t \tag{16.60e}$$

which, with low-pass filtering to eliminate the components at twice the Larmor frequency, leaves

$$\hat{B}_1'(t) = \hat{x}'\frac{B_1}{2} \tag{16.60f}$$

a component of B_1 that appears to be stationary (not time varying) in the rotating frame, as expected.

Flip Angles and Decay

In MRI, the ability to control the position of the magnetization vector is important. A practical configuration to accomplish this magnetic precession is shown in Figure 16.34 for rotated coordinates. First, the magnetization vector is shown aligned with the static field in Figure 16.34a. Second, a radio frequency (rf) magnetic field B_1 along the x'-axis at the Larmor frequency, v_L, is applied as a tone burst (a gated sinusoid) of pulse time duration, t_p. The application of this pulse creates a force to pull the magnetization $\vec{M}_0$ (usually initially aligned in an equilibrium position along the z-axis) down into a precession angle φ from the z-axis, called the flip or tip angle,

$$\varphi = \gamma\left(\frac{B_1}{2}\right)t_p = \gamma B_1' t_p \tag{16.61}$$

The precessing magnetization vector can also be broken down into its Cartesian components in the rotated frame, as shown in Figure 16.34b. In addition to the component along the z'-axis, $M_{z'}^0$, there are two components in the x'–y' plane with a radial magnitude, $M_{x'y'}^0$. It is convenient to look at the complicated resulting changes in time separately along the two components, $M_{z'}^0$ and $M_{x'y'}^0$, even though they occur simultaneously.

EXAMPLE PROBLEM 16.13

Find the pulse length necessary to achieve a phase rotation of π for the field calculated in Example Problem 16.10. Repeat for a coil that generates a field 100 times greater.

Solution

From Eq. (16.60), solve for t_p using Eq. (16.56), $\gamma = 2\pi\gamma'$, and Table 16.2.

$B_1 = B_\phi = 8 \times 10^{-6}$ T from Example Problem 16.10, and $\gamma = 42.58 \times 10^6$ MHz/T. Solve for t_p in Eq. (16.61):

$$t_p = \frac{\varphi}{2\pi\gamma'(B_1/2)} = \frac{\pi}{2\pi \times 42.58 \times 10^6(/s - T)(4 \times 10^{-6}T)} = 2.936 \text{ msec}$$

For the second case, the pulse is smaller by a factor of 100, or 29.36 microseconds.

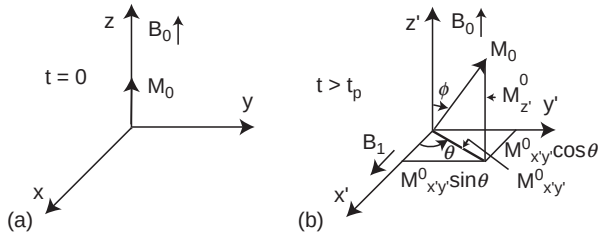


FIGURE 16.34 (a) Initial alignment of magnetization at $t = 0$. (b) Tipping of the net magnetization from the z-axis through an angle ϕ by the application of rf B_1 field at the Larmor frequency aligned along x'-axis in rotated coordinates. The projections of the net magnetization onto the axes are shown.

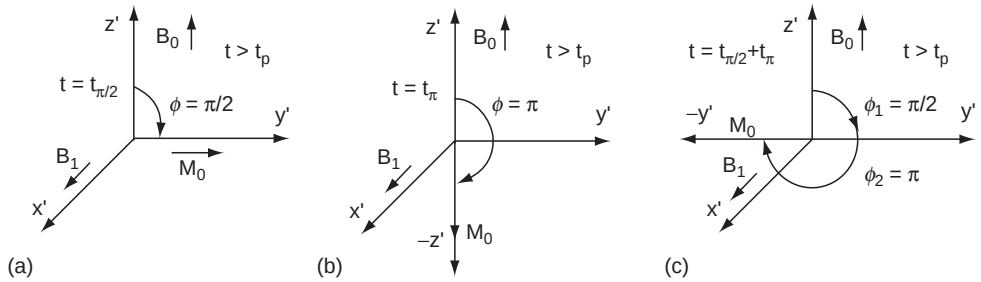


FIGURE 16.35 (a) Tipping through an angle of $\pi/2$ radians to the y'-axis. (b) Tipping through an angle of π radians to the $-z'$-axis. (c) Tipping sequence: first tip through an angle of $\pi/2$ radians to the y'-axis and then tip through an angle of π radians to the $-y'$-axis.

If the flip angle is chosen to be 90° or $\pi/2$ radians, the magnetization is rotated from an initial value of M_0 along the z'-axis to a position lying along the y'-axis, as shown in Figure 16.35a. The vertical component is brought a value of zero in the x'–y' plane and then rebounds back to its original initial value of M_0 over a period of time. This recovery can be described by the following equation and Figure 16.36.

$$M_{z'}(\theta = \pi/2) = M_0[1 - \exp(-t/T_1)] \tag{16.62}$$

where T_1 is the spin lattice recovery time and t starts after the rf pulse. It is also called the longitudinal time constant because of its association with the vertical direction. This relaxation time is affected by temperature and the viscosity of the tissue or material (it is longer as viscosity increases).

If the flip angle is chosen to be 180° or π radians, the magnetization is rotated from an initial value of M_0 along the z'-axis to a position lying along the $-z'$-axis. In this antiparallel orientation, the appropriate relaxation equation is

$$M_{z'}(\theta = \pi) = M_0[1 - 2\exp(-t/T_1)] \tag{16.63}$$

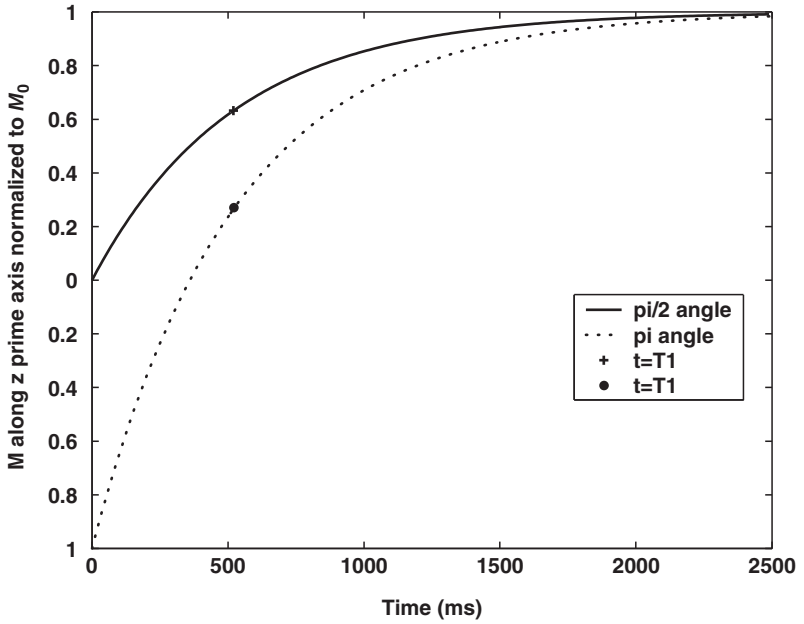


FIGURE 16.36 Plots of the recovery of a magnetization vector component along the z'-axis for a relaxation time T_1,s versus time for a $\pi/2$ rotation, and a π rotation for gray matter of the brain.

This change is illustrated by Figure 16.35b. A generalization of these equations to an arbitrary value of the z– component of magnetization following the rf pulse, M_z^0, is

$$M_{z'}(t) = M_0\left[1 - \left(1 - \frac{M_{z'}^0}{M_0}\right)\exp\left(-t/T_1\right)\right] \tag{16.64}$$

To determine the magnetization components in the x–y plane, refer back to the situation of the $\pi/2$ flip angle where the magnetization is rotated to the y'-axis at time $t = t_{\pi/2}$. In order for equilibrium to be restored, the initial $M_{x'y'}^0$ component at a time t beginning after the rf pulse must decay back to a zero value so the original value of the purely vertical M_0 can be recovered,

$$M_{x'y'}(\theta = \pi/2) = M_{x'y'}^0[\exp\left(-t/T_2\right)] \tag{16.65}$$

in which T_2 is the spin–spin relaxation time as plotted in Figure 16.37. Because of the motion and orientation of this component, its interaction with the material or tissue is different from that of T_1. As Figure 16.38 shows, a dephasing process occurs as vectors sweep over an increasing sector over time so the net sum of vectors gradually goes to zero.

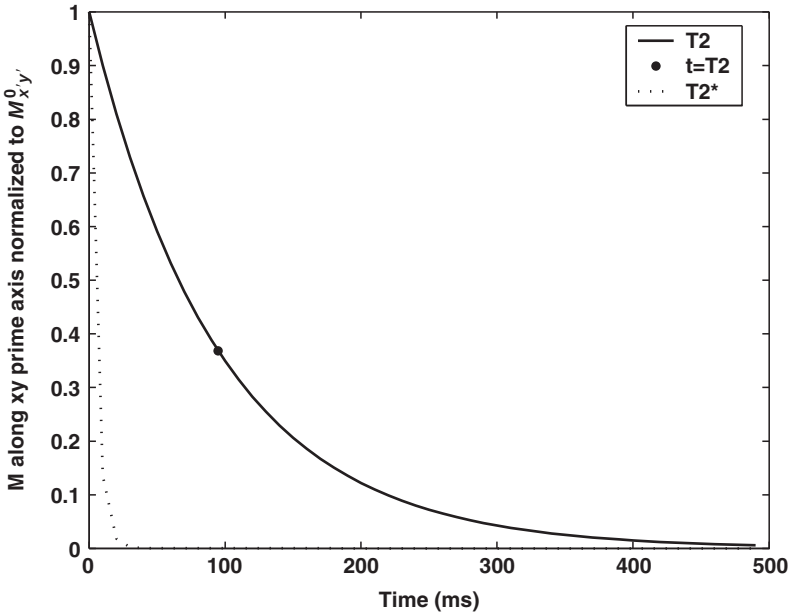


FIGURE 16.37 Plots of the magnetization vector component in the x'–y' for a decay time T_2 versus time for a $\pi/2$ rotation for gray matter and $T_2^* = 5$ ms.

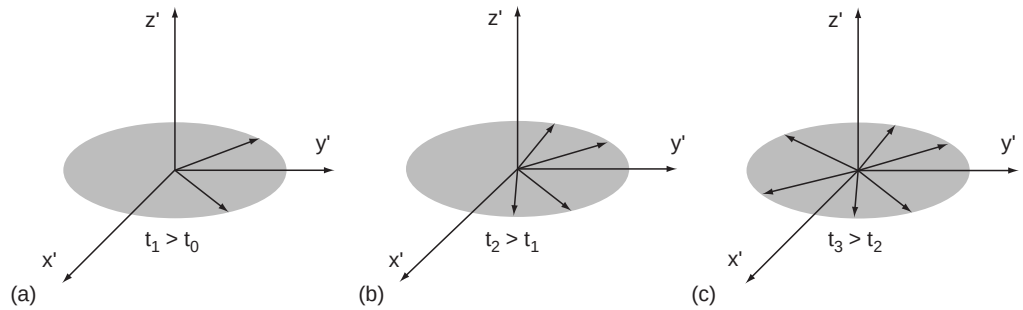


FIGURE 16.38 Time sequence of dephasing plots for spin–spin decay with time constant T_2 after $\pi/2$ rotation pulse occurring at $t = t_p$.

T_2 is sensitive to molecular interactions and inhomogeneities in the applied field B_0. A more realistic definition of this relaxation constant includes the effects of variations ΔB_0:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{\gamma \Delta B_0}{2} \tag{16.66}$$

Through a comparison of Figures 16.36 and 16.37, it is evident that the T_2 constants are much shorter than T_1 and T_2^* is the shortest of all. Typical values for relaxation times can be found in Table 16.3.

TABLE 16.3 T_1 and T_2 Values for Common Tissue Types

Tissue	T_1 (ms)	T_2 (ms)
Cerebrospinal fluid	2,000	1,000
Fat	160	100
Gray matter	520	95
Malignant tumor	800	200
Typical edema or infarction	600	150
White matter	380	85

EXAMPLE PROBLEM 16.14

After a $\pi/2$ rotation, determine the percentage of recovery of the magnetization vertical component associated with T_1 in fat when the horizontal magnetization vector component has decayed to a $1/e$ value.

Solution

From Table 16.3, $T_2 = 100$ ms, which would be the time required for the horizontal magnetization component to decay to a value of $1/e$ value. With the value of $T_1 = 160$ ms and Eq. (16.62), the relative recovery value is

$$1 - \exp(-T_2/T_1) = 0.465$$

Detected Response Waveforms

Because sufficient background information has been given to describe the excitation and decay of a nuclear resonance, it is now possible to describe the kinds of signals detected. A setup like that of Figure 16.33a is used with a detection loop in a plane perpendicular to the x-axis. A $\pi/2$ rf excitation pulse for tipping the field 90° consists of a tone burst t_p long at the Larmor frequency according to Eq. (16.61). Since the duration of the pulse is much longer than any of the recovery and decay time constants, a sinusoidal excitation in an exponential form will be used to simplify analysis. The resulting transverse magnetization vector in the x–y plane is

$$\hat{M}_{xy}(t) = M_{xy}^0 e^{-|t|/T_2} e^{j\omega t}(\hat{x} + \hat{y}) \qquad t < t_p \tag{16.67}$$

The effect of this excitation is to cause a resonance and a change in the spin states of hydrogen electrons. Faraday's law can be applied to the detection of the overall net magnetization response from Section 16.3.2, case 3. The area S of a loop of diameter d can be rewritten in terms of magnetization M and an equivalent current I:

$$S = \frac{M}{I} \tag{16.68}$$

If the direction of the area is taken to be the x'-axis and the corresponding component of M is used in Faraday's law, Eq. (16.51), for the cubic volume, V_{voxel}, of an image voxel

$$v(t) = -V_{\text{voxel}} \frac{\partial}{\partial t} \left[\left| B_1'(t) \times \frac{M_{xy}}{I} \right| \right]$$
$$v(t) = -V_{\text{voxel}} \frac{\partial}{\partial t} \left[\left| B_1'(t) \times \frac{\hat{M}_{xy}}{I} \right| \right] = -V_{\text{voxel}} \frac{\partial}{\partial t} \left[\left| B_1'(t) \right| \left| \frac{M_{xy}}{I} \right| \sin\theta \right] \tag{16.69a}$$

From Eq. (16.67) for M_{xy} and Eq. (16.60f) for $\hat{B}_1'(t)$, and since the angle between $\hat{B}_1'$ and $\hat{M}_{xy}$ is $\theta = 90^\circ$, the voltage becomes

$$v(t) = \left(\frac{B_1}{2I} \right) \cdot \left(j\omega V_{\text{oxel}} M_{xy}^0 e^{-|t|/T_2} e^{j\omega t} \right) \tag{16.69b}$$

where V_{oxel} is the cubic volume of an image voxel and the second-order terms in dM/dt are dropped. This detected voltage is called the free induction decay (FID), the real part of which is shown in Figure 16.39.

The major reason the FID waveform is unsuitable for detection is that it is severely affected by variations in the magnetic field that dominate the response over the signals of the desired spin states. Also, the waveform weakens rapidly with the shorter time decay constants. As shown in Figure 16.38, spin states experience different magnetic fields because of their spatial location; therefore, the angular frequencies of their magnetization vectors vary, some moving faster than others, according to Eq. (16.56).

A clever alternative is called the spin echo method. If a certain time, $T_e/2$, elapses so that the spin vectors are severely but not totally dispersed, as shown in Figure 16.35c, a second π pulse with an amplitude twice that of the original $\pi/2$ pulse rotates the leading vectors back to the $+y$–axis, as shown in Figure 16.40a. The individual magnetization vectors rotate in an opposite direction toward realignment. As the vectors draw into coherent phase, the detected signal increases to an alignment peak at T_e and then decays, as shown in the spin echo waveform in Figure 16.40b. The beauty of this approach is that the effects of magnetic field variations are canceled out, resulting in a truer $T2$ tissue response. Following a derivation similar to that of the FID waveform, we can obtain an expression for the spin echo:

$$v(t) = j(B_1/2I)\omega V_{\text{oxel}} M_{xy}^0 e^{-|t/T_2|} e^{jwt} \tag{16.70}$$

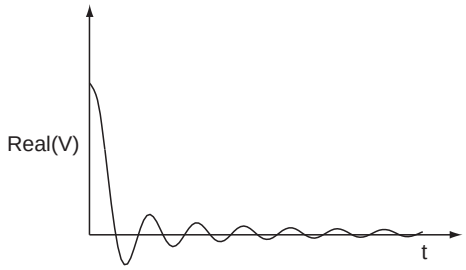


FIGURE 16.39 Free induction decay waveform versus time. This waveform starts at $t = 0$.

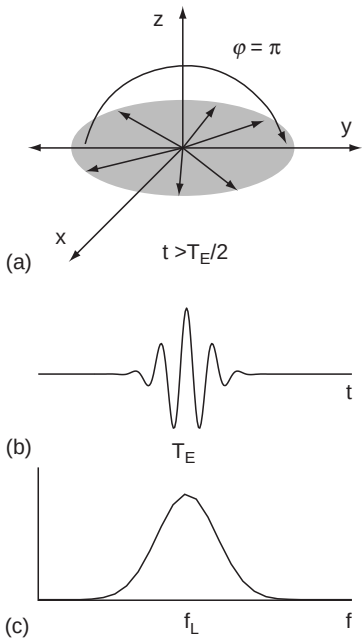


FIGURE 16.40 (a) A π pulse rotation causes a reversal in the rotation direction of the magnetization vectors so they move toward realignment: (b) spin echo waveform versus time, and (c) spin echo spectral magnitude versus frequency.

To find the spectrum of the spin echo, we use the following transform pair

$$\Im[\exp(-|t|)] = \frac{2}{1+(2\pi f)^2} \tag{16.71a}$$

along with the transform pair for the exponential (note in Eq. (16.70), $\omega = \omega_L = 2\pi f_L$, a constant, the electrical frequency corresponding to the physical Larmor frequency),

$$\Im(e^{j\omega_L t}) = \delta(f - f_L) \tag{16.71b}$$

and the scaling theorem, Eq. (16.4a), to obtain the echo spectrum of Eq. (16.70)

$$V(f) = \Im\left(\frac{jB_1 M_{xy}^0 \omega_L V_{\text{oxel}} e^{-|t/T_2|} e^{j\omega_L t}}{2I}\right) = \frac{jB_1 M_{xy}^0 \omega_L V_{\text{oxel}} T_2}{2I}\left\{\frac{1}{1+[2\pi T_2(f-f_L)]^2}\right\} \tag{16.71c}$$

and the magnitude of which is shown in Figure 16.39c. If a function $S(f)$ is defined as

$$S(f) = \frac{jB_1 M_{xy}^0 \omega_L V_{\text{oxel}} T_2}{2I}\left\{\frac{1}{1+[2\pi T_2 f]^2}\right\} \tag{16.71d}$$

then the right-hand side of Eq. (16.71c) becomes

$$V(f) = S(f - f_n) \tag{16.71e}$$

in which $f_n = f_L$ for this case. In general, the index n can be associated with the excitation Larmor frequency at a location x_n. How this mapping is carried out is the subject of the next section.

16.3.5 Setup for Imaging

To create images from detected signals, a means of spatially localizing the detected waveforms must be used. Three methods are used to encode the waveform data and position, one for each Cartesian coordinate. The three-dimensional image is discretized into tiny cubes called voxels, each with a volume $V_{\text{oxel}} = \Delta x \Delta y \Delta z$.

The first step in setting up the image is to fix the desired location of the slice plane, one voxel thick, along the z-axis, as shown in Figure 16.41. The person to be imaged is slid into a large superconducting magnet that creates a strong static field, B_0. Recall from cases 1 and 2 from Section 16.3.2 that a current in a wire or loop can generate a magnetic field. Electrical coils are arranged within the magnet to create an electrically controlled linear magnetic field gradient along the z-axis

$$\vec{B}_z = \hat{z}(B_0 + G_z z) \tag{16.72a}$$

where G_z is a gradient constant in T/m. From Eq. (16.56), a unique Larmor frequency is associated with each spatial location z,

$$\omega = \gamma B_z = \gamma B_0 + \gamma G_z z \tag{16.72b}$$

The relative position $z = 0$ is called the isocenter. The reference frequency associated with this center is $\omega_0 = \gamma B_0$, which can be subtracted from ω by electronic mixing.

By applying a time excitation pulse with an appropriate center frequency as shown in Figure 16.41, a specific location is selected. The envelope of the sinusoid used for time

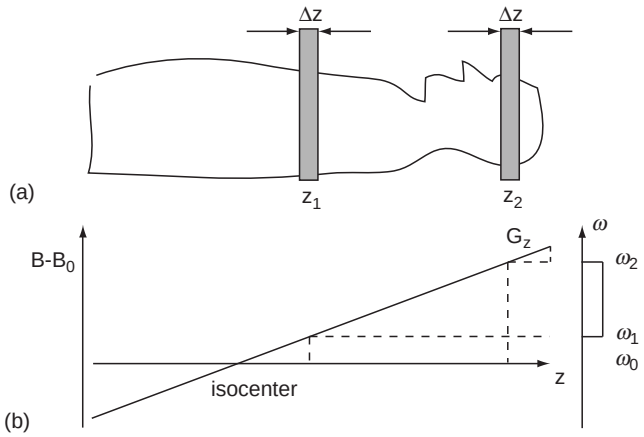


FIGURE 16.41 (a) Side view of selected axial image plane locations. (b) Two regions corresponding to selected locations on the z-axis magnetic field gradient. On the right vertical axis are regional gradient spectra from rf excitation pulses whose Larmor center frequencies are selected for z-axis positions. Scale is exaggerated for clarity.

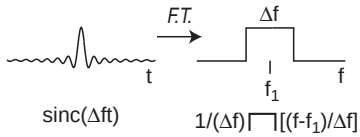


FIGURE 16.42 Sinc modulated excitation pulse with a selected Larmor center frequency and its Fourier transform representing its spectrum or the shape of the applied field.

domain rf pulse excitation is shaped to obtain a flatter field. A sinc-shaped pulse would have a flat rectangular spectral shape, a consequence of its Fourier transform relation, but, ideally, it would have to be infinitely long. This pulse and the corresponding field-spectral shape are shown in Figure 16.42. A sinc-time pulse has a Fourier transform, $\Im$, that can be expressed as a rect function of width Δf:

$$\Im[\sin c(\Delta ft)\exp(j2\pi f_1 t)] = (1/\Delta f)\Pi[(f - f_1)/\Delta f] \tag{16.73}$$

As a compromise, a truncated or tapered or modified sinc pulse of finite length is applied instead, such as the tapered shape shown in Figures 16.40b and 16.40c.

EXAMPLE PROBLEM 16.15

Find the bandwidth needed to achieve a z-slice thickness of 10 mm. Assume $G_z = 0.352\ T/m$. For $B_0 = 1$T, determine the frequencies needed to move 10 cm and 20 cm from the isocenter.

Solution

From a version of Eq. (16.72b) with γ', the difference in frequency is proportional to thickness, or

$$\Delta f = 42.58(\text{MHz}/T) \times 0.352(T/m) \times .010\ m = 0.15\ \text{MHz}$$

The same equation can be used with a value of the offset from the isocenter, $f_1 = 42.58$ (MHz) $[1T + 0.352\ (T/m) \times 0.1\ m] = 44.08$ MHz. Similarly, $f_2 = 45.58$ MHz corresponds to the frequency needed to move 20 cm from the isocenter.

In order to scan the x–y slice plane located by the z-axis gradient control, as seen in Figure 16.43, the second step is to locate data points along the axis. A gradient is also applied along the x-axis for spatial localization,

$$\omega = \gamma B_x - \gamma B_0 = \gamma G_x x \tag{16.74a}$$

Again, with mixing, $\omega_{\text{mix}} = \omega_0$ for $B = B_0$, the difference frequency is proportional to the position x,

$$\omega_\Delta = \omega - \omega_{\text{mix}} = \gamma G_x x \tag{16.74b}$$

For z slice selection, a specific frequency is applied that corresponds to a particular desired location, according to Eq. (16.72b). Although a similar principle is used for the locations along x, the implementation is somewhat different. Here, the detected signal

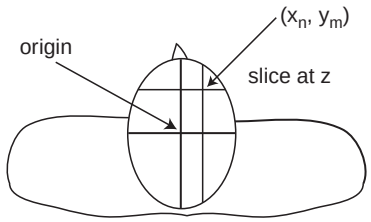


FIGURE 16.43 Depiction of an x–y slice plane of thickness Δz at a distance z, like the positions shown in Figure 16.41.

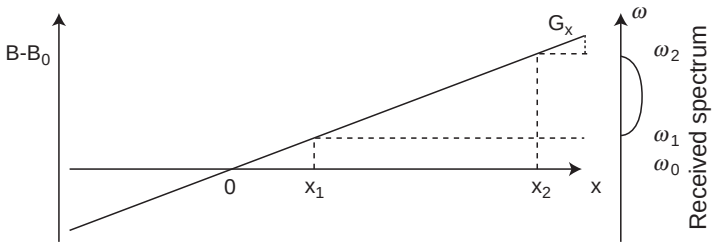


FIGURE 16.44 Frequency encoding method for x-axis with two locations, x_1 and x_2, shown.

frequency varies with distance x, a consequence of the linear magnetic field gradient, as shown in Figure 16.44. A number of excited resonances positioned along the scan line x can contribute to the overall detected signal and are encoded in the received signal as different frequencies.

Unlike diagnostic ultrasound imaging in which specific echo delays in time are associated with reflectors at different round-trip distances, MRI signals for different spatial locations are frequency encoded and added together in the same time signal position. In this case,

$$V(f) = \sum_n S(f - f_n) \tag{16.75}$$

The means of decoding these positions will be explained in a later section. Although MRI systems have employed yet another linear magnetic gradient along the y-axis to provide localization for this coordinate, a different phase-encoding scheme is now more common. The mechanism employed is that of the flip angle, Eq. (16.61), with a different twist

$$\varphi(y) = \omega t = -(\gamma B_0 t + \gamma G_y t_p y) \tag{16.76}$$

in which $y = y' + y_0$, where y_0 is the value at the center. In the previous discussion of the flip angle, an rf magnetic field with an appropriate pulse width, t_p, was applied to the spins to rotate a magnetization vector through a desired angle. In this context, Eq. (16.76) shows that either the pulse length, t_p, or the pulse gradient slope, G_y, can be used to change the phase.

The preferred method, as shown in Figure 16.45, is not to change pulse widths but rather to alter amplitudes and signs of the gradient slope, which can be changed by voltage amplitudes applied to the y-axis gradient coil. In order to fill out the x–y plane, a means of

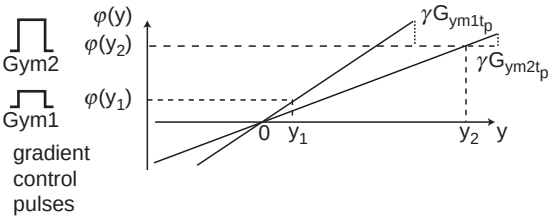


FIGURE 16.45 Phase encoding method for the relative y-axis for two locations, y_1 and y_2.

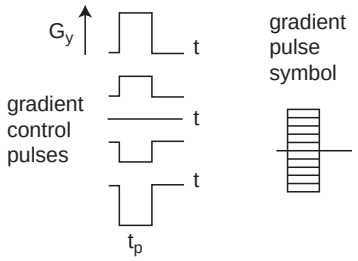


FIGURE 16.46 Pulse sequences for several consecutive lines. Note the amplitude of G_y pulses changes on sequential lines. TR is the repetition time between lines. Usually the amplitudes for G_y range from a maximum value, G_{ymax}, through a negative value of the same magnitude.

scanning along y is needed. Typically, 128 or 256 different excitations are utilized, and each one has to be uniquely linked with a specific y position. Pulse and excitation sequences for several consecutive lines are shown in Figure 16.46.

16.3.6 The Magnetic Resonance Imaging Sequence

Almost all the necessary bits and pieces needed for the imaging process have been introduced. Now, step by step, the different controls and detection methods will be reviewed and assembled into a sequence of events for imaging. The process will be simplified slightly for clarification. Figure 16.43 shows an x–y image plane that contains the isocenter, or center of magnetic field coordinate system, for a head scan. How net magnetization signals from specific locations in this plane translate into magnetic fields, detected waveforms, and ultimately into an image will be demonstrated.

Once the subject is within the MRI magnet, a series of sequence pulses are sent, as shown in Figure 16.47. In the figure, two rf pulses, one a $\pi/2$ pulse and one a π pulse, force the magnetization vectors to precess into the x–y plane to set up spin echoes, as described in Section 16.3.4. The $G_z z$-axis gradient pulses occur over the same intervals as the rf pulses. They turn on the gradient coils, which, in combination with the frequency selected for the rf pulses, select the z location of the x–y image plane, as detailed in Section 16.3.5.

The gradient pulses for the x-axis occur during the precession phase and also over the interval during which the spin echoes are formed. The field must be on during these

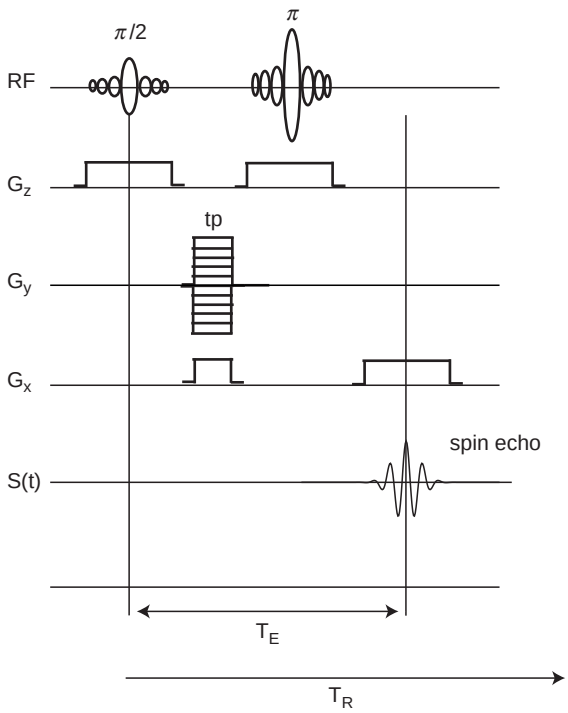


FIGURE 16.47 Essential rf excitation, and gradient pulse sequences, G_z, G_y, G_x, for spin-echo detection, $s(t)$.

intervals so the positions of the spins are translated into different frequencies that vary with position along the x-axis. Finally, a y gradient pulse excites the coils along the y-axis with an amplitude appropriate to phase encode the position y. Spin echoes are sensed at $t = T_E$. The sequence is repeated at intervals of T_R.

To form an image, the set of acquired signals and the object to be imaged (a distribution of net magnetization vectors) are assumed to be related by Fourier transforms. If a particular slice plane has been selected at z, then the object is a function of position (x, y), which implies that a double Fourier transform (for two dimensions) is involved. It will be easier to consider each dimension separately (x or y) and then combine them into a 2D transform. Signals detected from a net magnetization distribution at a position $(\underline{x}_n, y_m)$ will be considered.

Before starting with the y-axis, it will be helpful to review the role of the impulse in Fourier transform theory described in Section 16.2. The impulse and its transform play an essential role in the understanding of MRI. The impulse function is a generalized function that has the unusual property of sampling the integrand, as introduced in Section 16.1.1:

$$\int_{-\infty}^{\infty} \delta(f - f_0) R(f) df = R(f_0) \tag{16.77a}$$

When the inverse Fourier transform of the impulse function is taken, the result is an exponential

$$g(t) = \int_{-\infty}^{\infty} \delta(f - f_0) e^{i2\pi ft} df = e^{i2\pi f_0 t} \tag{16.77b}$$

which has the forward Fourier transform

$$G(f) = \int_{-\infty}^{\infty} \left(e^{i2\pi ft_0}\right) e^{-i2\pi tf} dt = \int_{-\infty}^{\infty} e^{i2\pi t(f-f_0)} dt = \delta(f - f_0) \tag{16.77c}$$

If $w(t)$ has the Fourier transform $W(f)$, then

$$\int_{-\infty}^{\infty} \left[w(t) e^{i2\pi ft_0}\right] e^{-i2\pi ft} dt = \int_{-\infty}^{\infty} w(t) e^{i2\pi t(f-f_0)} dt = W(f - f_0) \tag{16.77d}$$

The simplest place to start is the y-axis. For the y dimension, phase encoding according to Eq. (16.76) is applied and expressed with the constant phase term suppressed after mixing,

$$\varphi(y) = -\gamma G_y t_p y \tag{16.78a}$$

The phase associated with a particular location is

$$\varphi_m(y_m) = -(\gamma G_{y,m} t_p y_m) \tag{16.78b}$$

where the phase is indexed to a position y_m as well as a specific gradient slope coefficient, $Gy_{,m}$. The encoded time signal can be described by

$$s_m(0, G_y) = e^{-i\gamma t_p y_m G_y} \tag{16.78c}$$

The Fourier transform pair of variables for the y dimension will be the varying amplitude slope, G_y, as indicated in Figure 16.45 and a new variable called u. Analogous to the paired variables t and f, which have reciprocal units, u will have the upside-down or inverse units of m/T. These variables, G_y and u, can be used to construct a specific phase term that later will be associated with y_m in a Fourier transform format:

$$S_m(0, u) = \int_{-\infty}^{\infty} \left(e^{i2\pi u_m G_y}\right) e^{-i2\pi u G_y} dG_y = \int_{-\infty}^{\infty} e^{-i2\pi(u-u_m)G_y} dG_y = \delta(u - u_m) \tag{16.79}$$

Here, u appears in the argument of the second exponent in the integrand. The overall argument is recognized as a phase according to Eqs. (16.77c) and (16.78a) so an explicit expression for u can be obtained by equating phases:

$$2\pi u G_y = -\gamma t_p y G_y \tag{16.80a}$$

$$u = -\gamma t_p y/(2\pi) \tag{16.80b}$$

This evaluation of the Fourier transform of phase encoding follows directly from the discussion of the impulse function. As a consequence of this change in variables, the spectrum S_m in Eq. (16.79) can be restated in terms of the scaled variable from Eq. 16.80b,

$$S_m(0,u) = \delta(u - u\mathrm{m}) = \delta[(-\gamma t_p/(2\pi))(y - y_m)] \tag{16.81a}$$

$$S_m(0,y) = [2\pi/\gamma t_p)]\delta(y - y_m) \tag{16.81b}$$

where use has been made of the property $\delta(au) = \delta(u)/|a|$ from Section 16.2.

The derivation of this key equation reveals an important principle in magnetic resonance imaging. First, a local value of a gradient slope, $G_{y,m}$, encodes a position, y_m, into a phase according to Eq. (16.78b). Second, this phase, in terms of Fourier transforms, ensures that the signal function has a spectrum centered on the value of u associated with the location of the corresponding net magnetization. Through simple scaling, the scaled spectra are translated into spatial locations along the y-axis (Figure 16.48a). This process maps a scaled spectral magnitude into its locations along the y-axis.

To formulate the Fourier transform relation for the x-axis, a single spin echo waveform is generated, for example, at the position $y = 0$ in Figure 16.40b, so that $G_y = 0$. From Eq. (16.70), the signal waveform, delayed by time T_E and decaying with a unique time constant T_{2n} associated with location x_n can be expressed as

$$s_n(t,0) = 2S_0 e^{-|(t-T_E)/T_{2n}|} e^{j\omega_n(t-T_E)} \tag{16.82}$$

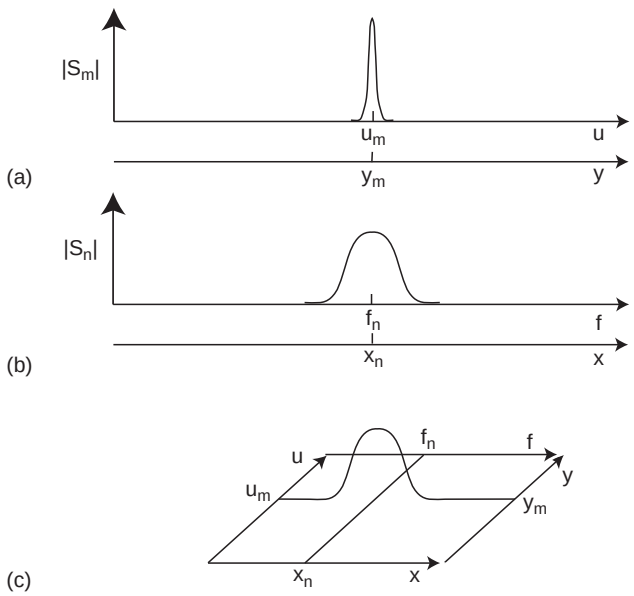


FIGURE 16.48 (a) Spectrally phase-encoded magnitudes plotted versus multiple scaled y- and u-axes. (b) Spectral frequency-encoded magnitudes plotted versus multiple scaled x- and f-axes. (c) Resulting pattern in xy or uf plane. Scale exaggerated for clarity.

in which the s is a function of t and $G_y = 0$, S_0 is a constant, and the Larmor frequency associated with location x_n is

$$\omega_n = \gamma G_x x_n \tag{16.83}$$

after mixing removes ω_0 at the isocenter.

The one-dimensional Fourier transform of Eq. (16.82) is

$$S_n(f,0) = \int_{-\infty}^{\infty} s(t,0)e^{-i2\pi ft}dt = \int_{-\infty}^{\infty} \left[S_0 e^{-|t-T_E|/T_{2n}}\right]\left[e^{-i2\pi f_n T_E} e^{-2\pi(f-f_n)t}\right]dt \tag{16.84a}$$

The form of the complex exponents is recognizable from Eq. (16.77d) as the transform of an impulse function centered on f_n. This relation implies that the Fourier transform of the function in the left brackets is centered on f_n. To show this result, it is necessary to determine if the Fourier transform of the first term in brackets is of the form of Eqs. (16.71a) and (16.71d) with S_0 as the constant,

$$W_n(f,\ 0) = \int_{-\infty}^{\infty} w_n(t,0)e^{-i2\pi ft}dt = \int_{-\infty}^{\infty} \left[S_0 e^{-|t-T_E|/T_{2n}}\right]e^{-i2\pi ft}dt = \frac{S_0}{1+(2\pi f T_{2n})^2} \tag{16.84b}$$

then $S_n\ (f,\ 0)$ can be expressed as

$$S_n(f,0) = e^{-i2\pi f_n T_E} W_n(f - f_n) \tag{16.84c}$$

From the relation for the Larmor frequency, this equation can be rewritten in terms of a scaled variable for frequency from Eq. (16.83) as

$$S_n(x,0) = e^{-i\gamma G_x x_n T_E} W_n\left[\frac{\gamma G_x}{2\pi}(x - x_n)\right] \tag{16.84d}$$

The derivation of this key equation reveals a second important principle in magnetic resonance imaging. First, a locally excited resonance encodes a position into a frequency f_n as part of the phase of a time waveform. Second, this phase, in terms of Fourier transforms, ensures that the signal function has a spectrum centered on the frequency associated with the location of the net magnetization density. Third, through simple scaling, the scaled spectra are translated into spatial locations along the x-axis (Figure 16.48b). This process maps scaled spectral magnitudes into their spatial locations along the x-axis.

To advance to two dimensions, the one-dimensional inverse Fourier transforms are combined in a two-dimensional relation:

$$s(t,G_y) = \Im\Im^{-1}[I(f,u)] = \iint I(f,u)e^{i2\pi ft}e^{i2\pi uG_y}dfdu \tag{16.85}$$

To find the object distribution, I, from the measured set of signals from a location (x_n, y_m), a forward 2D Fourier transform is

$$I(f,u) = S_{mn}(f,u) = \Im\Im[s_{mn}(t,G_y)] = \iint s_{mn}(t,G_y)e^{-i2\pi ft}e^{-i2\pi uG_y}dtdG_y \tag{16.86}$$

If the set of signals is taken to be simply the product of the one-dimensional signals for this simplified example

$$s_{mn}(t, G_y) = \left[S_0 e^{-|(t-T_E)/T_{2mn}|} e^{j\omega_n(t-T_E)}\right] e^{i2\pi y_m G_y} \tag{16.87}$$

then inserting this set into the 2D transform of Eq. (16.86) yields

$$S_{mn}(x, y) = \left\{ e^{-i\gamma G_x x_n T_E} W_{mn}\left[\frac{\gamma G_x}{2\pi}(x - x_n)\right] \right\} [2\pi/(\gamma t_p)]\delta(y - y_m) \tag{16.88}$$

In this case, as depicted in Figure 16.48c, the intersection of these one-dimensional functions places a scaled spectrum at the location (x_n, y_m). The role of u_m is to localize the position of the spectrum along y through scaling. In an MRI image, this spectrum, S_{mn}, is displayed as amplitude mapped on a gray scale with full white as maximum amplitude in the image plane at location z. What are displayed, then, are not the detected time signals but the amplitudes of their spectra, which through encoding methods are mapped into their proper spatial locations from anatomy.

In reality, during actual measurements, what are sensed are the signals from all the active net magnetization vectors from all the locations being scanned. The detected voltage is the sum of all the signals picked up in a specific z plane

$$V(t, G_y) = \sum_{m,n} s_{m,n}(t,\ G_y) \tag{16.89}$$

where the signals are not the ideal forms derived in Eqs. (16.87) and (16.88) but signals reflecting the characteristics and organization of tissue structures.

A convenient way of discussing the signals in terms of Fourier transforms is to describe the acquired signals as belonging to a k-space domain and the transformed signals as being in the space (x, y) domain. As shown in the following, this perspective is simply accomplished as a change in variables in the previous transform equations. The simplest place to start is in the exponential arguments of the transform equations, Eqs. (16.83) and (16.80a), that contain the key variables. These variables are set equal to the desired new ones:

$$2\pi f_n t = k_x x_n \tag{16.90a}$$

$$2\pi u G_y = k_y y_m \tag{16.90b}$$

From previous definitions, it is straightforward to show that

$$k_x = \gamma G_x t \tag{16.91a}$$

$$k_y = -\gamma G_y t_p \tag{16.91b}$$

or

$$t = k_x/\gamma G_x \tag{16.92a}$$

$$G_y = -k_y/\gamma t_p \tag{16.92b}$$

Equations (16.92a) and (16.92b) can be substituted in Eq. (16.85) to yield

$$s(k_x/\gamma G_x,\ -k_y/\gamma t_p) = \iint \left\{ \left(\frac{\gamma G_x}{2\pi}\right)\left(\frac{-\gamma t_p}{2\pi}\right) I\left[\left(\frac{\gamma G_x}{2\pi}\right)x,\ \left(\frac{-\gamma t_p}{2\pi}\right)y\right] \right\} e^{ik_x x} e^{ik_y y} dx dy \tag{16.93}$$

or by redefining the functions in terms of the new variables, $\tilde{s}$ and $\tilde{I}$,

$$\tilde{s}(k_x, k_y) = \iint \tilde{I}(x,\ y) e^{ik_x x} e^{ik_y y} dx dy \tag{16.94}$$

where $\tilde{I}\ (x, y)$ is equal to the term in the braces outside the exponentials in Eq. (16.93) and $\tilde{s}$ is redefined as well. Similarly, Eq. (16.86) becomes

$$\tilde{I}(x,\ y) = \iint (\tilde{s}(k_x,\ k_y) e^{-ik_x x} e^{-ik_x x} e^{-ik_y y} dk_x dk_y \tag{16.95}$$

Previous results have now been recast in *xy*-space and *k*-space variables.

16.3.7 Magnetic Resonance Imaging Systems

The main parts of an MRI system are shown in Figure 16.49. A large superconducting magnet provides a homogeneous static B_0 magnetic field within its interior. Typical clinical field values vary from 0.15 to 2 T, but there are systems available in the 3 to 5 T range. Inserted within the main magnet are pairs of surface coils for producing the *x*-, *y*-, and *z*-axis gradients. These typically produce variations of only a few percent of the value of the static field for spatially localizing the magnetic spin signals of the body. These signals are picked up by rf transmit/receiver coils that are sometimes in the form of surface coils placed close to the body. The rf signals are directed by a computer (CPU) through the rf pulse generator and sequencer. A switch routes the received rf signals through an amplifier and into a data acquisition unit, after which they are processed and Fourier transformed to create an image. Modern systems allow the operator considerable flexibility to alter the gradients and pulse sequences to achieve different image effects and weighting. Typical image acquisition times are on the order of 50 to 100 ms, though faster acquisitions are becoming available.

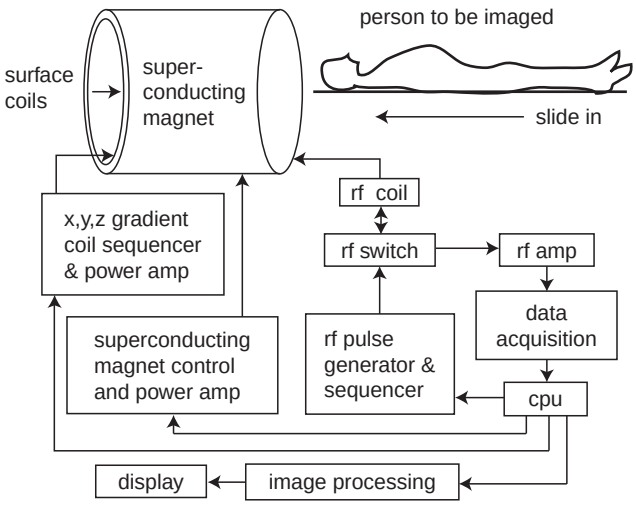


FIGURE 16.49 Block diagram of an MRI system.

Different signal characteristics can be emphasized by a manipulation of the basic pulse timing sequences (see Figure 16.47). For example, a short pulse repetition interval, *TR*, and a short spin echo time, *TE*, results in the commonly used T_1 weighting (*T*1W) shown in Figure 16.50b. For example, typical values are $TE = 12$ ms and $TR = 300$ ms. Prolonging *TR* and the delays before and after *TE* produces a T_2-weighted (*T*2W) image as in Figure 16.50a. An example of this is $TE = 100$ ms and $TR = 3{,}000$ ms. A combination of a long *TR* and a short *TE* creates what is called a proton density weighted (PD) image, shown in Figure 16.50c. For example, typical values are $TE = 30$ ms and $TR = 3{,}000$ ms. Recall that T_1 is more sensitive to local thermal properties and T_2 depends more on local transverse magnetic field inhomogeneities. Many other alternative timing and pulse sequences are possible and are useful for different clinical applications.

An important aspect of imaging is contrast among different types of tissues. In Figure 16.51 longitudinal magnetization is shown for two tissues as a function of time, indicating their different T_1 rise times. Note that the contrast among the tissues varies at different times, as indicated by the vertical separation between the curves at different times. For example, compare the separations for times of 400 ms, 800 ms, and 1,600 ms. Here it can be seen that contrast can be changed by timing intervals. Another way of enhancing contrast is to inject a paramagnetic medium or contrast agent such as gadolinium.

16.3.8 Magnetic Resonance Imaging Applications

Functional MRI (fMRI) is the use of MRI to detect localized changes in brain activity, usually in the form of changes in cerebral metabolism, blood flow, volume, or oxygenation in response to task activation. These changes are interrelated and may have opposite effects. For example, an increase in blood flow increases blood oxygenation, whereas an increase in metabolism decreases it. The most common means of detection is measuring the changes in the magnetic susceptibility of hemoglobin. Oxygenated blood is dimagnetic and deoxygenated blood is paramagnetic. These differences lead to a detection method called blood oxygen level dependent contrast (BOLD). Changes in blood oxygenation can be seen as

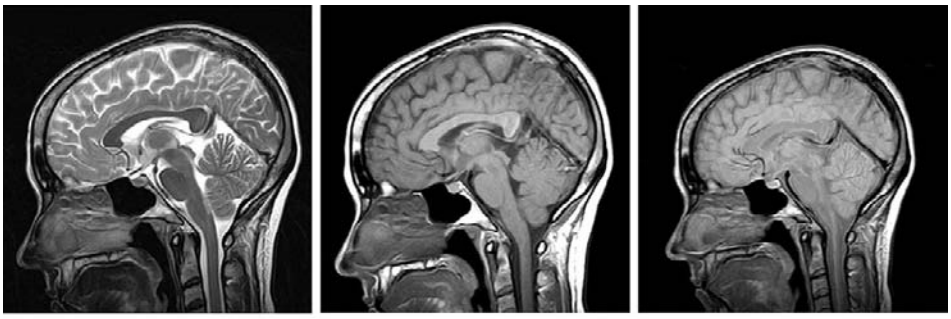


FIGURE 16.50 MR images of sagittal brain cross sections using three types of weighting planes: (a) T_2, (b) T_1, and (c) proton density. *Courtesy of Philips Medical Systems.*

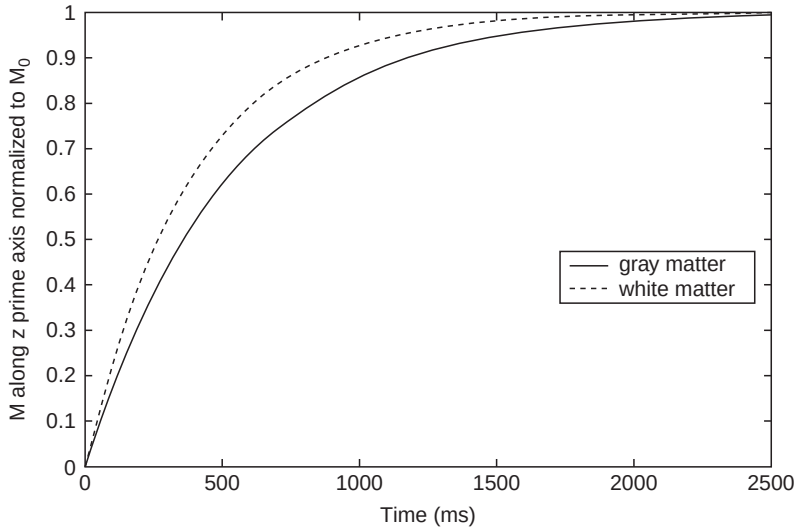


FIGURE 16.51 Longitudinal magnetization versus time for gray matter and white matter.

differences in the T^*_2 decay constant, so T_2 weighted images are used. Because these changes are very small, images before and after task initiation are subtracted from each other and the resultant difference image is overlaid on a standard image. Special care must be taken when obtaining these images because the effect is small and can be corrupted by several sources of noise. Typically, hundreds of images are taken for each slice plane position and statistical analysis is used to produce the final image. Sources of noise errors are thermal noise, head movement, and respiratory and cardiac cycles. An example of an fMRI image is shown in Figure 16.52 [5].

Functional MRI has been used extensively to map regions of the brain connected with stimuli, activities, and higher-level cognition. Another example of this type of neuroimaging—biological motion perception—was studied in humans [7]. Subjects were placed in an MRI scanner so they could view moving images while they were being imaged in real-time. Experiments involved biological motion (BM) perception, face recognition, and nonrigid motion (NRM) perception. In Figure 16.53, results from experiments involving these variables are displayed as color-coded regions in three-dimensional reconstructions of a brain. Note the degree of spatial resolution, localization, and discrimination achieved for distinct activities.

16.4 MAGNETOENCEPHALOGRAPHY

Magnetoencephalography (MEG) is a form of neuroimaging that maps the tangential components of magnetic fields associated with scalp potentials produced by the brain [3]. These potentials are the same ones that can be recorded as electroencephalograms (EEGs), but the dynamic magnetic components of these potentials contain different information

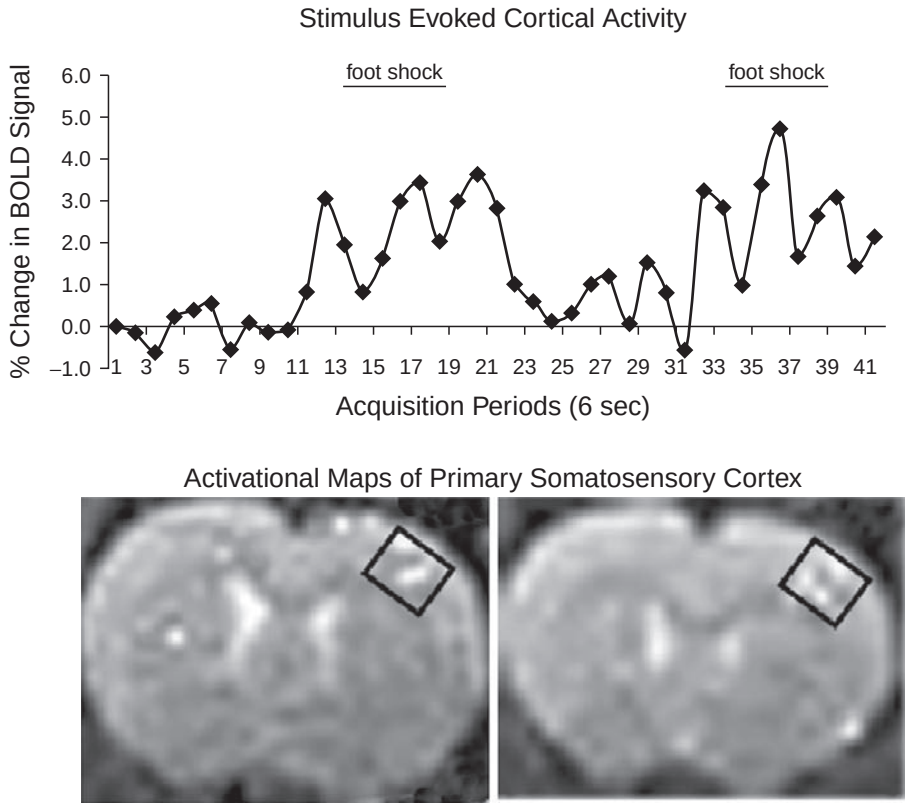


FIGURE 16.52 Functional imaging with fast spin echo. Images show percent change in BOLD signal in the right somatosensory cortex (boxed area in activational maps) in response to electrical stimuli to the left hind limb paw of a rat. *From [5], with permission from Elsevier.*

with spatial sampling. Unlike fMRI or PET images that provide indirect or delayed clues to synaptic events through changes in the metabolism or blood flow, MEG locates regions of the brain that respond directly to stimuli and yields corresponding time responses at various spatial locations.

To accomplish these tasks, a subject is placed into a helmet-like enclosure that houses a semispherical array of hundreds of magnetic sensing coils. Because the brain's dynamic magnetic field is weaker by five orders of magnitude than magnetic background noise, special noise-suppressing coil configurations, as well as ultrasensitive detectors, superconducting quantum interference devices (SQUIDs), and shielded chambers are needed. The signals from the coils as well as their locations on the semispherical grid are employed in solving the inverse problem of determining the source location of the signals in the brain. These calculations are repeated for different combinations of signals until active regions of the brain are identified with their associated time waveforms. These waveforms are combined with MRI and/or fMRI imaging to aid in localization. A resulting MEG neuroimage is shown in Figure 16.54. In this experiment, parallel localization ("where") and recognition ("what")

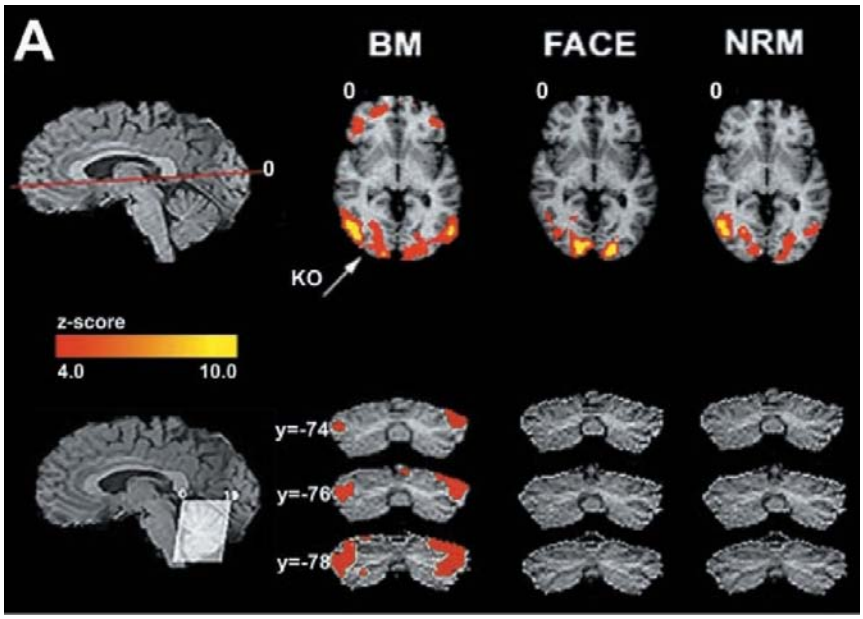


FIGURE 16.53 Example of functional MRI imaging of human brain responses. (A) The grey nearly horizontal line through the top sagittal slice of the brain at the left (denoted by a "0") shows the position of the three axial slices to the right. On the upper right are group z maps showing statistically significant, group averaged neural activity for each experiment: biological movement (BM), face-gender, and nonrigid motion (NRM). To generate these z maps for each experiment, a t test was performed to contrast each experiment's task with the control. The grey scale represents the z score of the activation. Note most of the activity is in the KO region and the lateral cerebellum. On the bottom left, the lines through the cerebellum in the sagittal slice indicate the angulation of the coronal cut. To the right, the three coronal slices (arranged vertically) show that only BM had activity in the lateral cerebellum. *From [4].*

pathways are verified, as well as the ability of a subject to selectively focus attention on either a "where" or "what" aspect. Dipole locations as calculated from received MEG signals on the semispherical grid are indicated on corresponding MRI images.

MEG is used to explore the characteristics of mental patterns such as epilepsy and schizophrenia. It is also useful in brain research to understand cognitive functioning and to learn which parts of the brain are involved in different tasks.

16.5 CONTRAST AGENTS

Contrast agents emphasize or magnify physiological features or functions that would otherwise be invisible, weak, or obscured in images. One of the most frequent applications of these agents is in the vasculature. Another is the uptake of an agent by an organ, indicating a degree of functional or metabolic activity. Each agent is designed to work with the particular physics of the intended imaging modality as well as the region of indication in the body.

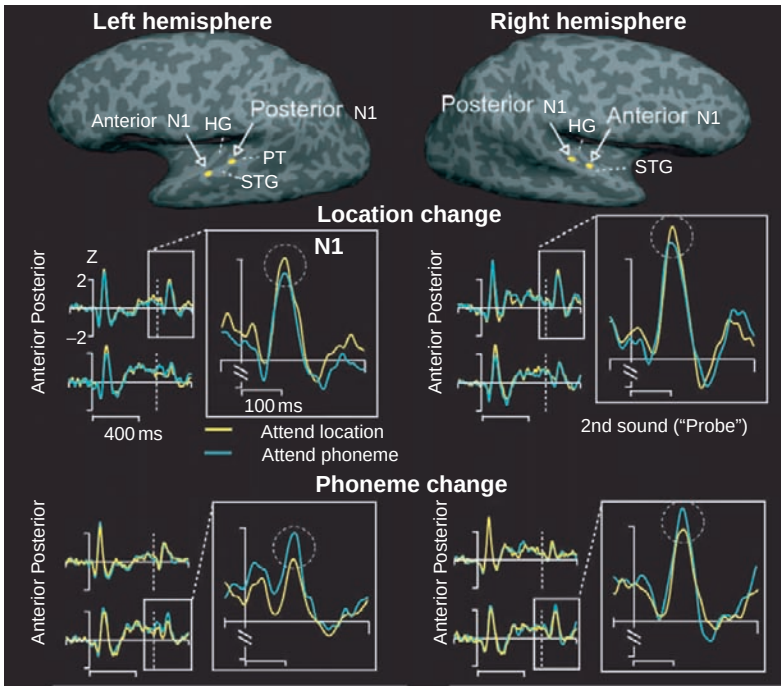


FIGURE 16.54 (Top) MEG locations of equivalent-current dipoles (ECD) sources to two different words, "Adaptor" and "Probe," sequentially spoken from different external spatial locations, displayed on the inflated brain hemispheres, MRI images, of one subject. (Middle and bottom) Amplitude-normalized source waveforms in the middle indicate that the "where" pathways are associated with the N1 locations, whereas waveforms at the bottom for parallel "what" pathways are connected with the N1 locations. Differences (peak changes encircled) in the averaged waveforms indicate the effects of selective attention to either sound or location. The N1 response amplitudes to Probes (encircled) are modulated task-dependently. The posterior N1 activity to Probes following spatially different Adaptors is enhanced by spatial attention. The anterior N1 activity to Probes following phonetically different Adaptors is enhanced by phonetic attention. *From [1].*

16.5.1 Computed Tomography Agents

Intravenously administered CT contrast agents containing iodine or gadolinium are employed in the visualization of blood vessels. For example, for a CT pulmonary angiogram, emboli and obstructions can be seen through attenuation of the x-rays or opacification, which brightens the appearance of flow in a vessel. Through the enhancement of blood vessels, other applications include visualization of the internal structures of the brain, spine, liver, and kidney and other organs. Afterward, the kidney and liver normally aid in passing the agents out of the system. In some cases, these agents may involve some risk and possible allergic reactions such as those of the kidney.

For the diagnosis of abdominal diseases and injuries, CT contrast agents are administered either orally or rectally. A dilute suspension of barium sulfate is used most often. Applications include the diagnosis and characterization of cancer and the investigation of sources of severe abdominal pain and trauma.

16.5.2 Magnetic Resonance Imaging Agents

For MRI, agents distort or alter the magnetic fields locally, where they are concentrated so there is a relative change in a region. The effects achieved can either intensify or weaken the signal close by, depending on the image weighting selected and the context of the agent; they work by reducing T_1 and T_2 relaxation times nearby. The agents tend to be either paramagnetic or supermagnetic materials that are modified to be ingested or infused into the body safely and passed out. The major type of agent used is chelated gadolinium in various forms. The main applications are intravascular and gastrointestinal.

16.5.3 Positron Emission Tomography Agents

For positron emission tomography (PET, introduced in Section 15.3.3), gamma rays are detected that come from a positron-emitting radionuclide material. The location and concentration of these tracer materials in the body are reconstructed from multiple detectors into a 3D image. The tracer most often used is 18F-FDG, a glucose analog. It is concentrations of this substance that are taken up by tissues and indicate local metabolic activity. Only concentrations of this contrast agent are displayed, hanging in space, so it is difficult to determine spatial landmarks to determine locations; therefore, PET images are often combined with another imaging modality such as CT (described in Section 16.7).

16.5.4 Ultrasound Agents

Ultrasound contrast agents are small microspheres that are used to highlight the passage of blood in regions otherwise difficult to image, such as blood, which is nearly invisible to ultrasound without special processing. These agents are usually microspheres filled with air or gases such as perfluorocarbon, with thin, flexible shells typically made of human serum albumin or surfactants, and are about 4 μm in diameter, similar to the size of red blood cells. They are injected into the venous system to act as blood cell tracers. While in the blood, ultrasound agents significantly improve the ability and sensitivity of ultrasound to follow the flow of blood, especially in small vessels, and the perfusion of blood in muscle, especially those of the heart, and to identify walls of the moving chambers of the heart (opacification). After traveling along in blood, the spheres diffuse and are released as they pass through the lungs.

Ultrasound contrast agents are designed for different applications and effects and are matched to operate at certain insonifying frequencies. They act as miniature nonlinear resonators that greatly enhance their reflectivity for ultrasound. Under certain conditions, the unusual properties of these agents are changes in size, cavitation, fragmentation, or directed movement.

16.6 COMPARISON OF IMAGING MODES

How can imaging modalities be compared? Each major diagnostic imaging method is examined in the following sections, and the overall results are tallied in Table 16.4. Examples of three imaging modalities—CT, MRI, and ultrasound—are shown as different images of the right kidney in Figure 16.55.

TABLE 16.4 Comparison of Imaging Modalities

	Ultrasound	X-ray	CT	MRI
What is imaged	Mechanical properties	Mean tissue absorption	Tissue absorption	Biochemistry (T_1 & T_2, & PD)
Access	Small windows adequate	2 sides needed	Circumferential around body	Circumferential around body
Spatial resolution	Frequency & axially dependent; 3 to 0.3 mm	~1 mm	~1 mm	~1 mm
Penetration	Frequency dependent; 3 to 25 cm	Excellent	Excellent	Excellent
Safety	Very good	Ionizing radiation	Ionizing radiation	Very good
Speed	100 frames per second	Minutes	Half-minute to minutes	Minutes
Cost	$	$	$$$$	$$$$$$$$
Portability	Excellent	Good	Poor	Poor

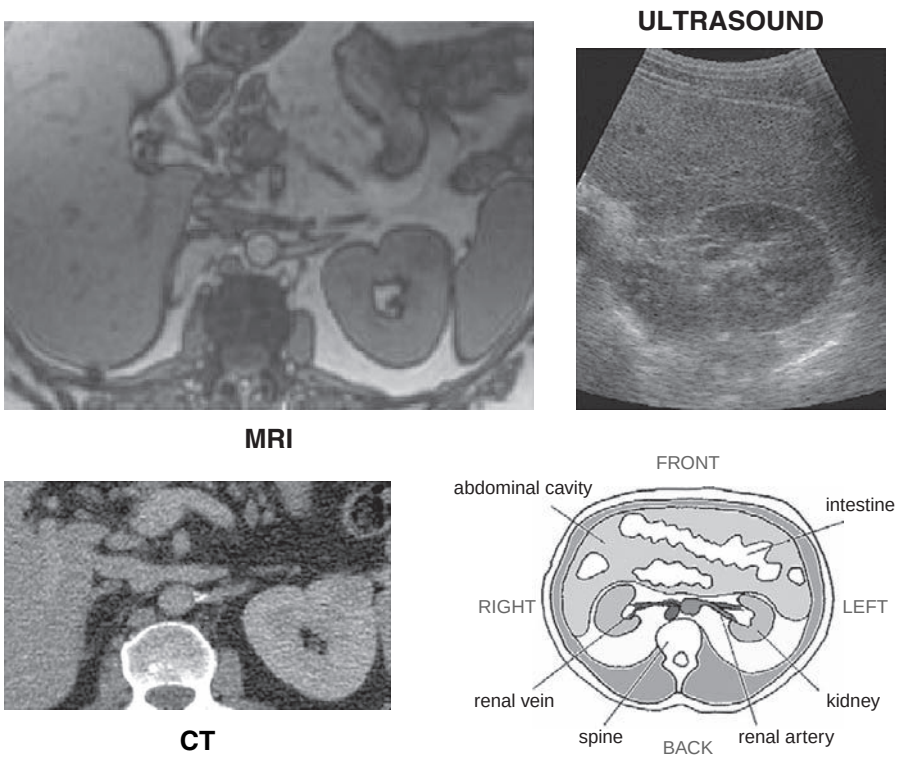


FIGURE 16.55 Images of the right kidney as viewed by different imaging modalities. Upper left: MRI; bottom left: CT; top right: ultrasound; bottom right: graphical depiction of abdominal cross section showing location of kidneys. *Courtesy of Dr. Marilyn Roubidoux.*

16.6.1 Ultrasound

Unlike other imaging modalities, ultrasound resolution and penetration depend on the center frequency and type of transducer selected. The resolution is spatially variant and depends on both the size of the active aperture and the center frequency (and bandwidth) of the transducer and the selected transmit focal depth. A commonly used focal depth to aperture ratio is five, so the half-power beam width is approximately two wavelengths at the center frequency, and thus the transmitted lateral spatial resolution is about two wavelengths. For typical frequencies in use ranging from 1 to 15 MHz, lateral resolution ranges from 3 mm to 0.3 mm and is the smallest in the focal region and varies elsewhere in a nonuniform way because of diffraction effects caused by apertures on the order of a few to tens of wavelengths. For a short pulse, axial resolution is approximately two wavelengths.

Another major factor in determining resolution is attenuation that limits penetration. Attenuation increases with higher center frequencies and depth; therefore, penetration decreases correspondingly, so fine resolution is difficult to achieve at deeper depths.

Ultrasound images are highly detailed and geometrically correct to first-order maps of the mechanical structures of the body according to their "acoustic properties" such as differences in characteristic impedance that depend on stiffness or elasticity and density. The dynamic motion of organs such as the heart can be revealed by ultrasound operating at up to hundreds of frames per second.

Diagnostic ultrasound is noninvasive. Ultrasound is also safe and does not have any cumulative biological side effects. Two other strengths of ultrasound imaging are its relatively low cost and portability. With the widespread availability of miniature portable ultrasound systems for screening and imaging, these two factors will continue to improve.

A high skill level is needed to obtain good images with ultrasound. This expertise is necessary because of the number of access windows, differences in anatomy, the many possible planes of view, and the experience required to find relevant planes and targets of diagnostic significance and to optimize instrumentation. Furthermore, a great deal of experience is required to recognize, interpret, and measure images for diagnosis.

16.6.2 Computed Tomography Imaging

Computed tomography (CT) (also known as computed axial tomography, or CAT) scanning involves x-rays and has been described in Chapter 15. As the x-rays pass through the body, they are absorbed by tissue so an overall "mean attenuation" image results along the ray path. Spatial resolution is not determined by wavelength but by focal spot size of the x-ray tube and scatter from tissue; a typical resolution is about 1 mm. Radioactive contrast agents can be ingested or injected to improve visualization of vessels. Though exposures are short, x-rays are a form of ionizing radiation, so dosage effects can be cumulative and extra precautions are needed for sensitive organs such as the eyes and for pregnancies.

CT equipment is large and stationary so a person can fit inside, and as a result, it is relatively expensive to operate. Consecutive pictures of a moving heart are now achievable through synchronization to ECG signals. The resolution of CT images is typically 1 mm. CT scanning creates superb images of the brain, bone, lungs, and soft tissue, making it complementary to ultrasound.

Although the taking of CT images requires training, it is not difficult. Interpretation of CT cross-sectional images demands considerable experience for a definitive diagnosis.

16.6.3 Magnetic Resonance Imaging

For magnetic resonance imaging, the patient is placed in a strong static magnetic field created by a large enclosing electromagnet. The resolution is mainly determined by the gradient or shape of the magnetic field, and it is typically 1 mm. Images are calculated by reconstruction algorithms based on the sensed voltages proportional to the relaxation times. Tomographic images of cross-sectional slices of the body are computed. The imaging process is fast and reasonably safe, since no ionizing radiation is used. Care must be taken to keep ferromagnetic materials away from the powerful magnets used, and there are limits to the strength of the applied magnetic fields and how quickly they are switched. Because the equipment needed to make the images is expensive, exams are costly.

MRI equipment has several degrees of freedom such as the timing, orientation, and frequency of magnetic fields; therefore, a high level of skill is necessary to acquire diagnostically useful images. Diagnostic interpretation of images involves both a thorough knowledge of the settings of the system and experience.

16.7 IMAGE FUSION

As implied by the comparison of images and the discussion of PET imaging in Section 16.5.2, the physics of each imaging modality reveals different characteristics of tissue properties and functions. In order to obtain a more complete picture, two (or more) imaging modalities can be combined. Image fusion is the simultaneous display of two different types of images either side by side or superimposed. To demonstrate these different points of view, Figure 16.56 compares four types of imaging views of a metastases and is explained next.

One of the leading types of image fusion is PET/CT, where specialized scanners, spatially coregistered, take two sets of images of the same subject in one convenient instrument [2]. As an example, images of a patient with cancer given FDG are shown in Figure 16.56. The top left PET images in frontal and transverse views (A and B) display the uptake of FDG as a small, dark region (see arrow), signifying hypermetabolic activity and the likelihood of a metastases; however, pinpointing the anatomical location of the cancer nodule is ambiguous in these views. The CT view (D), displays good resolution as well as tissue structures but not a clear indication of the nodule. The fusion image (E), a combination of

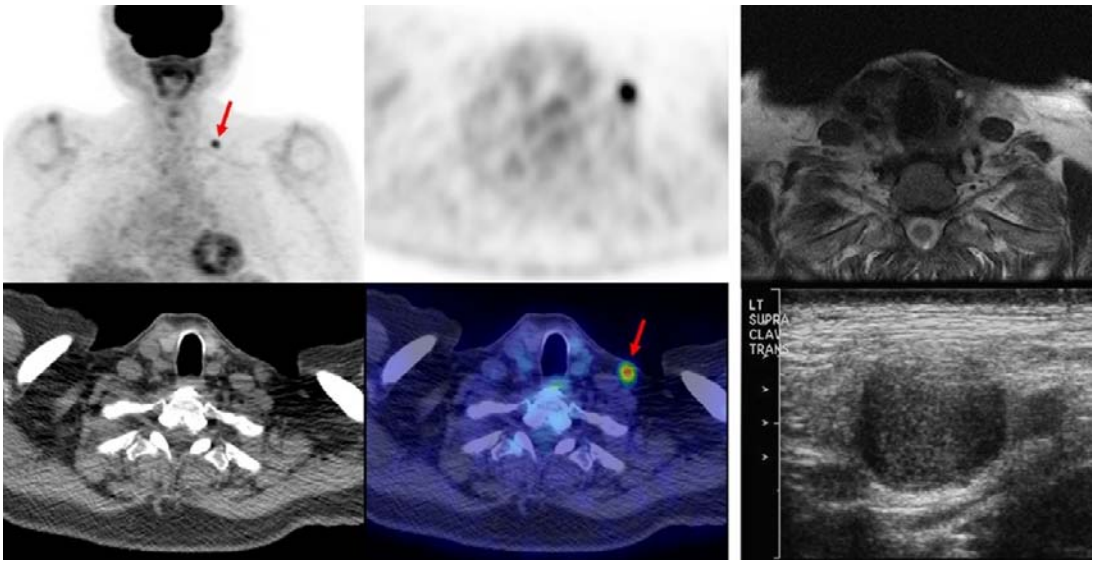


FIGURE 16.56 Images of different imaging modalities. (A) Upper left: PET frontal; (B) upper middle: PET transverse; (C) upper right: MRI transverse; (D) bottom left: CT transverse; (E) bottom middle: PET/CT transverse fused image; and (F) bottom right: ultrasound. Images of a 74-year-old male patient with nasal cavity esthesioneuroblastoma (a form of nasal cancer) who was restaged on a routine follow-up. PET identified a left supracalvicular hypermetabolic nodal metastases (A, B, D, and E) that was not identified in the MRI (C). An ultrasound-guided biopsy of the node was positive for esthesioneuroblastoma and then surgically resected (F). *Courtesy of Dr. Rathan Subramaniam, 2010.*

the PET and CT transverse, clearly emphasizes the metastases. For comparison, the MRI image (C) of the same view, while providing spatial organization of the organs, is by itself not as informative. Finally, the ultrasound image (F) provides a different view and was useful for a guided biopsy.

16.8 SUMMARY

The different features and characteristics of three major imaging modalities are summarized in Table 16.4 and shown in Figures 16.55 and 16.56. The tissue properties being imaged are considerably different among the different modalities and can be complementary. Image fusion offers new opportunities to reconcile limitations of individual imaging modalities. Different fusion combinations of imaging modalities have been attempted, but PET/CT fusion is rapidly becoming the most popular [2]. Both CT and MRI can image the whole body with consistent resolution and contrast; however, they are expensive methods of imaging and are slow and not portable. Ultrasound, on the other hand, has high but variable resolution and penetration and limited access to certain portions of the body (intestines, lungs, and bones). It is used to image soft tissues only, but it has the advantages of low cost, portability, and real-time interactive imaging.

16.9 EXERCISES

1. Find the Fourier transform of $\prod[(t - t_0)c/L]$ given the Fourier transform pair $\Im[\prod(t)] = \sin(\pi f)/(\pi f) = \text{sinc}(f)$.
2. Select the correct answer. Ultrasound imaging for medical applications began to grow significantly when
   (a) Wild and Reid demonstrated real-time imaging
   (b) Edler and Hertz started the science of echocardiography
   (c) Stewart found a connection between cancer in children and prenatal x-ray exposure
   (d) Satomura and his colleagues detected blood flow using Doppler techniques
3. Compute the pressure reflection factor between blood and muscle at normal incidence in dB (use Table 16.1). Compare the reflection factor at 45° incidence to this result. Which is greater?
4. Derive the normal reflection and transmission factors in terms of intensity rather than amplitude. Determine the normal transmission intensity factor between blood and (a) the liver, (b) an ideal rigid boundary, and (c) air. How do these compare to the corresponding amplitude results?
5. If the Q of the electrical part of a transducer equivalent circuit is defined as $Q = f_0/\Delta f = 1/[\omega_0 C_0 R_A(f_0)]$, where Δf is the –6 dB bandwidth and f_0 is 5 MHz, compare the fractional –6 dB bandwidths (in %) for the following two transducer materials: A $k_T = 0.7$, $\varepsilon_R = 1470$ and $k_T = 0.9$ and $\varepsilon_R = 680$. (Hint: See Eq. (16.27c).)
6. Calculate the electrical impedances of the following two transducer designs at resonance using the constants from Example Problem 16.3: one with a direct water load and one with a matching layer between the crystal and the water load, both with $Z_B = Z_C$. (Hint: Use Eq. (16.34).) What is the improvement in acoustic loss by including the matching layer?
7. Ultrasound is not used to image bone directly. (a) To estimate the strength of the reflection, determine the normal amplitude reflection factor from muscle to bone. (b) Determine a matching layer impedance to match the muscle better to bone. (c) What are the amplitude reflection and transmission factors with the matching layer at resonance?
8. A company can only afford either a tuning inductor or a matching layer for their transducer product. Which would you recommend and why? Material details: area: A = 400 mm^2; thickness: d = 0.87 mm; crystal data: c = 4.35 km/s; $\varepsilon_R^S/\varepsilon_0 = 830$; $\varepsilon_0 = 8.85$ pF; $k_T = 0.49$. Electrical source impedance: $R_g = 50$ ohms.
9. Compare the frame rates for the following two cases: a 3 MHz cardiac sector scanner with 128 lines per frame for a depth of 10 cm, and a small parts 10 MHz linear array with 300 lines per frame and a 1 cm depth.
10. If a 5 MHz transducer in an imaging system has an overall round-trip dynamic range of 100 dB, what is the greatest distance it can detect a pressure reflection between a liver-bone boundary? Between a liver-muscle boundary? Assume average tissue properties to the boundary are the same as that of liver. (Use Table 16.1 for relevant tissue and absorption data.)
11. A small lesion about 2 mm in diameter at a depth of 6 cm must be found. Which of the following rectangular array transducers with apertures and center frequencies would you use and why?
    (a) L = 15 mm and 6 MHz
    (b) L = 11 mm and 5 MHz
    (c) L = 12 mm and 3 MHz

    Assume c = 1.5 mm/μs and a focal length of F = 60 mm in the scan plane.

12. For a spherically focusing transducer with an aperture a = 5 mm and a frequency of 5 MHz, show which of the following combinations places the effective focal length for a target at a depth of 40 mm.
(a) F = 50 mm
(b) F = 40 mm
(c) F = 80 mm

13. For the Doppler detection of blood, a certain 2 MHz ultrasound system cannot sense signals below a threshold of –35 dB. Assume a layer of muscle with negligible absorption loss above a blood vessel where blood is flowing at a velocity of 1 m/s. Find the range of incident angles (to the nearest degree) that will provide adequate signal strength and the corresponding Doppler frequencies. Use Table 16.1 and note the Doppler angle = 90°—incident angle.

14. For the arrangement of the whirling magnet in Figure 16.29, as d is increased, does the intercepted voltage increase or decrease? What is a physical explanation for this effect? (Note: Volt = Weber/s.)

15. Determine the nuclear magnetic dipole moments, μ_z, for ^{1}H, ^{19}F, and ^{23}Na. Use Table 16.2.

16. Calculate the net magnetization moment M_0 for water at a temperature of T = 300°K and a magnetic field of $B_0 = 0.5$ T and a proton density of N = 6.7×10^{19} protons/mm^3. Use a form of Eq. (16.54), $\hat{M}_0 = \sum_n \hat{\mu}_n = \hat{\mu}(n_+ - n_-)$ and assume $n_+ = N/2$.

17. If $\hat{B}_y(t) = \hat{y}B_1 \sin(\omega t)$, what will the field be in the rotated frame? After low-pass filtering to eliminate 2ω frequencies, what components are left?

18. (a) For a flip angle of $\pi/2$ radians and $B_1 = 10^{-5}$ T, what is t_p for ^{1}H? (b) Find the total flip time for a $\pi/2$ angle followed by a π angle rotation.

19. Determine the relative decay for two individual rotations of $\pi/2$ and π in white matter and a malignant tumor at the time $t = 250$ ms. Which angle provides more contrast between the two tissue types? Repeat at $t = 500$ ms, and compare the contrast to the previous case.

20. (a) Derive an expression for the spectrum of an F. I. D. pulse, $v'(t) = v(t)\,H(t)$, where $H(t)$ is the step function, $H(t) = 1$ for $t > 0$, and $H(t) = 0$ for $t < 0$. Utilize the Fourier transform pair, $\Im\left[e^{-a|t|}H(t)\right] = \dfrac{1 - i2\pi f/a}{|a|\left[1 + (2\pi f/a)^2\right]}$. (b) What is the complex ratio of the F.I.D. spectrum to that of the spin echo? (c) Find the value of this ratio at the Larmor frequency. (d) Can you find an explanation for this ratio based on the time waveforms involved?

21. (a) If $T_2 = T_2^* = 5$ ms for the F. I. D. pulse of problem 20, find the spectral component at the Larmor frequency of 5 MHz for the conditions $B_1 = 20\ \mu T$, I = 1 A, $V_{oxel} = 1$ cm^3, and $M_{xy}^0 = 5 \times 10^{-12}$ J/(T-mm^3). (b) How does this value compare for a similar spectral magnitude of a spin echo with T_2 for fat at the same Larmor frequency?

22. Find the –6 dB width of the envelope of a time pulse and the two end frequencies needed to scan a 1-mm-thick slice from –15 cm to +15 cm around an isocenter on the z-axis. Assume $G_z = 0.5\ T/m$.

23. (a) In order to cover a range of ±10 cm about an isocenter along the y-axis, phase encoding is applied with a pulse length $t_p = 10\ \mu s$ for ^{1}H and an overall phase shift of $\pi/2$. Find the slopes, G_{ym}, at the end points and the slope resolution for 256 steps.

Continued

24. Determine the relative position (x_n, y_m) from the center for ^{1}H, given $f_n = 78.5$ MHz, $G_x = 10^{-4}\ T/m$, $t_p = 5\,\mu s$, and $u_m = 21.3\ m/T$.
25. Derive $s_{nm}(k_x, k_y)$ based on Eq. (16.87).
26. Answer the following:
(a) What is displayed in an MRI image?
(b) How is a location from a set of signals in the body determined?
(c) Of the three imaging axes, which are actively controlled and which are passively sensed?
27. Explain the basic physical principles underlying the differences in appearance (image content, and tissue differentiation) among T_1, T_2, and PD weighted images.
28. What are the complementary aspects of PET and CT images that make them suitable for image fusion?

References

[1] J. Ahveninen, et al., Task-modulated "what" and "where" pathways in the human auditory cortex, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 14608–14613.
[2] T.M. Blodgett, C.C. Meltzer, D.W. Townsend, PET/CT: Form and Function, Radiology 242 (2007) 360–385.
[3] D. Cohen, E. Halgren, Magnetoencephalography, Encyclopedia of Neuroscience 5 (2009) 615–622.
[4] P.C. Lauterbur, Image formation by induced local interactions: Examples employing nuclear magnetic resonance, Nature 242 (1973) 190–191.
[5] R. Ludwig, G. Bodgdanov, J. King, A. Allard, C.F. Ferris, A dual Rf resonator system forhigh-field functional magnetic resonance imaging of small animals, J. Neurosci. Methods 132 (2004) 125–135.
[6] R.K. Panda, Development of novel piezoelectric composites by solid freeform fabrication techniques, Dissertation, Rutgers, New Brunswick, NJ, 1998.
[7] L.M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, Belliveau, Functional neuroanatomy of biological motion perception in humans, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 11658–11661.

Suggested Readings

F. Bitter, H.A. Medicus, Fields and Particles: An Introduction to Electromagnetic Wave Phenomena and Quantum Physics, American Elsevier, New York, 1973.
D.W. Chakeres, P. Schmalbrock, Fundamentals of Magnetic Resonance Imaging, Williams & Wilkins, Baltimore, MD, 1992.
Z.H. Cho, J.P. Jones, M. Singh, Foundations of Medical Imaging, Wiley, New York, 1993.
J.A. Jensen, Estimation of Blood Velocities Using Ultrasound, Cambridge Univ. Press, Cambridge, 1996.
E.E. Kim, E.F. Jackson, Molecular Imaging in Oncology, Springer-Verlag, Berlin, 1999.
G.S. Kino, Acoustic Waves: Devices, Imaging, and Analog Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1987.
V. Kuperman, Magnetic Resonance Imaging Physical Principles and Applications, Elsevier Academic, Boston, 2000.
Z.P. Liang, P.C. Lauterbur, Principles of Magnetic Resonance Imaging: A Signal Processing Approach, IEEE Press, New York, 2000.
K.K. Shung, M.B. Smith, B.M.W. Tsui, Principles of Medical Imaging, Academic, San Diego, CA, 1992.
T.L. Szabo, 1998, Transducer arrays for medical ultrasound imaging, in: A. Duck, A.C. Baker, H.C. Starritt (Eds.), Ultrasound in Medicine, Medical Science Series, F, Institute of Physics Pub., Bristol, UK.
T.L. Szabo, Diagnostic Ultrasound Imaging: Inside Out, Elsevier Academic, Boston, 2004.

CHAPTER

17

Biomedical Optics and Lasers

Gerard L. Coté, PhD, LiHong V. Wang, PhD, and Sohi Rastegar, PhD

OUTLINE

AT THE CONCLUSION OF THIS CHAPTER, STUDENTS WILL BE ABLE TO:

- Understand essential optical principles and fundamentals of light propagation in tissue, as well as other biological and biochemical media.
- Utilize basic engineering principles for developing therapeutic, diagnostic, sensing, and imaging devices with a focus on the use of lasers and optical technology.
- Describe the biochemical and biophysical interactions of optic and fiber optic systems with biological tissue.
- Understand the photothermal interactions of laser systems with biological tissue.

The use of light for therapeutic and diagnostic procedures in medicine has evolved from the use of sunlight for heat therapy and as a simple tool for the inspection of eyes, skin, and wounds to the current use of lasers and endoscopy in various medical procedures. The introduction of photonic technology into medicine over the last decade has revolutionized many clinical procedures and has the potential to continue to greatly impact health care. For example, coherent fiber optic bundles have been applied in laparoscopy cholecystectomy (minimally invasive removal of the gallbladder) to transform a once painful and expensive surgery into virtually an outpatient procedure. In the process, biomedical engineers have been instrumental in defining and demonstrating the engineering fundamentals of the interaction of light and heat with biological media, resulting in advances in dermatology, ophthalmology, cardiology, and urology.

Optical engineering has traditionally been taught as part of an electrical engineering curriculum and has been primarily applied in the development of defense and communications technologies. In biomedical optics applications, however, there are many issues related to the interaction of light with participating biological tissue/media that classical optics books and curricula do not cover. Therefore, the goal of this chapter is to provide students with a better understanding of the fundamental principles associated with the growing field of biomedical optics as well as advances in optically based therapeutic, diagnostic, and monitoring devices.

17.1 INTRODUCTION TO ESSENTIAL OPTICAL PRINCIPLES

Throughout the ages, great minds, including those of Galileo, Newton, Huygens, Fresnel, Maxwell, Planck, and Einstein, have studied the nature of light. Initially it was believed that light was either corpuscular or particle-like in nature or that it behaved as waves. After much debate and research, however, it has been concluded that there are circumstances where light behaves like waves and those where it behaves like particles. In this section, an overview of the behavior of light is described, beginning with the wave equation, which is commonly derived from Maxwell's equations. In addition, the optical spectrum and the fundamental governing equations for light absorption, scattering, and polarization are covered.

17.1.1 Electromagnetic Waves and the Optical Spectrum

It was in the late 1800s that J. Clerk Maxwell showed conclusively that light waves were electromagnetic in nature. This was accomplished by expressing the basic laws of electromagnetism and deriving from them the wave equation. The key to validating this derivation was that the free space solutions to the wave equation corresponded to electromagnetic waves with a velocity equal to the known experimental value of the velocity of light. Many introductory optics texts (such as the works of Hecht and Pedrotti) begin with the vectorial form for Maxwell's equations

$$\nabla^2 \mathbf{E} = \varepsilon_0 \mu_0 \, \partial^2 \mathbf{E} / \partial t^2 \tag{17.1}$$

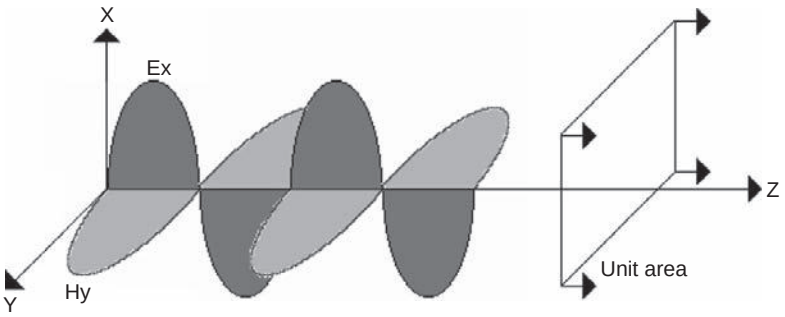


FIGURE 17.1 A sinusoidal electromagnetic wave.

where $\nabla^2\mathbf{E}$ is the partial second derivative of the electric field with respect to position ($\partial^2\mathbf{E}/dx^2 + \partial^2\mathbf{E}/dy^2 + \partial^2\mathbf{E}/dz^2$), ε is the electric permittivity of the medium, and μ is the magnetic permeability of the medium. In free space $\mu = \mu_0$ and $\varepsilon = \varepsilon_0$. From symmetry, there is a similar solution for the magnetic field (**H**). Since the wave velocity (c_0) is equal to the known value of the velocity of light, it can be expressed as follows:

$$c_0 = 1/(\varepsilon_0\,\mu_0)^{1/2} \tag{17.2}$$

Therefore, the change in the electric or magnetic field in space is related to the velocity and the change of the field with respect to time. Given Cartesian axes O_x, O_y, and O_z, shown in Figure 17.1, a typical solution to this wave equation is the sinusoidal solution given by

$$E_x = E_o \exp[j(\omega t - kz)] \tag{17.3}$$

$$H_y = H_o \exp[j(\omega t - kz)] \tag{17.4}$$

which states that the electric field oscillates sinusoidally in the X-Z plane, the magnetic field oscillates in the Y-Z plane (orthogonal to e-field and in phase), and the wave propagates in the O_z direction.

The frequency, f, and wavelength, λ, of the wave are related to the velocity, c, and given by

$$f = \omega/2\pi \tag{17.5}$$

$$\lambda = 2\pi/k \tag{17.6}$$

$$c = f\lambda = \omega/k \tag{17.7}$$

The ratio of the velocity of the wave in free space, c_o, to that in the medium through which the light propagates, c, is known as the index of refraction of the medium, $n = c_o/c$.

EXAMPLE PROBLEM 17.1

Given a red helium neon laser with a wavelength of 633 nm, determine the velocity of the light traveling through clear tissue, such as the cornea, that has an index of refraction of 1.33. How would this velocity change in glass that has an index of refraction of 1.5? Explain the significance of this result.

Continued

Solution

Rearranging the preceding equations,

$$c_{tissue} = c_0/n_{tissue} \simeq (2.998*10\text{^}8)/1.33 = 2.25*10\text{^}8 \text{ m/s for clear tissue}$$

while

$$c_{glass} = c_0/n_{glass} \simeq (2.998*10\text{^}8)/1.5 = 2.00*10\text{^}8 \text{ m/s for glass}$$

The significance of this result is that light travels faster through a material with a lower index of refraction, such as clear tissue (e.g., cornea, aqueous humor, and lens of the eye), compared to glass. This has many implications as you will see later in this chapter when it comes to determining the angle of reflection and refraction of light through differing materials and clear tissues.

The intensity or power per unit area of the wave can also be expressed in terms of the Poynting vector, **Π**—that is, defined in terms of the vector product of the electric and magnetic field as

$$\Pi = \mathbf{E} \times \mathbf{H} \tag{17.8}$$

The intensity of the wave is the average value of the Poynting vector over one period of the wave, so if E and H are spatially orthogonal and in phase, as shown in Figure 17.1, then

$$\mathbf{I} = \ < \Pi > \ = \text{power/unit area} = c\varepsilon\mathbf{E}^2 \tag{17.9}$$

where <**Π**> represents the average value of **Π**. For this equation it is clear that the intensity is proportional to the square of the electric field. This is true for the propagation of a wave through an isotropic dielectric medium. From Maxwell's equations it can be shown that the intensity is also proportional to the square of the magnetic field. Therefore, if the electric and magnetic fields were in phase quadrature (90 degrees out of phase), then the intensity would be zero, as shown following

$$\mathbf{I} =< \Pi >=< E_o \cos(\omega t)\, H_o \sin(\omega t) >= 0 \tag{17.10}$$

since the average value of a sine wave times a cosine wave is zero.

In addition to intensity, light can be characterized by its wavelength or, rather, electromagnetic spectrum, which ranges from low-frequency radio waves at 10^3 Hz to Gamma radiation at 10^{20} Hz. As depicted in Figure 17.2, the wavelength region of light can be divided into ultraviolet, which is 0.003 to 0.4 micrometers, to the visible region of 0.4 to 0.7 micrometers (7.5×10^{14} to 4.3×10^{14} Hz frequency range), to the near infrared region of 0.7 to 2.5 micrometers, to the mid-infrared region of 2.5 to 12 microns, and to the far infrared beyond 12 microns.

17.1.2 Optical Polarization

In the previous section the sinusoidal solution in terms of E_x and H_y is only one of an infinite number of such sinusoidal solutions to Maxwell's equations. The general solution for a sinusoid with angular frequency ω can be written as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}) \exp(j\omega t) \tag{17.11}$$

in which **E**(**r**,t) and **E**(**r**) are complex vectors and **r** is a real radius vector.

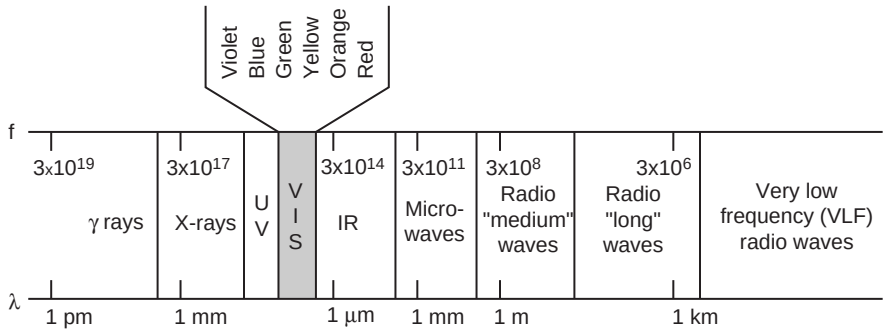


FIGURE 17.2 The electromagnetic spectrum.

For simplicity, consider the plane monochromatic (single wavelength) waves propagating in free space in the O_z direction; then the general solution to the wave equation for the electric field can be written as

$$E_x = e_x \cos(\omega t - kz + \delta_x) \tag{17.12}$$

$$E_y = e_y \cos(\omega t - kz + \delta_y) \tag{17.13}$$

in which δ_x and δ_y are arbitrary phase angles. This solution can be described completely by means of two waves (e-field in the X-Z plane and e-field in the Y-Z plane). If these waves are observed at a particular value of "z"—say, "z_o"—they take the oscillatory form

$$E_x = e_x \cos(\omega t + \delta'_x) \quad \delta'_x = \delta_x - kz_o \tag{17.14}$$

$$E_y = e_y \cos(\omega t + \delta'_y) \quad \delta'_y = \delta_y - kz_o \tag{17.15}$$

and the top of each vector appears to oscillate sinusoidally with time along a line. E_x is said to be linearly polarized in the direction O_x, and E_y is said to be linearly polarized in the direction O_y.

It can be seen from Eq. (17.15) that if the light is fully polarized, it can be completely characterized as a 2 by 1 matrix in terms of its amplitude and phase ($e_x \exp(j\delta'_x)$ and $e_y \exp(j\delta'_y)$). This vectoral representation is known as the Jones vector. When the optical system design includes the propagation of the light through nonscattering, and thus nondepolarizing, medium such as a lens or clear biological sample, the medium can be completely characterized by a 2 by 2 Jones matrix. Therefore, the output of the propagation of the polarized light can be modeled as a multiplication of the Jones matrix of the optical system and the input light vector. A system that yields both polarized and partially polarized light, such as that obtained from tissue scattering, can be characterized using 4 by 4 matrices known as Mueller matrices and a 4 by 1 matrix known as the Stokes vector.

Returning back to Eq. (17.15), the tip of the polarized light vector is the vector sum of E_x and E_y which, in general, is an ellipse as depicted in Figure 17.3 whose Cartesian equation in the X-Y plane at $z = z_o$ is

$$E_x^2/c_x^2 + E_y^2/c_y^2 + 2(E_xE_y/e_xe_y)\cos\delta = \sin^2\delta \tag{17.16}$$

in which $\delta = \delta_y' - \delta_x'$. It should be noted that the ellipse folds into a straight line, producing linear light if (1) $E_x \neq 0$ and $E_y = 0$, or (2) $E_y \neq 0$ and $E_x = 0$, or (3) $\delta = m\pi$ in which m is a positive or negative integer. The light becomes circular if $e_x = e_y$ and $\delta = (2m + 1)\pi/2$, since E_x and E_y would then have equal amplitudes and be in phase quadrature.

The polarization properties of light become particularly important for anisotropic media in which the physical properties are dependent on direction (i.e., E_x is different from E_y and thus values e_x, e_y, and δ will vary along the propagation path, as depicted with a birefringent crystal, polarization preserving optical fiber, or aligned tissue fibers such as collagen).

EXAMPLE PROBLEM 17.2

Consider a general solution to the wave equation represented as a plane electromagnetic wave (in SI units) given by the expressions

$$E_x = 0, \qquad E_z = 0, \qquad E_y = 2\cos[2\pi * 10^{14}\,(t - x/c) + \pi/2]$$

For the equations shown, determine the frequency (f), wavelength (λ), direction of motion (x, y, or z), amplitude (A), initial phase angle (ϕ), and polarization of the wave (linear, circular, elliptical).

Solution

The general form of a plane electromagnetic wave can be expressed as $E = A\cos(kx \pm \omega t + \phi)$, where $k = 2\pi/\lambda$, $\lambda = c/f$, $\omega = 2\pi f$, and $c = 3*10^8$ meters/second. The preceding equation can therefore be written in general form as $E_y = 2\cos[2\pi*10^{14}\,t - (2\pi*10^{14}\,x/c) + \pi/2]$. Thus, comparing the two equations yields the values $f = 1*10^{14}$ Hertz, $\lambda = 3*10^{-6} = 3$ micrometers, motion is in the positive x direction, A = 2 Volts/meter, $\phi = \pi/2$, and since $E_x = E_z = 0$, the wave must be linearly polarized in the *y* direction.

17.1.3 Absorption, Scattering, and Luminescence

Having discussed polarization, attention will now be focused on the optical properties of light that describe the changes in the light as it passes through biological media. The optical properties of the light are defined in terms of the absorption, scattering, and anisotropy of

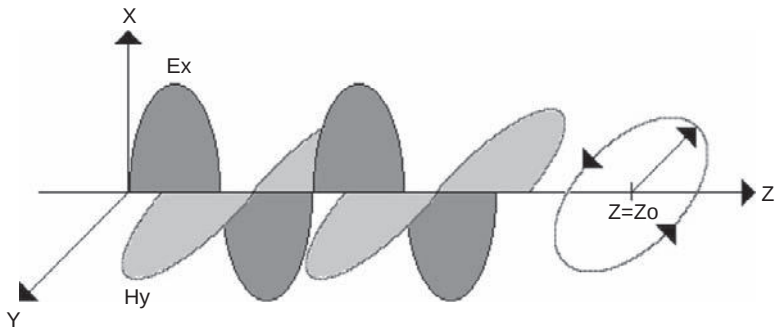


FIGURE 17.3 Polarization propagation of an **E** wave.

the light. Knowing the absorption, scattering, and direction of propagation of the light has applications for both optical diagnostics and sensing, as well as optical therapeutics. For instance, many investigators are trying to quantifiably and noninvasively measure certain body chemicals such as glucose using the absorption of light. In addition, by knowing the absorption and scattering of light, investigators are modeling the amount of coagulation or "cooking" of the tissue that is expected for a given wavelength of light, such as in the case of prostate surgery or cancer therapy.

The optical penetration depth and the volume of tissue affected by light is a function of both the light absorption and scattering. If it is assumed scattering is negligible or, rather, absorption is the dominant component, then the change in light intensity is determined by the absorption properties in the optical path and the illumination wavelength. For instance, for wavelengths above 1.5 micrometers, the light propagation through all biological tissues, since they have large water content, is absorption dominant. The intensity relationship as a function of optical path and concentration for this absorption dominant case can be described by the Beer-Lambert law as

$$I(z) = I_o \exp(-\mu_a z) \tag{17.17}$$

in which I is the transmitted light intensity, I_o is the incident light intensity, z is the path length of the light, and μ_a (1/cm) is the absorption coefficient. The absorption coefficient can be further broken down into two terms that include the concentration of the chemically absorbing species, C (mg/mL), times the molar absorptivity, v (cm^2/mg), which is a function of wavelength.

Luminescence, also known as fluorescence or phosphorescence, depending on the fluorescent lifetime, is a process that occurs when photons of electromagnetic radiation are absorbed by molecules, raising them to some excited state, and then, on returning to a lower energy state, the molecules emit radiation or rather luminesce. Fluorescence does not involve a change in the electron spin and therefore occurs much faster. The energy absorbed in the luminescence process can be described in discrete photons by the equation

$$E = hc/\lambda \tag{17.18}$$

in which E is the energy, h is Planck's constant (6.626×10^{-34} J-s), c is the speed of light, and λ is the wavelength. This absorption and reemission process is not 100 percent efficient, since some of the originally absorbed energy is dissipated before photon emission. Therefore, the excitation energy is greater than that emitted

$$E_{\text{excited}} > E_{\text{emitted}} \tag{17.19}$$

and

$$\lambda_{\text{emitted}} > \lambda_{\text{excited}} \tag{17.20}$$

The energy level emitted by any fluorochrome is achieved as the molecule emits a photon, and because of the discrete energy level, individual fluorochromes excited by a given wavelength typically only fluoresce at a certain emission wavelength. Originally fluorescence was used as an extremely sensitive, relatively low-cost technique for sensing dilute solutions. Only recently have investigators used fluorescence in turbid media such as tissue for diagnostics and sensing.

It is well known that tissues scatter as well as absorb light, particularly in the visible and near-infrared wavelength regions. Scattering, unlike absorption and luminescence, need not involve a transition in energy between quantized energy levels in atoms or molecules but rather is typically a result of random spatial variations in the dielectric constant. Therefore, the actual light distributions can be substantially different from distributions estimated using Beer's law. In fact, light scattered from a collimated beam undergoes multiple scattering events as it propagates through tissue. The transport equation that describes the transfer of energy through a turbid medium (a medium that absorbs and scatters light) is an approach that has been proven to be effective. The transport theory has been used to describe scattering, absorption, and fluorescence, but until recently, polarization has not been included, in part because it takes relatively few scattering events to completely randomize the polarization of the light beam. The transport theory approach is discussed later in this chapter and is a heuristic theory based on a statistical approximation of photon particle transport in a multiple scattering medium.

17.2 FUNDAMENTALS OF LIGHT PROPAGATION IN BIOLOGICAL TISSUE

In this section the propagation of light through biological media such as tissue is discussed, beginning with a simple ray optics approach for light traveling through a nonparticipating media, where the effect of the absorption and scatter within the media is ignored. The effects of absorption and scatter on light propagation are then discussed, along with the consideration of boundary conditions and various means of measuring the optical absorption and scattering properties.

17.2.1 Light Interactions with Nonparticipating Media

Previously we defined *light* as a set of electromagnetic waves traveling in the O_z direction. In this section light is treated as a set of "rays" traveling in straight lines that, when combined, make up the plane waves described using the preceding complex exponentials. This approach is important in order to study the effect on the light at the boundaries between two different optical media. In the treatment of geometrical or ray optics of the light, it is first assumed that the incident, reflected, and refracted rays all lie in the same plane of incidence. The propagation at the boundary between these two interfaces, as shown in Figure 17.4, is based on two basic laws—namely, the angle of reflection equals the angle of incidence, and the sine of the angle of refraction bears a constant ratio to the sine of the angle of incidence (Snell's law). It should be noted that ray propagation can be very useful for a good majority of the applications considered with the propagation of the light through bulk optics such as in the design of lenses and prisms. It has severe limitations, however, in that it cannot be used to predict the intensities of the refracted and reflected rays, nor does it incorporate the effects of scatter and absorption. In addition, when apertures are used that are smaller than the bulk combined rays, a wave process known as diffraction occurs that causes the geometrical theory to break down and the beams to diverge.

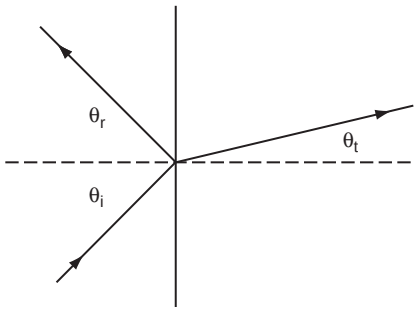


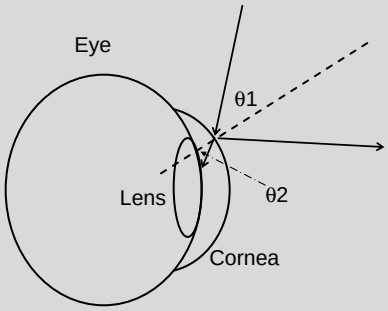
FIGURE 17.4 Ray propagation of light at the boundary of two interfaces. Clearly the angle of incidence equals the angle of reflection ($\theta_i = \theta_r$) and from Snell's law the sine of the angle of refraction bears a constant ratio to the sine of the angle of incidence ($n_i \sin(\theta_i) = n_t \sin(\theta_t)$).

EXAMPLE PROBLEM 17.3

A green light beam hits the cornea of the eye, making an angle of 25 degrees with the normal, as depicted following.

(a) Determine the output angle from the front surface of the cornea and into the cornea, given that the index of refraction of air is 1.000, and that of the cornea is 1.376.

(b) What can be said about how the cornea bends the light?



Solution

(a) Snell's law, n1 sin(θ1) = n2 sin(θ2), can be rearranged to calculate θ2 in the lens.

$$\theta 2 = \sin^{-1}(1.000 \sin(25)/1.376) = 17.89 \text{ degrees}$$

(b) It has been shown and can be said that the cornea tends to bend the light toward the normal as it passes through. This makes sense because the eye is made to bend the light so it can pass through the center iris and lens toward the retina to be imaged by the brain.

Wave theory is used to describe the phenomena of reflection and refraction in an effort to determine the intensity of the light as it propagates from one medium to another. If two nonconducting media are considered, as in Figure 17.4, the boundary conditions follow from Maxwell's equations such that the tangential components of **E** and **H** are continuous across the boundary $\mathbf{E}_i + \mathbf{E}_p = \mathbf{E}_t$, and the normal components of

B (the magnetic flux density) and **D** (the electric displacement) are continuous across the boundary. These conditions can only be true at all times and all places on the boundary if the frequencies of all waves are the same. The relative amplitudes of the waves will now be considered (note the values of **E**, **H**, **D**, and **B** will depend on the direction of vibration of **E** and **H** fields of the incident wave relative to the planes of incidence). In other words, they depend on the polarization of the wave. The wave can be divided into two orthogonal polarization states, and using Maxwell's equations, Snell's law, and the law of reflection as well as imposing the boundary conditions, the following Fresnel equations can be derived. The reflected and transmitted equations in the parallel configuration are

$$r_{\mathrm{p}} = \left(\frac{E_r}{E_i}\right)_{\mathrm{p}} = (-n_i \cos\theta_{\mathrm{t}} + n_t \cos\theta_{\mathrm{i}})/(n_i \cos\theta_{\mathrm{t}} + n_t \cos\theta_{\mathrm{i}}) \tag{17.21}$$

$$t_{\mathrm{p}} = \left(\frac{E_t}{E_i}\right)_{\mathrm{p}} = (2n_i \cos\theta_{\mathrm{i}})/(n_i \cos\theta_{\mathrm{t}} + n_t \cos\theta_{\mathrm{i}}) \tag{17.22}$$

The normal or perpendicular reflected and transmitted equations are

$$r_{\perp} = \left(\frac{E_r}{E_i}\right)_{\perp} = (n_i \cos\theta_{\mathrm{i}} - n_t \cos\theta_{\mathrm{t}})/(n_i \cos\theta_{\mathrm{i}} + n_t \cos\theta_{\mathrm{t}}) \tag{17.23}$$

$$t_{\perp} = \left(\frac{E_t}{E_i}\right)_{\perp} = (2n_i \cos\theta_{\mathrm{i}})/(n_i \cos\theta_{\mathrm{i}} + n_t \cos\theta_{\mathrm{t}}) \tag{17.24}$$

Using the preceding equations, there are two noteworthy limiting cases. The first case is as follows. If $n_t > n_i$, then $\theta_{\mathrm{i}} > \theta_{\mathrm{t}}$, $r_{\perp}$ is always negative for all values of θ_{i}, and r_{p} starts out positive at $\theta_{\mathrm{i}} = 0$ and decreases until it equals 0 when $\theta_{\mathrm{i}} + \theta_{\mathrm{t}} = 90^{\mathrm{o}}$. This implies that the refracted and reflected rays are normal to each other, and from Snell's law this occurs when $\tan\theta_{\mathrm{i}} = n_t/n_i$. This particular value of θ_{i} is known as the polarization angle or, more commonly, the Brewster angle. At this angle only the polarization with **E** normal to the plane of incidence is reflected, and as a result, this is a useful way of polarizing a wave. Now, as θ_{i} increases beyond θ_{p}, r_{p} becomes progressively negative, reaching –1.0 at 90^{o}, which implies that the surface performs as a perfect mirror at this angle. On the other hand, at normal incidence $\theta_{\mathrm{i}} = \theta_{\mathrm{r}} = \theta_{\mathrm{t}} = 0$, then $t_{\mathrm{p}} = t_{\perp} = 2n_i/(n_i + n_t)$ and $r_{\mathrm{p}} = -r_{\perp} = (n_t - n_i)/(n_i + n_t) = -(n_i - n_t)/(n_i + n_t)$. The wave intensity, which is what can actually be measured by a detector, is defined proportional to the square of the electric field. Thus, $(I_{\mathrm{R}}/I_i) = (r_{\mathrm{p}})^2 = \left(\frac{E_r}{E_i}\right)^2 = (n_t - n_i)^2/(n_i - n_t)^2$ and $(I_t/I_i) = (t_{\mathrm{p}})^2 = \left(\frac{E_t}{E_i}\right)^2 = 4n_i^2/(n_i + n_t)^2$. The preceding expressions are useful because they represent the amount of light lost by normal specular reflection when transmitting from one medium to another.

The second limiting case is, if $n_t < n_i$, then $\theta_{\mathrm{t}} > \theta_{\mathrm{i}}$, $r_{\perp}$ is always positive for values of θ_{i}, and $r_{\perp}$ increases from its initial value at $\theta_{\mathrm{i}} = 0$ until it reaches +1.0 at what is known as the critical angle θ_{c}. At this angle $\theta_{\mathrm{i}} = \theta_{\mathrm{c}}$, then $\theta_{\mathrm{t}} = 90^{\mathrm{o}}$, and from Snell's law, $\sin\theta_{\mathrm{t}} = n_i/n_t \sin\theta_{\mathrm{i}}$. Clearly for $n_i > n_t$ (less dense to more dense medium), $\sin\theta_{\mathrm{t}}$ could be greater than one, according to the preceding equation. This is not possible for any real value of θ_{t},

and thus the angle for which $\theta_i = \theta_c$ is when $(n_i/n_t) \sin \theta_c = 1$ or, rather, $\theta_t = 90^o$. Thus, for all values of $\theta_i > \theta_c$, the light is totally reflected at the boundary. This phenomenon is what is used for the creation of optical fibers, in which the core of the fiber (where the light is supposed to propagate) has an index of refraction slightly higher than the cladding surrounding it, so the light launched into the fiber will totally internally reflect allowing for minimally attenuated propagation down the fiber.

17.2.2 Light Interaction with Participating Media: The Role of Absorption and Scattering

When light is incident on tissue either from a laser or from other light transmitting devices, tissue acts as a participating medium by reflecting, absorbing, scattering, and transmitting various portions of the incident wave of radiation. Ideally, an electromagnetic analysis of the light distribution in the tissue would be performed. Unfortunately, this can be quite cumbersome, and a reliable database of electrical properties of biological tissues would be required. An alternate practical approach to the problem is to use transport theory that starts with the construction of the differential equation for propagation of the light intensity. In the following section, the equation for light intensity distribution in a purely absorbing medium will be derived. With this motivation, the next section introduces the general equation of transfer for a medium that scatters as well as absorbs the light.

The Case of Pure Absorption

In order to describe the distribution and transport of laser energy in a nontransparent participating medium, the medium can be viewed as having two coexisting "phases": a material phase for all the masses of the system and a photon phase for the electromagnetic radiation. Figure 17.5 shows the material phase as circles and the photon phase as curved arrows that strike the material phase. The energy balance equation for the material phase is introduced and discussed in the thermodynamic descriptions in Sections 17.3 and 17.5. The energy balance equation for the photon phase is discussed following.

Consider an infinitesimal volume of the material under irradiation (Figure 17.6). The rate of change of radiative energy, $U^{(rad)}$, with time is the difference between the incoming and outgoing radiative fluxes in the element minus the rate of energy absorption by the material phase. The difference between the incoming and outgoing fluxes is, in the limit, the negative of the divergence of the radiative flux. Therefore, denoting the energy absorbed by the material phase, which is the laser heat source term, by Q_L and the radiative flux by $\vec{q}^{(rad)}$ the governing differential equation for energy rate balance in the photon phase can be written as

$$\partial U^{(rad)} / \partial t = -\vec{\nabla} \cdot \vec{q}^{(rad)} - Q_L \tag{17.25}$$

Any possible photon emission or scattering by the material phase has been ignored.

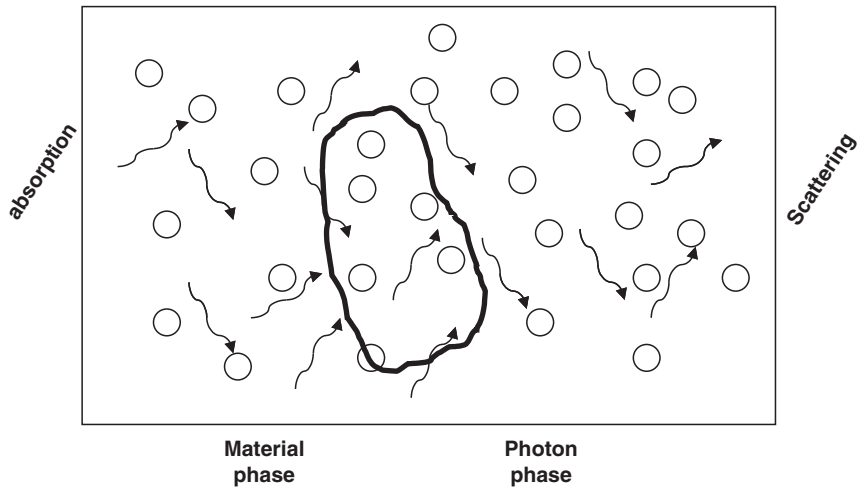


FIGURE 17.5 Two coexisting phases, the photon phase (laser) and the material phase (tissue) govern laser-tissue interactions. Photons are distributed within the tissue by the processes of absorption and scattering. Energy balance over a control volume (shown with thicker line) provides the governing equation for light and heat distribution and exchange of energy between the photon and material phase as described later in the chapter.

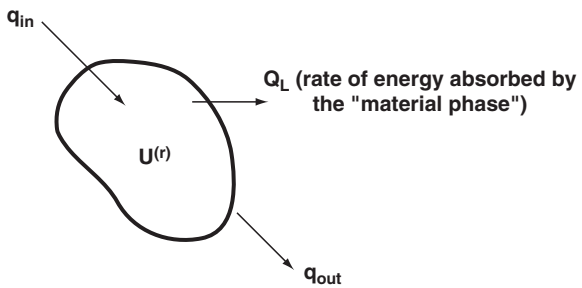


FIGURE 17.6 Control volume of the photon phase with radiative fluxes in and out, radiative energy, and rate of energy lost to the material phase by absorption.

For a steady-state system in which radiation is collimated and monochromatic, laser light travels only in the positive z direction—that is, assuming only one forward flux—Eq. (17.25) reduces to

$$dq_z^{(rad)}/dz = -Q_L. \tag{17.26}$$

Considering the speed of light and the dimensions of a biological tissue, the assumption of a steady-state condition is a reasonable assumption for most applications except when very fast light sources are used and/or when time-resolved analyses are considered. The important step at this point is to use the phenomenological relation

$$Q_L = \mu_a I \tag{17.27}$$

where μ_a is defined as the absorption coefficient and I is the total light intensity. The total intensity at a point is the sum of radiative fluxes received at that point. In the case presented here, using purely absorbing tissue, a single radiative flux is used, and thus,

$$I = q_z^{(rad)} \tag{17.28}$$

Proceeding by replacing Eqs. (17.27) and (17.28) into Eq. (17.26), the following differential equation is obtained

$$\frac{dI}{dz} = -\mu_a I \tag{17.29}$$

which has the simple solution

$$I(z) = I_0 \exp(-\mu_a z) \tag{17.30}$$

where I_0 denotes the intensity at the surface, $z = 0$. This equation is the well-known *Beer-Lambert* law of absorption, as given in Eq. (17.17) for a purely absorbing medium, which was mentioned earlier in the chapter.

The laser heat source term can now be written (Eqs. (17.27) and (17.30)) as

$$Q_L = \mu_a I_0 \exp(-\mu_a z). \tag{17.31}$$

This equation can be generalized to an axisymmetric three-dimensional case to include the effect of the radial beam profile of a laser light incident orthogonally on a slab by writing it as

$$Q_L(r, z) = \mu_a I_0 \exp(-\mu_a z) f(r), \tag{17.32}$$

where $f(r)$ is the radial profile of an axisymmetric laser beam. For a Gaussian beam profile, which is a common mode of laser irradiation,

$$f(r) = \exp\left(-\frac{2r^2}{\omega_o^2}\right) \tag{17.33}$$

where ω_o is known as the "$1/e^2$ radius" of the beam, since at $r = \omega_o$, $f(r) = 1/\mathrm{e}^2$.

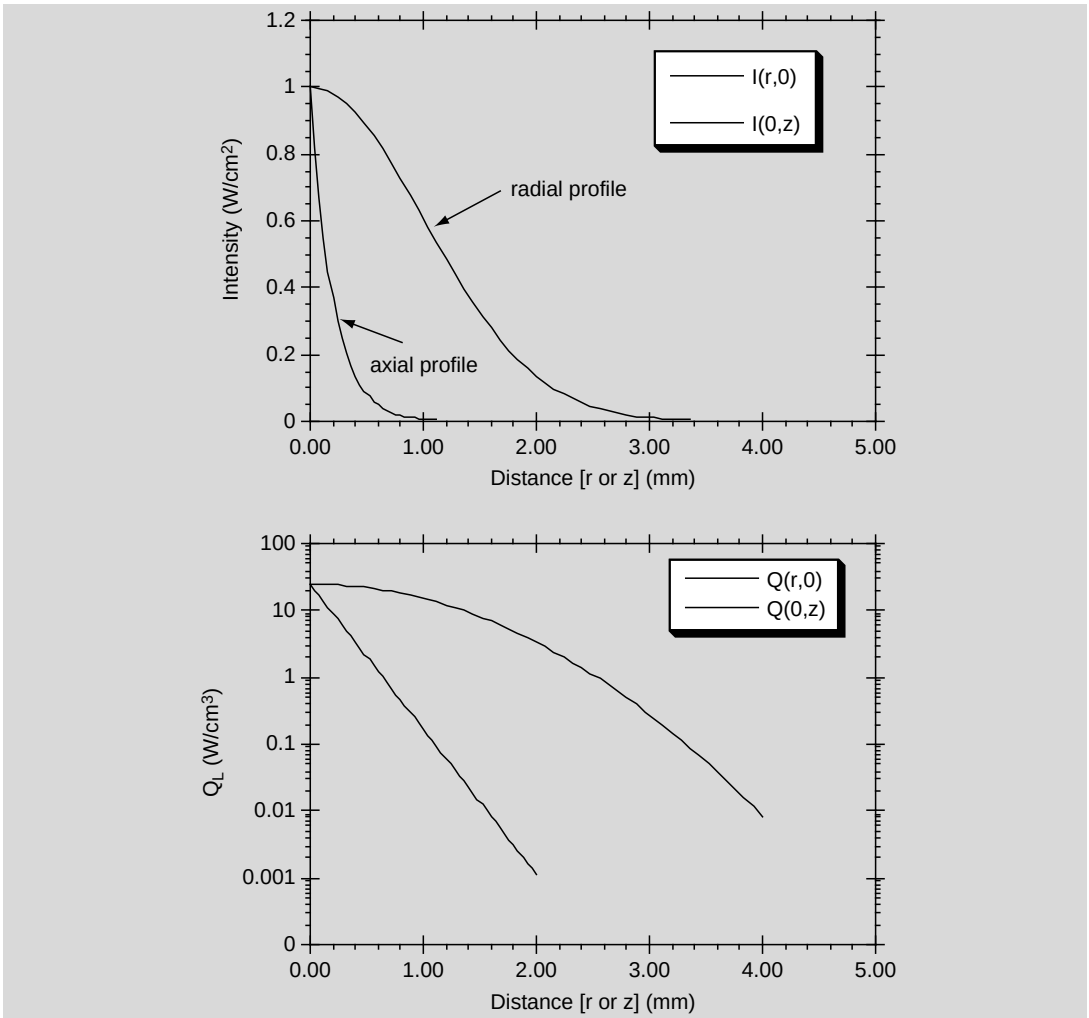
EXAMPLE PROBLEM 17.4

For a Gaussian laser beam irradiating at a wavelength of 2.1 μm, the absorption coefficient is estimated to be 25 cm^{-1} and scattering is negligible. The radial profile of light intensity and the rate of heat generation at the tissue surface, $z = 0$, and its axial profile along the center axis of the beam, $r = 0$, can be found using Eqs. (17.32) and (17.33). Graph $I(r, 0)$ for r ranging from $-2\omega_o$ to $+ 2\omega_o$ and $I(0, z)$ for $z = 0$ to $z = 5/\mu_a$, as well as graph $Q_L(r, 0)$ and $Q_L(0, z)$. For simplicity, $I_o = 1\ \mathrm{W/cm}^2$ can be used.

Solution

Plot of the radial and axial profile of light intensity, I, and volumetric rate of absorption, Q_L, in a purely absorbing tissue.

Continued



Methods for Scattering and Absorbing Media

The fundamental quantity in the transport theory approach is the total specific intensity, L, which is a function of position $\vec{r}$ for light in the direction given by a unit vector $\hat{s}$ and its units are $W\ .cm^{-2}\ .s^{-1}$. The equation of transfer satisfied by L can be written as

$$\frac{dL(\vec{\mathrm{r}},\hat{\mathrm{s}})}{ds} = -(\mu_{\mathrm{a}}+\mu_{\mathrm{s}})L(\vec{\mathrm{r}},\hat{\mathrm{s}}) + \mu_{\mathrm{s}}\int_{4\pi} p(\hat{\mathrm{s}},\hat{\mathrm{s}}')L(\vec{\mathrm{r}},\hat{\mathrm{s}}')d\omega' + \mathrm{S}(\vec{\mathrm{r}},\hat{\mathrm{s}}) \tag{17.34}$$

Equation (17.34) shows the decrease in L due to scattering μ_s and absorption μ_a and the increase in L due to scattering from L coming from another direction $\hat{s}'$ (Figure 17.7). The sum

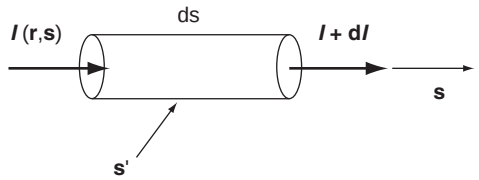


FIGURE 17.7 Light intensity change in the transport theory approach including scattered light from other directions.

of scattering and absorption, $\mu_t = \mu_a + \mu_s$, is defined as the attenuation coefficient. The function $p(\hat{s}, \hat{s}')$ is called the "phase function" and is related to the scattering amplitude of a particle, a scaled form of the probability distribution of scattering angles. Note that by ignoring scattering and assuming a collimated light source, Eq. (17.34) reduces to Eq. (17.29), which was derived in the last subsection. $S(\vec{r}, \hat{s})$ is the source term that could be irradiation on the surface, fluorescence generated inside the tissue, or an internal volumetric light source.

The equation of transfer is an integro-differential equation for which a general solution does not exist. However, several approximate solutions have been found, such as the two-flux and multiflux models, the discrete ordinate finite element method, the spherical harmonic method, the diffusion approximation, and the Monte Carlo method. Each of these is subject to certain limitations and assumptions. In this chapter the focus is on the diffusion approximation, which is one solution.

Diffusion Approximation

The diffusion approximation is a second-order differential equation that can be derived from the radiative transfer equation (17.34) under the assumption that the scattering is "large" compared with absorption. The solution to this equation provides a useful and powerful tool for the analysis of light distribution in turbid media. The governing differential equation for the diffusion approximation is

$$\nabla^2 \phi_d - 3\mu_a \mu_t' \phi_d = -\frac{\mu_s'}{D} \phi_c \tag{17.35}$$

where ϕ_d is the diffuse fluence rate and the parameters of the equation are $\mu_s' = \mu_s(1 - g)$, $\mu_t' = \mu_a + \mu_s'$ and $D = 1/3\ \mu_t'$, in which g is defined as the anisotropy of the medium.

The total light fluence rate, ϕ [W/cm^2], is the sum of the collimated part, ϕ_c, and the diffuse part, ϕ_d. The total fluence rate, as given by the following equation, is a key parameter in laser-tissue interaction. It can be thought of as the total light received at a point in space, or within the medium, through a small sphere divided by the area of that sphere.

$$\phi(r, z) = \phi_c(r, z) + \phi_d(r, z). \tag{17.36}$$

The collimated fluence rate is given by

$$\phi_c = I_o(r)\ (1 - r_{sp})\ \exp(-\mu_t' z), \tag{17.37}$$

where I_o is the surface flux density [W/cm^2] of the incident beam and r_{sp} is the specular reflectance. The optical boundary condition at the beam axis ($r = 0$) is

$$\left.\frac{\partial \phi_d}{\partial r}\right|_{r=0} = 0. \tag{17.38}$$

The boundary condition elsewhere is

$$\phi_d - 2AD\,\nabla\phi_d \bullet \tilde{n} = 0 \tag{17.39}$$

where A is the internal reflectance factor and $\tilde{n}$ is the inward unit normal vector. The internal reflectance factor A can account for the effect of mismatch in the index of refraction between the boundary and the surrounding medium and is given by

$$A = \frac{1 + r_i}{1 - r_i} \tag{17.40}$$

where r_i is evaluated by an empirical formula

$$r_i = -1.440\, n_{\text{rel}}^{-2} + 0.710\, n_{\text{rel}}^{-1} + 0.688 + 0.0636\, n_{\text{rel}}, \tag{17.41}$$

and n_{rel} is the ratio of the refractive indices of the tissue and the medium. The internal reflectance factor A reduces to 1 in cases where the boundary is matched—that is, when $n_{\text{rel}} = 1$.

Equations (17.35) through (17.41) provide the governing differential equations and boundary condition for the diffusion approximation. A solution of these, either by analytical or numerical methods, allows the calculation and analysis of fluence rates within a scattering and absorbing media such as biological tissues. As an example, for an isotropic point source inside an infinite medium, the solution for fluence rate as measured by a detector embedded inside the medium at a "large" distance r from the fiber can be derived using the Green's functions solution of the preceding equations to be

$$\phi(r) = \frac{\phi_o}{4\pi D}\,\frac{e^{-r/\delta}}{r} \tag{17.42}$$

where δ ($= \sqrt{D/\mu_a}$) is the penetration depth.

EXAMPLE PROBLEM 17.5

It is desired to measure the concentration of an absorber in a scattering medium with known scattering coefficient. If the reduced scattering coefficient μ_s' is known and the relative intensity at a distance r_o from an isotropic source can be measured, solve an algebraic equation for the absorption coefficient based on the diffusion approximation given r_o is large enough for diffusion approximation to be valid.

Solution

Given scattering coefficients μ_s', r_o, and $\phi(r_o)/\phi_o$, an algebraic equation for μ_a based on the diffusion approximation is required.

$$\begin{aligned} \phi(\mathrm{r}) &= \phi_o \exp(-\mathrm{r}/\delta)/(4\,\pi\,\mathrm{D}\,\mathrm{r}) \\ \log[\phi(\mathrm{r})/\phi_\mathrm{o}] &= -\mathrm{r}/\delta - [\log 4 + \log \pi + \log \mathrm{D} + \log \mathrm{r}] \\ \mathrm{D} &= 1/3\mu_{t'} \\ \delta &= \sqrt{\mathrm{D}/\mu_\mathrm{a}} \end{aligned} \tag{ex.1}$$

The left-hand side of Eq. (ex.1) is a known value—say, k_1:

$$\begin{aligned} \therefore \mathrm{k}_1 &= -\mathrm{r}/\delta - [\log 4 + \log \pi + \log \mathrm{D} + \log \mathrm{r}] \\ \therefore \mathrm{k}_1 + \log 4 + \log \pi + \log \mathrm{r} &= -\mathrm{r}/\delta - \log \mathrm{D} \end{aligned} \tag{ex.2}$$

The left-hand side of Eq. (ex.2) is again a known value—say, k_2:

$$\begin{aligned} \therefore \mathrm{k}_2 &= -\mathrm{r}/(\sqrt{\mathrm{D}/\mu_a}) - \log \mathrm{D} \\ \therefore \mathrm{k}_2 &= -\mathrm{r}\,(\sqrt{\mu_a/\mathrm{D}}) - \log \mathrm{D} \\ \therefore \mathrm{k}_2 &= -\mathrm{r}\,(\sqrt{\mu_a/1}/(3\mu_t')) - \log(1/3\mu_t') \\ \therefore \mathrm{k}_2 &= -\mathrm{r}\,(\sqrt{3\mu_a\mu_t'}) - \log\,(1/3\mu_t') \\ \therefore \mathrm{k}_2 &= -\mathrm{r}\,(\sqrt{3\mu_a(\mu_a+\mu_s')}) - \log\,(1/3(\mu_a+\mu_s')) \end{aligned} \tag{ex.3}$$

Solving Eq. (ex.3), the value of the coefficient of absorption can be found.

17.2.3 Measurement of Optical Properties

The measurement of optical properties—namely, absorption coefficient (μ_a), scattering coefficient (μ_s), and scattering anisotropy (g)—of biological tissues is an important problem in the field of biomedical optics. Knowledge of these parameters is important in both therapeutic and diagnostic applications of light in medicine. For example, optical properties are required to predict fluence distributions for irradiation procedures such as photodynamic therapy, photocoagulation, and tissue ablation. Also, in addition to having a profound impact on in vivo diagnostics such as fluorescence spectroscopy and optical imaging, the optical properties themselves can potentially be used to provide metabolic information and diagnose diseases.

To date, a number of methods have been developed to measure tissue optical properties. The collimated transmission technique can be used to measure the total interaction coefficient ($\mu_a + \mu_s$). In this technique, a collimated light beam illuminates a thin piece of tissue. Unscattered transmitted light is detected, while the scattered light is rejected by use of a small aperture. The unscattered transmitted light can be calculated based on the Beer-Lambert law, which is an extension of Eq. (17.30). The Beer-Lambert law for absorbing and scattering media is $I(z) = \mathrm{I}_0 \exp[-(\mu_a + \mu_s)\, z]$, where $I(z)$ is the unscattered transmitted light intensity after penetrating a depth of z. In collimated transmission measurements, I_0, $I(z)$, and z are measured. Therefore, $\mu_a + \mu_s$ can be deduced.

EXAMPLE PROBLEM 17.6

A 5 mW collimated laser beam passes through a 4 cm nonabsorbing, scattering medium. The collimated transmission was measured through a small aperture to be 0.035 mW. Calculate the scattering coefficient.

Solution

Based on Beer's law, $\mu_s = \ln(5/0.035)/4 = 1.2\ \text{cm}^{-1}$.

One of the earliest techniques developed for experimentally measuring absorption and scatter is the integrating sphere measurement method. In this technique, a thin slice of tissue is sandwiched between two integrating spheres (spheres with an entrance and exit port whose inner surface is coated with a diffuse reflecting material). A collimated beam is incident upon the tissue sample. Both the diffuse reflectance and transmittance are measured by integrating the diffusely reflected and transmitted light, respectively. These two measurements are used to deduce the absorption coefficient (μ_a) and the reduced scattering coefficient (μ_s') with a model. The model could be based on the adding-doubling method, the delta-Eddington method, the Monte Carlo method, or other light transport theories.

Another technique is normal incidence video reflectometry. A collimated light beam is normally incident upon a tissue. The spatial distribution of diffuse reflectance is collected using either a CCD camera or an optical fiber bundle. Diffusion theory is used to fit the measured spatial distribution of diffuse reflectance to determine the optical properties. The measured spatial distribution of diffuse reflectance must be in absolute units unless total diffuse reflectance is measured along with the diffuse reflectance profile. Calibration to absolute units is a sensitive procedure, so this method is not ideal for a clinical setting.

It is possible to use time-resolved or frequency-domain techniques to measure optical properties. But these techniques require instrumentation that may not be cost effective for nonresearch applications. One promising approach for in vivo optical property measurements is fiber optic–based oblique incidence reflectometry. It is a fairly accurate method for measuring the absorption and reduced scattering coefficients, μ_a and μ_s', providing the sample can be regarded as a semi-infinite turbid media, as is the case for most in vivo tissues. Therefore, this approach will be covered in more detail in this section.

Obliquely incident light produces a spatial distribution of diffuse reflectance that is not centered about the point of light incidence. The amount of shift in the center of diffuse reflectance is directly related to the medium's diffusion length, D. A fiber optic probe may be used to deliver light obliquely and sample the relative profile of diffuse reflectance. Measurement in absolute units is not necessary. From the profile, it is possible to measure D, perform a curve fit for the effective attenuation coefficient, μ_{eff}, and then calculate μ_a and μ_s'. Here, μ_{eff} is defined as

$$\mu_{eff} = \sqrt{\mu_a / D} \tag{17.43}$$

The spatial distribution of diffuse reflectance of normally incident light has previously been modeled using two isotropic point sources, one positive source located 1 transport

mean free path (mfp′) below the tissue surface and one negative image source above the tissue surface. The positive source represents a single scatter source in the tissue, and the height in z of the image source depends on the boundary condition. The transport mean free path is defined as

$$mfp' = 1/(\mu_a + \mu_s') \tag{17.44}$$

With oblique incidence, the buried source should be located at the same path length into the tissue, with this distance now measured along the new optical path as determined by Snell's law. It is assumed that (a) the angle of incidence and (b) the indexes of refraction for both the tissue and the medium through which the light is delivered are known. The net result is a change in the positions of the point sources, particularly a shift in the x direction. These two cases are shown in Figures 17.8a and b.

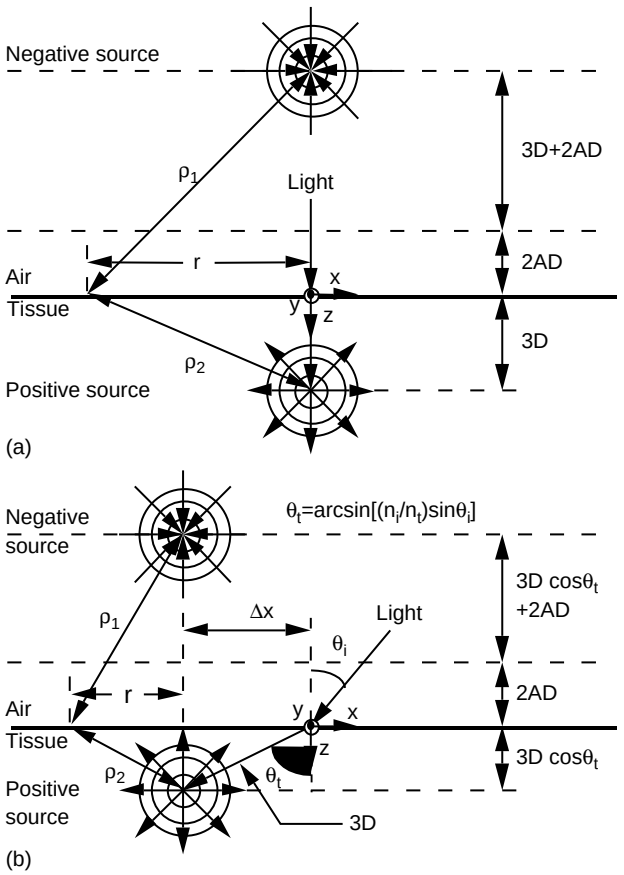


FIGURE 17.8 (a) Positions of point sources in diffusion theory model for normal incidence. (b) Positions of point sources in diffusion theory model for oblique incidence. The y-axis points out of the page. r_1 and r_2 are the distances from the positive and negative point sources to the point of interest on the tissue surface, at a radius of r from the axis of the sources. θ_t is determined from Snell's law.

The diffuse reflectance profile for oblique incidence is centered about the position of the point sources, so the shift, Δx, can be measured by finding the center of diffuse reflectance relative to the light entry point. The two source model gives the following expression:

$$R(x) = \frac{1}{4\pi}\left[\frac{\Delta z(1+\mu_{eff}\rho_1)\exp(-\mu_{eff}\rho_1)}{\rho_1^3} + \frac{(\Delta z + 2z_b)(1+\mu_{eff}\rho_2)\exp(-\mu_{eff}\rho_2)}{\rho_2^3}\right] \tag{17.45}$$

where $\Delta z = 3D\cos\theta_t$. Equation (17.45) can be scaled arbitrarily to fit a relative reflectance profile that is not in absolute units. The effective attenuation coefficient, μ_{eff}, is defined previously, while ρ_1 and ρ_2 are the distances from the two sources to the point of interest (the point of light collection; see Figure 17.8). As can be seen in Figure 17.8b, the diffusion coefficient can be calculated from Δx:

$$D = \Delta x/(3\ \sin\theta) \tag{17.46}$$

Optical properties of biological tissues depend on the tissue type and optical wavelength. For instance, the liver, with its reddish color, would have a much higher absorption coefficient from a green light source such as an argon laser than a tan piece of tissue such as chicken breast. Depending on the tissue type and exact wavelength, a typical set of optical properties for visible or near-infrared light would be 0.1 cm^{-1} for the absorption coefficient, 100 cm^{-1} for the scattering coefficient, and 0.9 for the anisotropy. In the UV region, light absorption is dominated by proteins. In the IR region, light absorption is dominated by water. The near IR (~730 nm) is considered a diagnostic window because of the minimal absorption and relatively low scattering in this region.

An important application of optical properties is the measurement of hemoglobin oxygen saturation, which is a critical physiological parameter. Because the oxygenated and deoxygenated hemoglobin molecules have different absorption spectra, the relative concentration ratio between the two forms of hemoglobin can be calculated once the optical properties of the tissue are measured.

17.3 PHYSICAL INTERACTION OF LIGHT AND PHYSICAL SENSING

After reading Section 17.2, it becomes quite apparent that most of the tissues in our bodies are not transparent to light—that is, light is typically either absorbed or scattered in tissues. In this section the physical interaction of the light with the tissue is described and its use in both sensing and therapeutics. The four fundamental physical interactions that will be described are the thermal changes induced or measured using light, pressure changes induced or measured using light, velocity changes manifested as Doppler frequency shifts in the light, and path length changes in the sample, which causes an interference pattern between two or more light beams.

17.3.1 Temperature Generation and Monitoring

All tissues in our body absorb light at various wavelengths, and the absorption process can convert the light energy into heat. In addition, our body tissues, as with any object above absolute zero temperature, also generate light radiation known as blackbody

radiation. Thus, light can be utilized to heat tissues for therapy or be measured from tissues to determine temperature.

Temperature Monitoring

The measurement of temperature, in general, has been traditionally performed reliably and fairly inexpensively using electrically based sensors such as thermistors or thermocouples. However, these sensors are not as useful when strong electromagnetic interference is present, which is common in a hospital-based setting. Also, when attempting to monitor temperature rise due to laser radiation, these sensors are inappropriate because they can absorb the laser radiation and thus can act as heat sinks. Therefore, to measure temperature in the body, several indirect optic and fiber optic approaches have been reported. For example, liquid crystal optrodes show a dramatic change with color due to temperature differences, interferometric sensors change phase or fringes with a change in the path length due to temperature variations, and the luminescence of many materials depends strongly on temperature.

The direct type of temperature sensor initially described is known as a thermographic or radiometer system. The primary application and engineering design for this type of system has been done by the military for tasks such as detecting vehicles, personnel, and even ships in total darkness. The formula for the total emission from a blackbody at temperature T is

$$I = \varepsilon \sigma T^4 \tag{17.47}$$

where ε is the emissivity and σ is the Stefan-Boltzmann constant. At room temperature, objects emit mainly in the far infrared region of the wavelength spectrum, but as temperature rises, the emission appears in the near infrared and finally in the visible spectrum. The emissivity is one ($\varepsilon = 1$) for biological tissue, and thus the total emission is dependent only on the temperature T. For the military applications, a thermographic picture generated from the IR emission can be used to display the temperature, and using this same technique, the surface temperature of the human body can be monitored for medical applications. For example, a thermal camera may be used with a TV monitor to reveal the temperature distribution on the human chest in an effort to reveal the pathologic condition of inflammation and/or breast cancer, since the affected tissue will be slightly warmer than the healthy tissue. It should be mentioned that although IR imaging is less reliable than x-ray mammography for breast screening, it is totally passive and the patients are not exposed to ionizing radiation. The thermographic systems used for this IR emission measurement require a line of sight between the warm surface and the detector. When no direct line of sight exists, an infrared-transmitting optical fiber can sometimes be used to make the connection to the detector. This type of fiber optic radiometer has been proposed for treatments using microwave heating and laser tissue heating treatments.

Light-Induced Heating

As mentioned previously, in addition to monitoring the light radiation from the body, light itself can be used to heat tissues that absorb the light. Light-induced heating of tissue can be used for a variety of applications, including biostimulation, sealing or welding

blood vessels, tissue necrosis, or tissue vaporization. Biostimulation has been claimed as a result of minute heating of the tissue whereby the light-induced heat and/or the electromagnetic field stimulates nerves and accelerates wound healing. Higher-energy absorbed laser light can facilitate the joining of tissues—in particular blood vessels (i.e., anastomosis)—or can be used to coagulate blood to stop bleeding during a surgical procedure. If the light-induced heating causes the temperature to rise above 45°C, tissue necrosis and destruction occur as would be desired for the treatment of cancer or enlarged prostate tissues. At even higher-power densities, ablation or vaporization of the tissue occurs, as is the case for the corrective eye surgery known as radial keratectomy.

The light-induced heating is typically performed with laser light. The lasers used have different wavelengths (from the ultraviolet to the infrared spectrum), power densities (i.e., ratio between beam power and irradiated area), and duration times. The amount of energy imparted to the tissue and therefore temperature rise can be changed by either varying the power density or the duration of the time pulse of the laser. For high-power densities, coagulation, necrosis, and vaporization of tissue can occur, while at low-power densities, minimal heating is observed. For some wavelengths—for instance, those of the known strong absorption bands of water—the laser beam is highly absorbed, since tissue is primarily made of water. At these wavelengths the energy is then highly absorbed in a relatively thin layer near the surface where rapid heating occurs (i.e., radial keratectomy). The absorption is less for other wavelengths used away from the water absorption bands, and this results in slower heating of a larger volume of the tissue (i.e., for prostate coagulation).

Temperature Generation and Rate of Photon Absorption

A thermodynamically irreversible mode of interaction of light with materials is the process of absorption in which the photon energy is absorbed by the material phase. In the absence of conduction, the temperature rise in tissue is governed by a thermodynamic equation of state. The equation of state requires that the change in internal energy of the system be proportional to temperature rise. The change in internal energy of the system, in the absence of conduction and other heat transfer processes, is equal to the rate of energy deposition by the laser. Expressing this relation in terms of time derivatives gives

$$\Delta U/\Delta t = \rho C \Delta T/\Delta t \approx Q_L \tag{17.48}$$

where ΔT (K) is temperature rise, Q_L (W/m^3) is volumetric rate of photon absorption by the tissue, ρ (kg/m^3) is mass density, C (J/Kg.K) is specific heat, ρC (J/m^3.K) is volumetric heat capacity, and t is the duration. A very important factor in temperature rise by photons is the rate of photon absorption Q_L, which is also known as the light/laser source term and as the rate of energy deposition. For ordinary interaction processes, the rate of absorption of photons by the material is proportional to irradiance, and the constant of proportionality is the absorption coefficient:

$$Q_L = \mu_a \phi \tag{17.49}$$

Irradiance ϕ is related to the spatial distribution of photons within the tissue.[1] For purely absorbing materials, for instance, the Beer-Lambert law applies, and if a Gaussian beam profile is additionally assumed, then

$$Q_L = \mu_a{}^*\phi_o \ \exp(-\mu_a z) \ \exp(-2r^2/\omega^2) \tag{17.50}$$

where ϕ_o is incident intensity, z is depth, and r is radial distance. In the presence of scattering, one of the scattering models discussed before can be used to describe ϕ_o.

Equation (17.48) assumes no other thermal interaction processes such as conduction, convection, or perfusion. If these processes can be ignored, as, for example, for very short laser pulses, then temperature rise can be estimated as

$$\Delta T \ \approx \ Q_L \Delta t/\rho C \tag{17.51}$$

where Δt is exposure duration. Other thermal effects of laser-tissue interaction will be considered in more detail in Section 17.5. Equation (17.51) is valid for very short irradiations during which heat diffusion is negligible. A criteria or an estimate of the upper limit for validity of Eq. (17.51) is

$$\Delta t_{MAX} = \frac{1}{\sqrt{4\mu_a^2 \alpha}} \tag{17.52}$$

where α is the thermal diffusivity of the material. For water, the value of α is about 1,400 cm^2/s.

17.3.2 Laser Doppler Velocimetry

In addition to temperature, another physical interaction of light is the Doppler phenomenon, which is based on a frequency shift due to the velocity of a moving object. For medical applications, these objects are typically the moving red blood cells with a diameter of roughly 7 μm. The Doppler approach is used for measuring blood flow velocity for a variety of medical applications, including heart monitoring, transluminar coronary angioplasty, coronary arterial stenosis, tissue blood flow on the surface of the body, and monitoring blood flow on the scalp of a fetus during labor.

When a light wave of frequency f and velocity c impinges on a stationary object, it is reflected at the same frequency. However, if the object moves with velocity v, the reflected frequency f' is different from f. This difference or shift in frequency, δf, from the original light wave is known as the Doppler effect or Doppler shift and can be described as

$$\delta f = f - f' \tag{17.53}$$

The Doppler shift is described in Eq. (17.54) in terms of the velocity, v, of the moving red blood cells, the refractive index of the medium, n, the speed of light in the tissue, c_o, the input frequency, f, and the angle between the incident beam and the vessel, θ:

$$\delta f/f = 2 \ v \ n \ \cos\theta/c_o \tag{17.54}$$

[1]For multiphoton light-tissue interaction processes, the exponent of ϕ is increased. For example, for a two-photon process $Q_L = \mu_a * \phi^2$.

Further, it is well known that the wavelength, λ, is equal to the speed of light divided by the frequency, so

$$\delta f = 2\, v\, n\, \cos\theta/\lambda_o \tag{17.55}$$

A comparable analogy for the Doppler effect with sound waves is a train moving toward an observer. As the train gets closer, the whistle sounds like a higher pitch.

EXAMPLE PROBLEM 17.7

Using a laser Doppler system, calculate the velocity of blood given an argon laser (514 nm) at an angle with the vessel of 30 degrees, an index of refraction of 1.33, and a frequency shift of 60 KHz. Does this make sense physiologically? Does it really matter if you were off in your probe to vessel measurement angle by 10 degrees?

Solution

Using Eq. (17.55), the velocity can be calculated to be

$$v = (\delta f^{*}\lambda)/(2n\, \cos(\theta)) = [(60 \times 10^{3})(514 \times 10^{-9})]/[2\, (\cos(30))\, (1.33)] = 1.34\ \text{cm/s}$$

This is reasonable for an average-sized vessel in the human body. If you were off by 10 degrees—for example, if you thought the angle was 40 instead of 30—this would make a pretty big difference with the newly calculated speed of 1.51 cm/s (a 13 percent error).

Laser Doppler is a good method for monitoring velocities, but one reason it has not gained wide use clinically is that the flow (cm^3/sec) of blood or rather the average velocities of all the particles in the fluid is the physical quantity of diagnostic value, and flow cannot be directly measured using the Doppler approach. Calculating the flow rate for a rigid tube filled with water by knowing the velocity is a straightforward problem to solve. However, determining the absolute flow rate of blood in the body by simply using the Doppler measured velocity is a much more difficult problem, particularly for narrow vessels. Blood is thicker than water, and the flow characteristics are more complex. Blood vessels are also not rigid, straight tubes, and the flow of blood is pulsatile. Finally, a fiber optic probe can be inserted in the blood vessel and scanned to get a series of velocity measurements to determine flow, but the probe itself can alter the flow measurement.

Overall, laser Doppler velocimetry is a simple concept that can, with some effort, be used to measure relative changes in flow, but it is very difficult for this approach to be calibrated for absolute flow measurements. Thus, the standard for blood mass-flow measurements has been by thermodilation and is widely used in clinical practice. The thermodilation method is used to determine flow by inducing a predetermined change in the heat content of the blood at one point of the circulation and detecting the resultant change in temperature at some point downstream, after the flow has caused a controlled degree of mixing across the vessel diameter.

17.3.3 Interferometry

The phenomenon of interference is another method by which physical light interaction takes place, and it depends on the superposition of two or more individual waves, typically originating from the same source. Since light consists of oscillatory E and M fields that are

vectors, they add vectorally. Thus, when two or more waves emanating from the same source are split and travel along different paths, they can then reunite and interfere constructively or destructively. When the constructive and destructive interferences are seen to alternate in a spatial display, the interference is said to produce a series or pattern of fringes. If one of the paths in which the light travels is altered by any small perturbation, such as that due to temperature, pressure, or index of refraction changes, then once recombined with the unaltered reference beam, the perturbation causes a shift in the fringe pattern that can be readily observed with optoelectronic techniques to about 10^{-4} of the fringe spacing. The useful information regarding the changing variable of interest can be measured quite accurately as a path length change on the order of one-hundredth of a wavelength or 5×10^{-9} meters for visible light.

There are several variations for producing light interference, including one of the first instruments known as the Rayleigh refractometer from which came the Mach-Zehnder interferometer. For the sake of brevity, this section focuses on two more sophisticated variations of the Rayleigh idea: the dual beam Michelson interferometer and the multiple beam Fabry-Perot interferometer. As shown in Figure 17.9, the Michelson interferometer begins with a light source that is split into two beams by means of a beam-splitting mirror or fiber optics, which also serves to recombine the light after reflection from fully silvered mirrors. A compensating plate is sometimes needed to provide equal optical paths before introduction of any perturbation or sample to be measured. The perturbation can take the form of a pressure or temperature change, causing a strain and thus a path length change in the sample arm of the fiber optic or as a change due to the addition of a tissue sample or replacement of the mirror with a tissue sample. For example, the Michelson interferometer has been investigated for measuring tissue thickness, particularly for corneal tissue as feedback for the radial keratectomy procedure (laser removal or shaving of the cornea to correct vision), and this interferometer, when used with a low coherent light source, has been researched for use in optical coherence tomographic imaging of superficial tissues. The governing equation for the irradiance of the fringe system of circles concentric with the optic axis in which the two interfering beams are of equal amplitude is given as

$$I = 4I_o \cos^2(\delta/2) \tag{17.56}$$

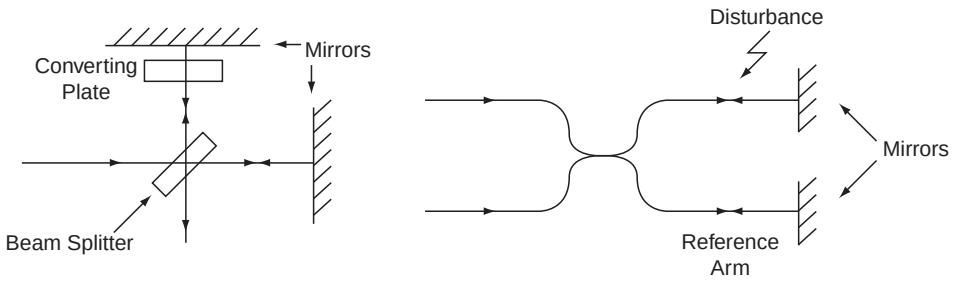


FIGURE 17.9 Michelson interferometer: (a) bulk optics and (b) fiber optics.

where I_o is the input intensity and the phase difference δ is defined as

$$\delta = 2\pi\Delta/\lambda \tag{17.57}$$

In this equation, λ is the wavelength, and the net optical path difference, Δ, is defined as

$$\Delta = 2d\cos(\theta) + \lambda/2 = (m + 0.5)\lambda \tag{17.58}$$

Clearly, from Eq. (17.58) the following equation can be derived

$$2d\cos(\theta) = m\lambda \tag{17.59}$$

in which $2d$ is the optical path difference or, rather, the difference in the two paths from the beam splitter, m is the number of fringes, and θ is the angle of incidence ($\theta = 0$ is a normal or on-axis beam). When a plate, a gas, or a thin minimally absorbing or scattering tissue slice is assumed with constant index of refraction and is inserted in one of the paths, then $d = (n_s - n_{air})L$, in which L is the actual length of the substance, n_s is the index of refraction of the substance, and n_{air} is the index of refraction of the air.

EXAMPLE PROBLEM 17.8

A thin sheet of clear tissue, such as a section of the cornea of the eye ($n \approx 1.33$), is inserted normally into one beam of a Michelson interferometer. Using 589 nanometers of light, the fringe pattern is found to shift by 50 fringes. Determine the thickness of the tissue section.

Solution

From Eq. (17.59), $2d\cos(\theta) = m\lambda$, and thus $d = (50\times.589)/2 = 14.72$ micrometers, which is the calculated optical path length. However, the physical length of the tissue must take into account the index of refraction of the sample and the air. Thus, as described in the paragraph after Eq. (17.59), the equation for physical path length is $L = d/(n_s - n_{air}) = 14.72/(1.33 - 1.0) = 44.6$ micrometers.

All the dual beam interferometry techniques such as the Michelson and Mach-Zehnder approaches suffer from the limitation that the accuracy depends on the location of the maxima (or minima) of a sinusoidal variation, as shown in Eq. (17.59). For very accurate measurements, such as precision spectroscopy, this limitation is severe. Rather than a dual beam, if the interference of many beams is utilized, the accuracy can be improved considerably. A Fabry-Perot interferometer uses a multiple beam approach, as shown in Figure 17.10. As can be seen in the figure, the interferometer uses a plane parallel plate to produce an interference pattern by combining the multiple beams of the transmitted light. The parallel plate is typically composed of two thick glass or quartz plates that enclose a plane parallel plate of air between them. The flatness and reflectivity of the inner surfaces are important and are polished generally better than $\lambda/50$ and coated with a highly reflective layer of silver or aluminum. The silver is good for wavelengths above 400 nm in the visible light region, but aluminum has better reflectivity below 400 nm. These film coatings must also be thin enough to be partially transmitting ($\sim$50 nm thickness for silver coatings). In many instances the outer surfaces of the glass plate are purposely formed at a small angle relative to the inner faces (several minutes of arc) to eliminate spurious fringe

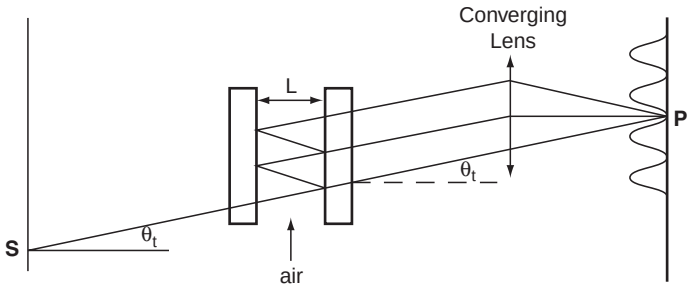


FIGURE 17.10 Fabry-Perot multiple beam interferometer.

patterns that can arise from the glass itself acting as the parallel plate interferometer. When L is fixed, the instrument is referred to as a Fabry-Perot etalon. The nature of the superposition at point P is determined by the path difference between successive parallel beams, so

$$\Delta = 2n_f\, L\, \cos\theta_t = m\lambda_{(\text{minimum})} = 2L\, \cos\theta_t \tag{17.60}$$

using $n_f = 1$ for air.

Other beams from different points of the source but in the same plane and making the same angle θ_t with the axis satisfy the same path difference and also arrive at P. With L fixed, Eq. (17.60) for Δ is satisfied for certain angles θ_t and the fringe system is the familiar concentric rings due to the focusing of fringes of equal inclination, as shown in Figures 17.11a and b. If the thickness of L varies with time, a detector D will record an interferogram as a function of time, as shown in Figure 17.11c.

Variation in the fringe pattern irradiance of the Fabry-Perot as a function of phase or path difference is known as the fringe profile. The sharpness of the fringes is important to the ultimate resolving power of the instrument. Using the trig identity $\cos\delta = 1 - 2\sin^2(\delta/2)$, the transmittance can be written as

$$T = I_T/I_i = 1/(1 + [4r^2/(1-r^2)^2]\, \sin^2(\delta/2)) \quad \text{(Airy function)} \tag{17.61}$$

where r is the reflectivity, and the term in the square brackets in the denominator is known as the coefficient of finesse, F, and thus,

$$T = 1/(1 + F\, \sin^2(\delta/2)) \tag{17.62}$$

It should be noted that as $0 < r < 1$, then $0 < F < \infty$. The coefficient of finesse, F, also represents a certain measure of fringe contrast:

$$F = ((I_T)_{\max} - (I_T)_{\min})/I_{T\,\min} = (T_{\max} - T_{\min})/T_{\min} \tag{17.63}$$

It should be noted here that $T_{\max} = 1$ when $\sin(\delta/2) = 0$, and $T_{\min} = 1/(1 + F)$ when $\sin(\delta/2) = \pm 1$. Given r, the fringe profile can be plotted as shown in Figure 17.12. As seen in the figure, $T = T_{\max} = 1$ at $\delta = m2\pi$ and $T = T_{\min} = 1/(1 + F)$ at $\delta = (m + ½)2\pi$. Further, $T_{\max} = 1$ regardless of r, and $T_{\min}$ is never zero but approaches zero as r approaches one. Also, the transmittance peaks sharply at higher values of r, remaining near zero for most of the region between the fringes. It should be noted that the Michelson interferometer

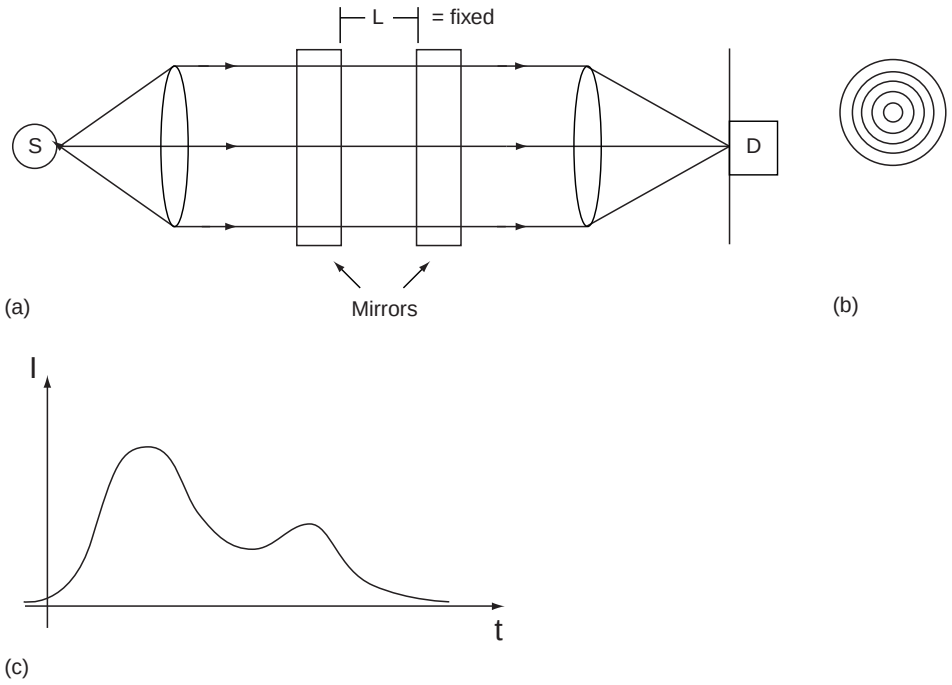


FIGURE 17.11 F-Petalon with L fixed.

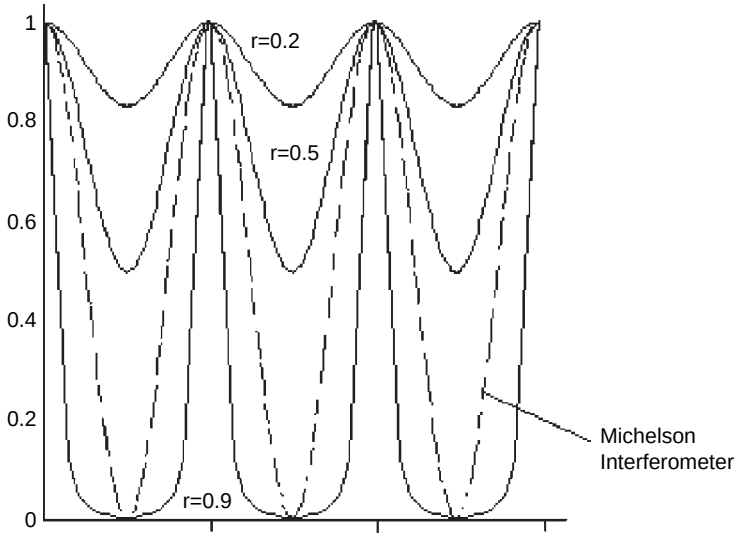


FIGURE 17.12 The difference in resolution between the multiple beam Fabry-Perot interferometer with different reflectivities and the dual beam Michelson interferometer.

produces broad fringes when it is normalized to the same max value, since it has a simple $\cos^2(\delta/2)$ dependence on the phase (dashed lines in the figure). An example of a fiber optic–based Fabry-Perot interferometer is the commercially available intracranial fluid pressure monitoring system for patients with severe head trauma or hydrocephalus (increased amount of cerebral spinal fluid in the ventricles and/or subarachnoid spaces of the brain).

17.4 BIOCHEMICAL MEASUREMENT TECHNIQUES USING LIGHT

In the last few years, there has been much enthusiasm and considerable effort from medical device companies and universities to perform diagnostic procedures such as cancer detection and quantifiably monitor blood chemicals, such as glucose, lactic acid, albumin, and cholesterol using various optical approaches. The most well-known optical monitoring approach used clinically in vivo is the pulse oximeter, which indirectly measures oxygen saturation and changes in blood volume by detecting changes in the optical absorption peaks of oxygenated and deoxygenated hemoglobin as it is pulsed through the arteries. It is often used to also measure heart rate.

The most common optical approaches for diagnostic and sensing applications include absorption, scattering, luminescence, and polarimetry. The primary variable for each of these approaches is a change in the light intensity as it passes through the medium, which can change as a function of the wavelength or polarization of the light. The lifetime and/or phase can also be used as a variable. For absorption and luminescence, the intensity will change nearly linearly for moderate analyte concentrations and nonlinearly for high concentrations.

17.4.1 Spectroscopic Measurements Using Light Absorption

Many investigators have suggested infrared absorption as a potential route to blood chemical monitoring and diagnostic sensing, in particular for glucose monitoring and cancer detection. The governing Eq. (17.17) for purely absorbing media as well as fluorescence was described previously using the Beer-Lambert law. Expressed logarithmically, the equation becomes

$$A = \ln(T) = \ln(I_o/I) = \mu\, z = z\, \Sigma\, \varepsilon_i\, C_i \tag{17.64}$$

in which A is absorbance, T is transmittance, I_o is the incident light, I is the transmitted light, z is the path length, and μ is the absorption coefficient, or rather the sum of the multiplication of the molar absorptivity times the concentration of all the different components in the analyte. For tissue and blood this equation is valid in the midinfrared wavelength region (wavelengths of 2.5–12 micrometers), in which the absorption peaks due to various chemicals are distinguishably sharp and the scatter is weak. Unfortunately, the absorption of the light due to water within tissue in this region is orders of magnitude stronger than any of the blood chemicals, which results in the possibility of only short path length sample investigations (on the order of micrometers). This brings about the possibility for surface or superficial investigations on the skin or, with the use of fiber optics, on internal body

cavities. Unfortunately, the optical fibers available in this region tend to be toxic, hydroscopic, and/or rigid. Consequently, the combination of the water absorption and lack of good fibers makes in vivo diagnostics and sensing very difficult in the mid-IR region. The light sources used in this region include the broad-based tungsten bulb, nernst glower, nichrome wire, and globar rod, and narrowly tunable, typically liquid nitrogen cooled, laser diodes. The detectors used include the cooled mercury cadmium telluride (MCT), thermopile, and thermistor. The optics include sodium chloride and potassium bromide, mirrors (typically gold coated), and gratings.

In the near-infrared (NIR) wavelength region, the spectrum is not affected by water to the same degree as the midinfrared region, allowing for path lengths within the tissue of 1 mm to 1 cm to be used. In addition, low OH silica fibers are quite transparent across this range and are nontoxic and nonhydroscopic. The NIR region (700–2,500 nm) exhibits absorptions due to low-energy electronic vibrations (700–1,000 nm), as well as overtones of molecular bond stretching and combination bands (1,000–2,500 nm). These bands result from interactions between different bonds (−CH, −OH, −NH) to the same atom. Typically, only the first, second, and third overtones of a molecular vibration are detectable and are broad in nature. Thus, only at high concentrations of the chemical species are these overtones qualitatively detectable with the intensity dropping off rapidly as overtone order increases. The NIR absorption bands are also influenced by temperature, pressure, and hydrogen bonding effects and can overlap significantly. In the 700 to 1,200 nm region, they are also influenced by scattering effects. Thus, unlike mid-IR spectroscopy, NIR spectroscopy is primarily empirical and not well suited for qualitative work. However, using techniques such as multivariate statistics, quantitative analysis is possible with the NIR spectrum. The light sources used in this region include the broad-based tungsten bulb, light emitting diodes (LEDs), laser pumped solid-state lasers such as the tunable titanium sapphire laser, and laser diodes. The detectors used include silicon (good to 1 micrometer), germanium (good to 1.7 micrometers), and indium antiminide (good to 5.5 micrometers). The optics include low OH glasses, quartz, glass, gratings, and mirrors.

The means for separating the wavelength of the light of a broadband source include dispersive and nondispersive methods. The dispersive approach uses either a ruled reflective grating or transmissive prism to separate the wavelengths of light. The nondispersive systems include a series of wavelength selection filters or a Fourier transform infrared (FTIR)-based instrument. The FTIR method uses an interferometer similar to the Michelson interferometer shown in Figure 17.9 to collect the entire spectrum and then deconvolutes it, using Fourier transform techniques. Both approaches can be configured to cover the NIR and mid-IR ranges. The advantages of the dispersive systems include higher resolution and separation of closely spaced wavelength bands, while the nondispersive systems generally have better throughput, since all the light passes through the sample.

17.4.2 Monitoring Approaches Using Scattered Light

There are fundamentally two types of optical scattering for diagnostics and monitoring: elastic and inelastic. The elastic scattering can be described using Mie theory (or Rayleigh scattering for particles that are small compared to the wavelength), in which the intensity of the scattered radiation can be related to the concentration, size, and shape of the

scattering particles. The inelastic scattering, in which the polarization of the particle is not constant, can be described as Raman scattering. Note that fluorescence, although an absorption reemission process, is also inelastic.

When using the light scattering phenomenon for sensing, the intensity of the reflected light is usually considered. The reflection of light, however, can be divided into two forms. Specular reflection, or a "mirror" type of reflection, occurs at the interface of a medium. The returned light yields little information about the material other than its roughness, since it never penetrates the medium. Thus, for applications other than surface roughness, the specularly reflected light is typically minimized or eliminated with the design of the optical sensor. Diffuse reflection, however, occurs when light penetrates into a medium, becomes absorbed and multiply scattered, and makes its way back to the surface of the medium. The model describing the role of diffuse scattering in tissue is based on the radiative transfer theory, as was described in Section 17.2. The same theory applies for sensing as well.

The use of elastically scattered light has been suggested for both diagnostic procedures such as cancer detection and for monitoring analytes such as glucose noninvasively for diabetics. For use as a monitoring application of chemical changes such as glucose, researchers have used an intensity-modulated frequency domain NIR spectrometer, capable of separating the reduced scattering and absorption coefficients to detect changes in the reduced scattering coefficient showing correlation with blood glucose in human muscle. This approach was promising at first as a relative measure over time, since clearly an increase in glucose concentration in the physiologic range decreases the total amount of tissue scattering. However, the drawbacks of the light elastic scattering approach for analyte monitoring are still quite daunting. The specificity of the elastic scattering approach is the biggest concern with this method, since other physiologic effects unrelated to glucose concentration could produce similar variations of the reduced scattering coefficient with time, and unlike the absorption approach, elastic light scattering as a function of the molecule is nearly wavelength independent.

The measurement precision of the reduced scattering coefficient and separation of scattering and absorption changes is another concern with this approach. It is difficult to measure such small changes and be insensitive to some of the larger absorption changes in the tissue, particularly hemoglobin. This approach also needs to take into account the different refractive indices of tissue. Tissue scattering is caused by a variety of substances and organelles (membranes, mitochondria, nuclei, etc.), and all of them have different refractive indices. The effect of blood glucose concentration and its distribution at the cellular level is a complex issue that needs to be investigated before this approach can be considered viable. An instrument of this type would require in vivo calibration against a gold standard, since the reduced scattering coefficient is dependent on additional factors such as cell density. Last, there is a need to account for factors that might change the reduced scattering coefficient, such as variations in temperature, red blood cell concentration, electrolyte levels, and movements of extracellular and intracellular water.

As a diagnostic screening tool for cancer detection, measurement of the scatter in thin tissues or cells may hold promise. Many of the changes in tissue due to cancer are morphologic rather than chemical and thus occur with changes in the size and shape of the cellular and subcellular components (membranes, mitochondria, nuclei, etc.). Thus, the changes in elastic light scatter should be larger with the morphologic tissue differences.

If the wavelength of the elastic light scattering is carefully selected so as to be outside the major absorption areas due to water and hemoglobin and if the diffusely scattered light is measured as a function of angle of incidence, there is potential for this approach to aid in pathologic diagnosis of disease.

Inelastic Raman spectroscopic scattering has been utilized over the past few decades initially by physicists and chemists. Raman spectroscopy has become a powerful tool for studying a variety of biological molecules, including proteins, enzymes and immunoglobulins, nucleic acids, nucleoproteins, lipids and biological membranes, and carbohydrates, but with the advent of more powerful laser sources and more sensitive detectors, it has also become useful as a diagnostic and sensing tool. The phenomenon of Raman scattering is observed when monochromatic (single wavelength) radiation is incident upon the media. In addition to the elastic scatter of the transmitted light, a portion of the radiation is inelastically scattered. Thus, some of the incident light of frequency w_o exhibits frequency shifts $\pm w_m$, which is associated with transitions between rotational, vibrational, and electronic levels that are specific to a particular analyte of interest. Most studies utilize the Stokes type of scattering bands that correspond to the $w_o - w_m$ scattering. Therefore, the Raman bands typically used are those shifted by the interaction with the analyte to longer wavelengths relative to the excitation wavelength.

As with infrared spectroscopic techniques, Raman spectra can be utilized to identify molecules, since these spectra are characteristic of variations in the molecular polarizability and dipole moments. Raman spectroscopy can be considered as complementary to absorption spectroscopy because neither technique alone can resolve all of the energy states of a molecule. In fact, for certain molecules, some energy levels may not be resolved by either technique. Due to the anharmonic oscillator model for dipoles, overtone frequencies exist in addition to fundamental vibrations. It is an advantage of Raman spectroscopy that the overtones are much weaker than the fundamental tones, thus contributing to simpler spectra as compared to absorption spectroscopy.

One advantage to using Raman spectroscopy in biological investigations is that the Raman spectrum of water is weaker and therefore, unlike infrared spectroscopy, only minimally interferes with the spectrum of the solute. Thus, the spectrum can be obtained from aqueous solutions at reasonable path lengths. However, the Raman signal is also weak, and only recently, with the replacement of slow photomultiplier tubes with faster CCD arrays as well as the manufacture of higher power near infrared laser diodes, has the technology become available to allow researchers to consider the possibility of distinguishing normal and abnormal tissue types as well as quantifying blood chemicals in near real time. In addition, investigators have applied statistical methods such as partial least squares (PLS) to aid in the estimation of biochemical concentrations from Raman spectra.

As with elastic scatter, Raman spectroscopy has been used for both diagnostics and monitoring. The diagnostic approaches look for the presence of different spectral peaks and/or intensity differences in the peaks due to different chemicals present in, for instance, cancerous tissue. For quantifiable monitoring it is the intensity differences alone that are investigated. In tissue, one problem is the high fluorescence background signal as a result of autofluorescence incurred in heavily vascularized tissue due to the high concentration of proteins and other fluorescent components. Instrumentation to excite in the NIR

wavelength range has been proposed to overcome the autofluorescence problem, since the fluorescence component falls off with increasing wavelength.

Excitation in the NIR region offers the added advantage of longer wavelengths that pass through larger tissue samples with lower absorption and scatter than other spectral regions, such as visible or ultraviolet. However, in addition to fluorescence falling off with wavelength, the Raman signal also falls off to the fourth power as wavelength increases. Thus, there is a tradeoff between minimizing fluorescence and maintaining the Raman signal. The eye has been suggested as a site for analyte concentration measurements using Raman spectroscopy to minimize autofluorescence, but the disadvantage to using the eye for Raman spectroscopy is that the laser excitation powers must be kept low to prevent injury and this significantly reduces the signal-to-noise ratio. Finally, like infrared and NIR absorption, to quantifiably determine the inherently low concentrations of analytes in vivo, the presence of different chemicals must be accounted for that yield overlapping Raman spectra.

17.4.3 Use of the Luminescence Property of Light for Measurement

As described previously, luminescence is the absorption of photons of electromagnetic radiation (light) at one wavelength and reemission of photons at another wavelength. The photons are absorbed by the molecules in the tissue or medium, raising them to some excited energy state, and then, upon returning to a lower energy state, the molecules emit radiation or, rather, luminesce at a different wavelength. The luminescent effect can be referred to as fluorescence or phosphorescence. Fluorescence is luminescence that has energy transitions that do not involve a change in electron spin, and therefore the reemission occurs much faster. Consequently, fluorescence occurs only during excitation, while phosphorescence can continue after excitation (i.e., after the light source is turned off). As an example, a standard television while turned on is producing fluorescence, but for a very short time after it is turned off the screen will phosphoresce.

The measurement of fluorescence has been used for diagnostic, monitoring, and research purposes. Fluorescent microscopes are now being used for research that produce outstanding images of cells and tissues that provide a variety of information about them. Obtaining diagnostic information, in particular with respect to cancer diagnosis or the total plaque in arteries, has also been attempted using both intrinsic fluorescence of tissue and extrinsic dyes. The intrinsic fluorescence is due to the naturally occurring proteins, nucleic acids, and nucleotide coenzymes within the tissue, while extrinsic fluorescence is induced by the uptake of certain dyes in the tissue. Extrinsic fluorescence has also been investigated, for instance, to monitor such analytes as glucose, intracellular calcium, proteins, and nucleotide coenzymes. Unlike the use of fluorescence in chemistry on dilute solutions, the intrinsic or autofluorescence of tissue, as well as the scattering and absorption of the tissue, acts as a noise source for the extrinsic approach.

The response of a fluorescence sensor can be described in terms of the output intensity as

$$I_f = \Phi_f \, (I_o - I) \tag{17.65}$$

in which I_f is the radiant intensity of fluorescence, Φ_f is the fluorescence efficiency, I_o is the radiant intensity incident on the sample, and I is the radiant intensity emerging from the

sample. The fluorescence efficiency can be described as the combination of three factors. First is the quantum yield, which is the probability that an excited molecule will decay by emitting a photon rather than by losing its energy nonradiatively. This parameter varies from 1.00 to 0.05 and varies with time on the order of nanoseconds. Thus, in addition to intensity measurements, time-resolved fluorescence measurements are possible using pulsed light sources and fast detectors. The second parameter is the geometrical factor, which is the solid angle of fluorescence radiation subtended by the detector and that depends on your probe design. Last is the efficiency of the detector itself for the emitted fluorescence wavelength.

Since fluorescence is an absorption/reemission technique, it can be described in terms of Beer's law as

$$I_f = \Phi_f \, I_o \, [1 - \exp(-\varepsilon C l)] \tag{17.66}$$

in which C is the concentration of the analyte, l is the path length, and ε is the molar absorptivity. Equation (17.66) can be described in terms of a power series, and for weakly absorbing species ($\varepsilon C l < 0.05$) only the first term in the series is significant. Therefore, under these conditions, the response of the fluorescence sensor becomes linear with analyte concentration and can be described as

$$I_f = \Phi_f \, I_o \, \varepsilon C l \tag{17.67}$$

The primary fluorescent sensors are based on the measurement of intensity, but lifetime measurements in the time or frequency domain are also possible. To gain the most information, particularly in a research or teaching setting, dual monochromators (grating-based wavelength separation devices) are used with either a photomultiplier tube as the detector or a CCD array detector. In a typical benchtop fluorimeter, a broad, primarily ultraviolet/visible xenon bulb is used as the light source. The light is coupled first through a monochromator, which is a wavelength separator that can be set for any excitation wavelength within the range of the source. The light then passes through the sample and is collected by a second monochromator. The light reflected from the grating within the second monochromator can be scanned so a photomultiplier tube (PMT) receives the different wavelengths of light as a function of time. Alternatively, all the wavelengths from the grating can be collected simultaneously on a CCD detector array.

The advantage of the CCD array is that it provides for real-time collection of the fluorescence spectrum. The advantage of the PMT is that it is typically a more sensitive detector. In many systems a small portion of the beam is split at the input and sent to a reference detector to allow for correction of fluctuations in the light source. Once the optimal configuration for a particular biomedical application, such as cervical cancer detection or glucose sensing, has been investigated using the benchtop machine, an intensity measurement system can be designed with a simpler, more robust configuration. Such a system can be designed with lasers and/or wavelength-specific filters instead of monochromators and be made to work at two or more discrete wavelengths. In addition, optical fibers can be used for delivery and collection of the light to the remote area. Since the excitation wavelength and fluorescent emission wavelengths are different, the same fiber or fibers can be used to both deliver and collect the light. In any configuration it is important to match the spectral characteristics of

the source, dye if used, sample, and detector. For instance, depending on the tissue probed, there are strong absorbers, scatterers, and autofluorescence that need to be factored into any fluorescent system design.

17.4.4 Measurements Made Using Light Polarization Properties

Having discussed some of the fundamental electromagnetic theory of polarized light generation in Section 17.1.2, some of the applications of polarized light shall now be discussed. Applications of polarized light include biochemical quantification in vivo such as noninvasively monitoring glucose for diabetes, measuring birefringence in tissues using polarized light microscopy, and tissue characterization, in particular to aid in cancer identification or for use in the measurement of the nerve fiber layer of normal and glaucomatous eyes.

The rotation of linearly polarized light by optically active substances has been used for many years to quantify the amount of the substance in solution. A variety of both polarimeters, adapted to the examination of all optically active substances, and saccharimeters, designed solely for polarizing sugars, have been developed. The concept behind these devices is that the amount of rotation of polarized light by an optically active substance depends on the thickness of the layer traversed by the light, the wavelength of the light used for the measurement, the temperature, the pH of the solvent, and the concentration of the optically active material.

Historically, polarimetric measurements have been obtained under a set of standard conditions. The path length typically employed as a standard in polarimetry is 10 cm for liquids, the wavelength is usually that of the green mercury line (5,461 Angstroms), and the temperature is 20°C. If the layer thickness in decimeters is L, the concentration of solute in grams per 100 ml of solution is C, α is the observed rotation in degrees, and $[\alpha]$ is the specific rotation or rotation under standard conditions, which is unique for all chiral molecules, then

$$C = \frac{100\alpha}{L[\alpha]} \tag{17.68}$$

In the preceding equation, the specific rotation $[\alpha]$ of a molecule is dependent on temperature, T, wavelength, λ, and the pH of the solvent.

For polarimetry to be used as a noninvasive technique—for instance, in blood glucose monitoring—the signal must be able to pass from the source, through the body, and to a detector without total depolarization of the beam. Since the skin possesses high scattering coefficients, which causes depolarization of the light, maintaining polarization information in a beam passing through a thick piece of tissue (i.e., 1 cm), which includes skin, would not be feasible. Tissue thicknesses of less than 4 mm, which include skin, may potentially be used, but the polarimetric sensing device will encounter greater than 95 percent depolarization of the light due to scattering from the tissue.

As an alternative to transmitting light through the skin, the aqueous humor of the eye has been investigated as a site for detection of in vivo glucose concentrations, since this sensing site is a clear biological optical media. It is also known that glucose concentration of the aqueous humor of the eye correlates well with blood glucose levels, with a minor

time delay (on the order of minutes) in rabbit models. The approximate width of the average anterior chamber of a human eye is on the order of 1 cm. Therefore, an observed rotation of about 4.562 millidegrees per optical pass can be expected for a normal blood glucose level of 100mg/100ml, given a specific rotation of glucose at $\lambda = 633$ nm of $45.62^{\circ}\text{cm}^2\text{g}^{-1}$ and thickness of 1 centimeter. The eye as a sensing site, however, is not without its share of potential problems. For instance, potential problems with using the eye include corneal birefringence and eye motion artifact.

As shown in Eq. (17.68), the rotation is directly proportional to the path length, and thus it is critical that this length be determined or at least kept constant for each individual subject regardless of the sensing site. If the eye is used as the sensing site, the angle of incidence on the surface of the cornea must be kept relatively constant for each patient so not only the path length but alignment remains fixed each time a reading is taken. In most tissues, including the eye, the change in rotation due to other chiral molecules such as proteins needs to be accommodated in any final instrument. In addition, most other tissues also have a birefringence associated with them that would need to be accounted for in a final polarimetric glucose sensor.

It is the birefringence and retardation of the polarized light as well as polarized scattering of the tissue that are the signals rather than the noise when using polarized light for tissue characterization. For example, a scanning laser polarimeter has been used to measure changes in retardation of the polarized light impinging on the retinal nerve fiber layer. It has been shown that scanning laser polarimetry provides statistically significant higher retardation for normal eyes in certain regions over glaucoma eyes. Images generated from the scattering of various forms of polarized light have also been shown to be able to differentiate between cancer and normal fibroblast cells.

17.4.5 Micrometer and Nanometer Biosensing Applications

Developments in microtechnology and, in particular, nanotechnology are transforming the fields of biosensors, prosthesis and implants, and medical diagnostics. In terms of medical diagnostics, these devices are being used in combination with optical biosensing for external, lab-on-a-chip, high-throughput screening for analyzing blood and other samples. Inside the body many researchers and companies are developing optically based nanotechnology applications for cancer diagnosis and therapy.

One nanotechnology used in biomedical optics that has come to the forefront is that of quantum dots. Quantum dots are devices capable of confining electrons in three dimensions in a space small enough that their quantum (wave-like) behavior dominates over their classical (particle-like) behavior. At room temperature, confinement spaces of 20–30 nm or smaller are typically required. Once the electrons are confined, they repel one another, and no two electrons can have the same quantum state. Thus, the electrons in a quantum dot will form shells and orbitals highly reminiscent of the ones in an atom, and will exhibit many of the optical, electrical, thermal, and chemical properties of an atom. Quantum dots can be grown chemically as nanoparticles of semiconductor surrounded by an insulating layer nearly colloidal in nature. These particles can also be deposited onto a substrate, such as a semiconductor wafer patterned with metal electrodes, or they can be crystallized into bulk solids by a variety of methods. Either substance can be stimulated with electricity or

light (e.g., lasers) in order to change its properties. Fundamentally, when you excite a quantum dot with one wavelength, depending on the size of the dot, it will emit (fluoresce) at a narrow-band longer wavelength. The size-dependent behavior allows them to be used for multiplexing, since a single wavelength can be used to excite different-sized quantum dots to produce multiple wavelengths within a sample.

Although quantum dots are nanometer-scale crystals that were developed in the mid-1980s for optoelectronic applications, one application in the biomedical area has been to probe living cells in full color over extended periods of time. Such a technique could reveal the complex processes that take place in all living organisms in unprecedented detail, such as the development of embryos. Existing imaging techniques use natural molecules that fluoresce, as just discussed, such as organic dyes and proteins that are found in jellyfish and fireflies. However, each dye emits light over a wide range of wavelengths, which means that their spectra overlap, and this makes it difficult to use more than three dyes at a time in order to tag and image different biological molecules simultaneously. The fluorescence of dyes also tends to fade away quickly over time, whereas semiconductor nanocrystals—quantum dots—can get around these problems. In addition to being brighter and lasting longer than organic fluorophores, quantum dots have a broader excitation spectrum and thus, as mentioned, a mixture of quantum dots of different sizes can be excited by a light source with a single wavelength, allowing simultaneous detection and imaging in color.

Although the preceding example has focused on biomedical sensing using fluorescent quantum dots, these micro- and nanoparticles can also be made of various materials and used with all of the preceding light propagation methods. For instance, metal nanoparticles such as gold or silver can be used with Raman spectroscopy to produce an effect known as surface-enhanced Raman spectroscopy (SERS), which gives rise to signals a million times more sensitive than regular Raman signals. These same types of metal nano- or microparticles can be injected into cancerous tissue and used as absorbers that when hit with infrared light will absorb the energy, cook the cancerous tumor, and kill it. Since nanoparticle development is still in its infancy, where they go in the body, their toxicology, whether they can be functionalized to go to specific organs or cells like cancer, and what other biomedical applications are to come from the combination of these particles with light remain to be seen.

17.5 FUNDAMENTALS OF THE PHOTOTHERMAL THERAPEUTIC EFFECTS OF LIGHT SOURCES

The therapeutic application of light, including lasers, is mediated by conversion of photonic energy to absorbed energy within the material phase of the tissue. The primary mode of this energy conversion manifests itself as a nonuniform temperature rise that leads to a series of thermodynamic processes. These thermodynamic processes can then be exploited as a means to affect therapeutic actions such as photothermal coagulation and ablation of tissue. Another mode of interaction is the utilization of the absorbed energy in activation of endogeneous or exogenous photosensitizing agents in a photochemical process known as photodynamic therapy.

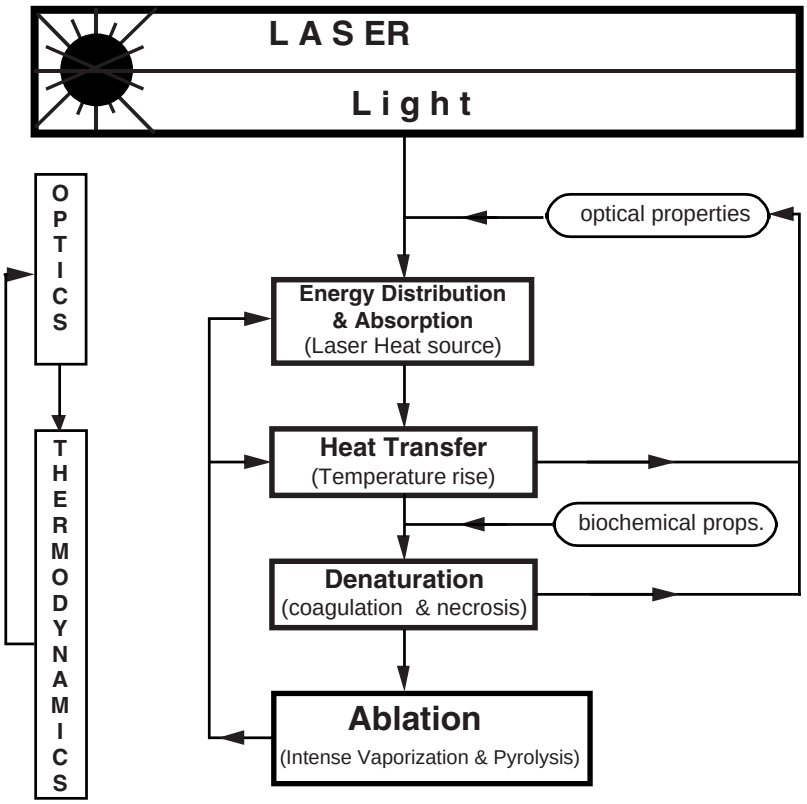


FIGURE 17.13 Flowchart of photothermochemical processes in thermal interaction of light with biological tissues.

Light including laser interaction with biological tissue is composed of a combination of optical and thermodynamical processes. An overview of these processes can be summarized as shown in Figure 17.13. Once light is irradiated on tissue, the photons penetrate into the tissue and, depending on the tissue optical properties such as scattering coefficient, absorption coefficient, and refractive index, the energy is distributed within the tissue. A portion of this energy is absorbed by the tissue and is converted into thermal energy, making the light act as a distributed heat source. This light-induced heat source in turn initiates a nonequilibrium process of heat transfer, manifesting itself by a temperature rise in tissue. The combined mechanisms of conduction, convection, and emissive radiation distribute the thermal energy in the tissue, resulting in a time- and space-dependent temperature distribution in the tissue. The temperature distribution depends on the thermal properties, conductivity, heat capacity, convective coefficients, and emissivity of the tissue.

As heat deposition and transfer continue, a certain threshold can be reached above which a process of irreversible thermal injury initiates. This process leads to coagulation of tissue caused by denaturation of enzymes and proteins and finally leads to necrosis of constituent cells. As a result of this thermochemical process of injury, properties of the tissue, especially

the optical properties, start to change. The change in properties in turn influences the process of energy absorption and distribution in the tissue.

The next stage in these processes is the onset of ablation. As the temperature continues to rise, a threshold temperature is reached, at which point, if the rate of heat deposition continues to exceed the rate at which the tissue can transport the energy, an intensive process of vaporization of the water content of the tissue combined with pyrolysis of tissue macromolecules initiates, which results in ablation or removal of tissue.

17.5.1 Temperature Field during Light/Laser Coagulation

In this section the governing equations are described for a thermodynamic analysis of light heating of biological tissues up to the onset of ablation. First, the heat conduction equation and typical boundary and initial conditions are described. Next the Arrhenius-Henriques model for quantitative analysis of irreversible thermal injury to biological tissue will be introduced.

Section 17.2 discussed how light energy is absorbed by a participating medium such as biological tissue. As light energy with a rate Q_L is absorbed by a material under irradiation, there is an immediate thermal energy flux traveling in different directions. This is caused by nonuniformity of Q_L in both the radial direction, due to beam profile, and the axial direction, due to absorption. The energy rate balance equation in the material can be found as follows.

Consider an infinitesimal element of the material under light/laser irradiation (Figure 17.14). The rate of thermal energy storage in the element, U, depends on the difference between incoming and outgoing thermal fluxes, the rate of light energy absorption, Q_L, and other energy rate interactions that are lumped together and labeled Q_o. The difference between influx and efflux is, in the limit, equal to the negative of the divergence of the thermal flux vector. So the energy rate balance can be written as

$$\partial U/\partial t = -\vec{\nabla} \cdot \vec{q} + Q_L + Q_o \tag{17.69}$$

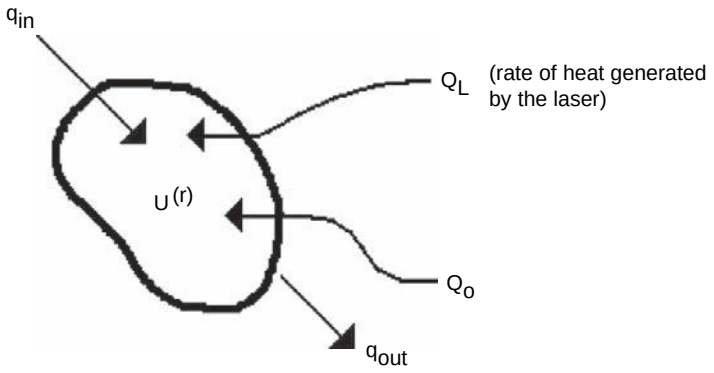


FIGURE 17.14 Thermal fluxes and energy in a control volume in the "material phase."

where $\vec{\nabla} \bullet \vec{q}$ is the change in thermal flux. The equation of state for thermal energy is

$$U = \rho C T \tag{17.70}$$

where T is an absolute temperature (K), ρ (kg/m^3) is the density, and C (joules/kg K) is the specific heat of the material.

At this point, as in the derivation of the photon phase, a phenomenological relation for the thermal flux is needed. The commonly used relation is known as Fourier's law, and it is written as

$$\vec{q} = -\kappa \vec{\nabla} T \tag{17.71}$$

where κ (W/m K) is the thermal conductivity of the material. Proceeding now by substituting Eqs. (17.71) and (17.70) into Eq. (17.69), the heat conduction equation is obtained:

$$\partial(\rho C T)/\partial t = \vec{\nabla} \cdot \kappa \vec{\nabla} T + Q_L + Q_o \tag{17.72}$$

This equation, solved with application of proper initial and boundary conditions, can, in theory, predict the temperature distribution in a material under laser irradiation.

In the case of biological tissue with other energy rates involved, Q_o can be written as

$$Q_o = Q_b + Q_m \tag{17.73}$$

where Q_b is the heat rate removed by blood perfusion, and Q_m is the rate of metabolic heat generation. These terms are not considered here. They are in general much smaller than the laser heat source term Q_L except when lower laser powers are used for longer times, as in hyperthermia treatments.

The material properties κ and ρC have been measured for many biological tissues, including human tissues. In general these properties may depend on temperature and water content, which could cause a nonlinear effect in Eq. (17.72). Another source of nonlinearity may be due to temperature dependence of the optical properties that, when present, has a more dominant role.

17.5.2 Thermal Coagulation and Necrosis: The Damage Model

The idea of quantifying the thermal denaturation process was first proposed in the 1940s. Using a single constituent kinetic rate model, this early work grossly incorporates the irreversible biochemical processes of coagulation, denaturation, and necrosis associated with thermal injury to biological tissue in terms of a single function. Defining an arbitrary nondimensional function, Ω, as an index for the severity of damage, the model assumes that the rate of change of this function follows an Arrhenius relation

$$\frac{d\Omega}{dt} = A \exp\left(-\frac{E}{RT}\right) \tag{17.74}$$

where R is the universal gas constant (cal/mole), T is the absolute temperature (K), and A (1/sec) and E (cal/mole) are constants to be determined experimentally. Total damage accumulated over a period t can be found by rearranging and integrating Eq. (17.74):

$$\Omega = A \int_0^t \exp\left(-\frac{E}{RT}\right) dt \tag{17.75}$$

The experimental constants A and E for pigskin were determined such that $\Omega = 1.0$ corresponded to complete transepidermal necrosis and $\Omega = 0.53$ indicated the minimum condition to obtain irreversible epidermal injury. The reported values are $A = 3.1 \times 10^{98}$ (1/sec), E = 150,000 (cal/mole). These values result in a threshold temperature of about 45°C. Different values for other tissues have been reported. For instance, for human arterial vessel walls, the values are $A = 4.1 \times 10^{44}$(1/sec), E = 74,000(cal/mole). For these values $\Omega = 1.0$ was defined as the threshold for histologically observed coagulation damage. It should be noted that the coagulation process—that is, collagen denaturation—is a different damage process than that seen in skin burns.

Once the heat conduction Eq. (17.72) is solved for the temperature field, the temperature history at every point in the tissue can be used in (17.75) to predict the accumulated damage at that point.

17.5.3 Photothermal Ablation

The temperature rise in a biological tissue under laser irradiation cannot continue indefinitely. A threshold temperature can be reached beyond which, if the rate of heat deposition by the light/laser continues to be higher than the rate heat can be transported by heat transfer mechanisms, intense vaporization of the water content of the tissue occurs. This results in creation of vacuoles—vacuolization—followed by pyrolysis at higher temperatures. The combined processes of intense vaporization—vacuolization—and pyrolysis of tissue macromolecule constituents is the "light/laser ablation process." This is the primary underlying mechanism for removal of tissue by laser surgery.

The large water content of most biological tissues suggests that the dominant part of the ablation process is the intense vaporization process. Apart from a short discussion of pyrolysis, intense vaporization is the main process of concern.

Following, a derivation is given for the ablation interface energy balance equation, which introduces the mathematical nonlinearity of the problem of ablation as a moving boundary problem. The section is concluded with a note on pyrolysis.

Ablation Interface Energy Balance Equation

The heat conduction equation alone cannot provide a mathematical model for determination of the dynamic behavior of the ablation front as a function of time. In fact, the motion of the ablation front influences the heat transfer process. An energy balance at the ablation front is needed in order to determine the motion of the front of ablation and its influence on the temperature field.

Consider a portion of a material under light irradiation with unit cross-sectional area and of thickness Δs, as shown in Figure 17.15, which is at the ablation threshold temperature

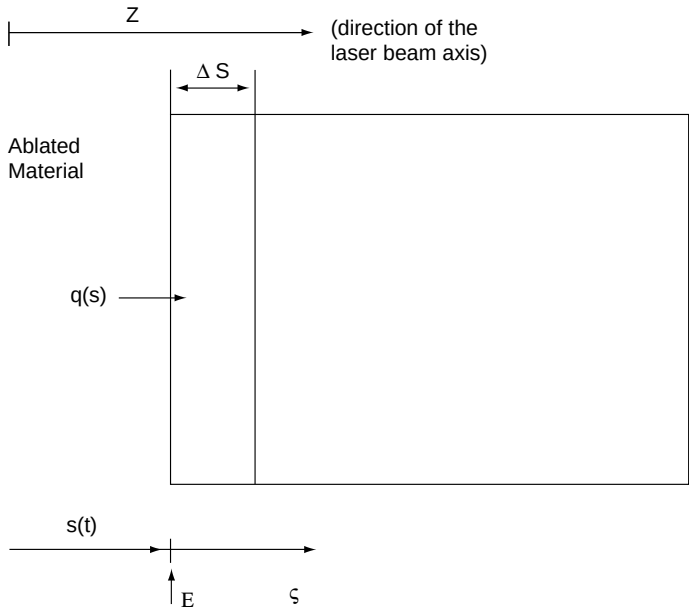


FIGURE 17.15 Ablation interface energy balance which is depicted as the amount of energy required to ablate a portion Δs of the tissue.

and is to be vaporized in the next interval of time Δt. Using Fourier's law for heat flux, Eq. (17.71), and after mathematical manipulations, the following equation for energy balance at the ablation interface can be derived

$$\kappa \frac{\partial T\,(s,\,t)}{\partial z} = \rho f_L L \frac{ds}{dt} \tag{17.76}$$

where f_L is the water fraction parameter. Nonlinearity inherent in the problem can be realized from this equation in that it couples the temperature gradient at the ablation interface to the rate of change of the front position, which is not known a priori. The solution requires an iteration procedure that requires the solution of the heat conduction equation simultaneously with the ablation front equation, with application of proper boundary and initial conditions.

Equation (17.76) is essential in providing information on the dynamics and thermodynamics of ablation. When solved simultaneously with the heat conduction equation and with the application of proper boundary and initial conditions, this equation provides the information on the position and velocity of the front of ablation. Furthermore, an important observation can be readily made about the temperature profile within the tissue, as discussed following.

Note that on the right-hand side of Eq. (17.76), ds/dt must be positive for the ablation process to proceed. This means that on the left-hand side, the temperature, $\partial T/\partial z$ must be positive. Consequently, the subsurface temperature must be higher than the surface temperature. That is, subsurface tissue must be superheated to temperatures higher than the

surface ablation temperature so thermal energy can be provided to the surface for the ablation process to proceed. This is in accordance with one of the corollaries of the second law of thermodynamics, stating that heat can be transported only from hotter points to colder points. In fact, the higher than ablation threshold temperature puts the subsurface tissue in a "metastable" equilibrium condition, which may be perturbed by internal tissue conditions and result in nucleation and vaporization initiating subsurface and manifesting itself by an "explosion" and mechanical tearing of the tissue surface. It should be noted at this point that initiation of the ablation process involves nonequilibrium nucleation processes that are not considered by the present approach.

17.5.4 Analytical Solution of Light Heating in a Purely Absorbing Medium

The governing equations for light irradiation with ablation were described in Section 17.5.3. Here, the governing equations will be repeated for a one-dimensional, semi-infinite, and purely absorbing medium. The governing equation for the preablation heating stage can be written as

$$\frac{\partial \rho c T}{\partial t} = \frac{\partial^2 T}{\partial z^2} + \mu_a I e^{-\mu_a z} \tag{17.77}$$

where μ_a is the absorption coefficient. This equation is valid up to the onset of ablation, which is assumed to occur when the surface temperature reaches the ablation threshold temperature, T_{ab}. The details of the analysis of ablation are beyond the scope of this chapter and will not be considered.

In the late 1970s, a nondimensionalization of the heat conduction equation for an axisymmetric three-dimensional case of preablation laser heating of tissue by a Gaussian beam in an absorbing medium was solved. The following relations will transform the governing equations into a dimensionless moving frame:

$$\theta = \frac{T - T_0}{T_{ab} - T_0} \qquad \xi = (I_0 c/kL)\, z \qquad \tau = (I_0 c/kL)(I_0/\rho L)t \tag{17.78}$$

The variables introduced are, respectively, dimensionless temperature θ, dimensionless coordinate ξ in the moving frame with origin at the ablation front, and dimensionless time τ. A dimensionless absorption parameter, B, and a dimensionless heating parameter λ are also defined as

$$\begin{aligned} B &= (kL/I_0 c)\alpha \\ \lambda &= c(T_{ab} - T_0)/L \end{aligned} \tag{17.79}$$

Analytical solutions of the nondimensional form of the governing equations can be found by Laplace transformation of the space variable ξ. The solution is as follows:

$$\theta(\xi, \tau) = \frac{1}{B\lambda}\left\{ 2B\sqrt{\tau}\ \text{ierfc}\left[\frac{\xi}{2\sqrt{\tau}}\right] - e^{-B\xi} + \frac{1}{2}e^{B^2\tau}\left(e^{-B\xi}\ \text{erfc}\left[B\sqrt{\tau} - \frac{\xi}{2\sqrt{\tau}}\right] + e^{B\xi}\ \text{erfc}\left[B\sqrt{\tau} + \frac{\xi}{2\sqrt{\tau}}\right]\right)\right\} \tag{17.80}$$

This equation describes the temperature field as a function of the dimensionless space variable ξ and dimensionless time τ and is valid up to the onset of ablation when $\theta = 1$. The symbol ierfc indicates the integral of the function erfc, $\text{ierfc}(z) = \int_z^\infty \text{erfc}(t)dt$. This equation can be implemented with relative ease using modern user-friendly computer tools such as MATLAB™ or Mathematica™. An example of the graph of this solution, which is the progression in time of the temperature profile as function of depth, is shown in Figure 17.16.

By letting $\theta = 1$ (i.e., $T = T_{ab}$) at $\xi = 0$ in Eq. (17.78), the following transcendental algebraic equation is obtained, which can be solved numerically for the time for the onset of ablation τ_{ab}:

$$B\lambda = \frac{2}{\sqrt{\pi}} B\sqrt{\tau_{ab}} + e^{B^2\tau_{ab}}\text{erfc}[B\sqrt{\tau_{ab}}] - 1 \tag{17.81}$$

Note that for large values of $B\ \lambda$, the behavior is almost linear with a slope of $\frac{2}{\sqrt{\pi}}$.

The present analysis did not consider the case of scattering tissue. An analytical solution for Q_L could be obtained using the diffusion approximation approach described in Section 17.2. However, implementation of that solution into Eq. (17.72) to solve analytically for temperature would be highly cumbersome at best. Therefore, a solution of temperature field would normally require a numerical approach, such as the finite difference or the finite element method.

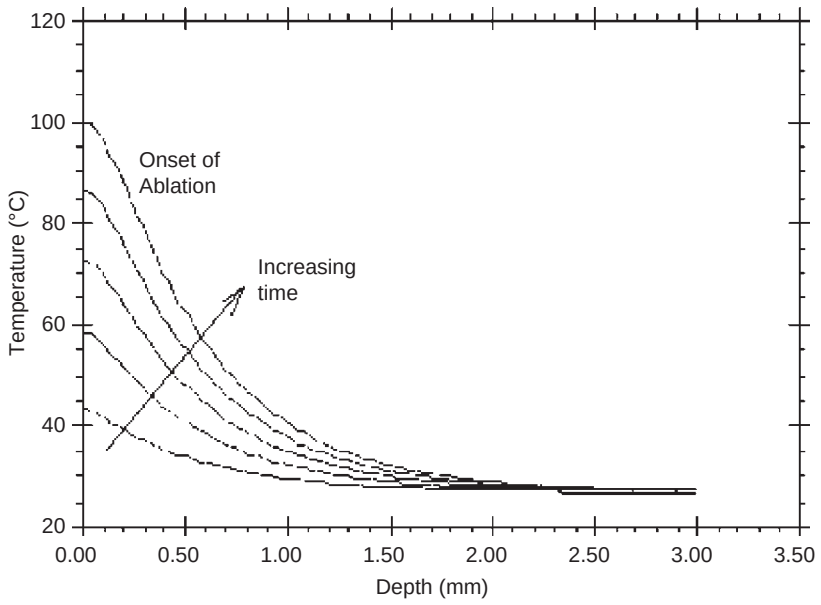


FIGURE 17.16 Nondimensional temperature as a function of nondimensional depth at various times prior to the onset of ablation.

17.5.5 Biochemical Damage Analysis

The Arrhenius model for prediction of damage was introduced previously. The damage function was defined in Section 17.5.2. For numerical computation, it is more convenient (due to large values of A and E) to rewrite the integrand of this function in an equivalent form so Eq. (17.83) becomes

$$\Omega(z,t) = \int_0^t \exp\left(\ln A - \frac{E}{RT(z,t)}\right) dt. \tag{17.82}$$

Two sets of values were used for A and E, one for pigskin and another for human vessel wall. The value of Ω(z,t) can be determined by replacing for T(z,t) from the analytical solution of the last section. The preablation solution should be used up to the onset of ablation, and after that, the ablation stage solution should be used.

The extent of damage or the position of the "damage front" at a given time t can be found by setting $\Omega = 1$ in Eq. (17.82) and solving for z. Alternatively, for every point z, Eq. (17.82) can be integrated by a controlled variable time-step procedure up to the time when $\Omega = 1$. The latter method was used to determine the position of the damage front as a function of time, and the difference in the value integrand at the upper and lower bounds of integration was monitored and used to change the time step. An example of such a calculation was done for a 5-watt laser pulse of 500 ms duration. The calculation was done using both values for A and E from both the pigskin and another for human vessel wall. The results, along with the position of the ablation front as a function of time (details of that are beyond the scope of this chapter), are shown in Figure 17.17.

The vertical line at about 0.0035 seconds indicates the onset of ablation, which helps to show that there had been biochemical damage in the tissue before the onset of ablation. After the onset of ablation, the difference between the damage front and ablation front is the actual extent of damage in the tissue intact. The important observation from this graph is the difference in prediction of the extent of damage using the two different sets of values for A and E. For the same irradiation conditions, thermal damage would propagate faster in skin than it would in vessel wall. This is expected, since the damage mechanisms for vessel wall and skin are different.

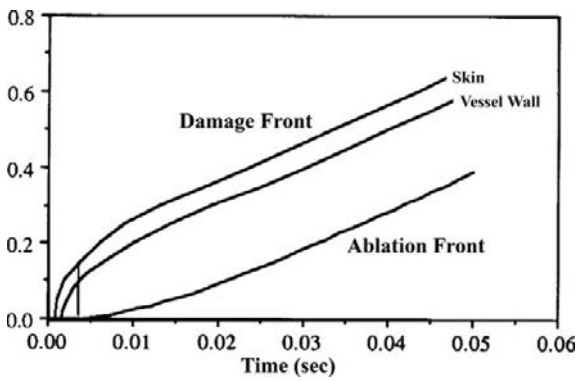


FIGURE 17.17 Damage front as a function of time.

EXAMPLE PROBLEM 17.9

Consider the skin subjected to a constant temperature heating T. For exposure durations t = 1 μs, t = 1 ms, t = 1s, t = 60s, and t = 30 minutes, find the critical temperature, T_c (in °C) required to achieve thermal damage, Ω = 1, during the exposure. Plot T_c versus exposure duration.

Solution

Given that the exposure durations are 1 μs, 1 ms, 1 s, 60 s, and 30 minutes, and R is the universal gas constant = 2 cal/gm-°C, then what is required for each case is for T_c (critical temperature at which point omega = 1) to be calculated.

Assumptions

Values of ln A and E are approximated to pig tissue values of 102.72 and 74,000, respectively.

Equation

$$\Omega(z,t) = \int_0^t \exp\left(\ln A - E/(RT(z,t))\right)dt$$

$$\therefore 1 = \int_0^t \exp\left(\ln A - E/(RT\_c)\right)dt \tag{ex1}$$

$$\therefore 1 = \int_0^t \exp\left(102.72 - 74000/(2{*}T\_c)\right)dt$$

$$\therefore 1 = \exp\left(102.72 - 74000/(2{*}T\_c)\right)\int_0^t dt \tag{ex2}$$

$$\therefore 1 = \exp\left(102.72 - 74000/(2{*}T\_c)\right)t \tag{ex3}$$

Solving Eq. (ex3) for all values of t, we can get T_c for those values. The tabulated values of t and T_c are as follows:

Exposure Time t (in s)	Critical Temperature T_c (in °C)
1 e-6	143.1548
1 e-3	113.153
1	87.2
60	74.56
1,800	72

17.5.6 Effect of Vaporization and Ablation Temperature

The coefficient f_L was introduced previously in relation to the heat of ablation as the water fraction parameter. Another nonvaporization phenomenon that can affect this coefficient is ejection of material due to subsurface nucleation. As was discussed, chunks of the material may be ejected without vaporization. Another factor that can affect the ablation

rate is the excess energy required for pyrolysis, which, as discussed before, manifests itself as a higher ablation threshold temperature.

Although the full solution of the ablation problem in this chapter was not considered, in order to understand how these phenomena affect the ablation rate, the steady-state ablation velocity may be used. This equation can be directly derived to be

$$v_{ss} = \frac{I}{\rho c \Delta T_{ab} + \rho f_L L}. \tag{17.83}$$

Fixing the values of I, ρ, c, and L, the effect of the parameters f_L and ΔT_{ab} on v_{ss} can be determined. In the following example, the physical properties of water are used, and I is chosen to be 267.8 W/cm2 so $v_{ss} = 1$mm/sec for $\Delta T_{ab} = 100$ °C and $f_L = 1$. Figure 17.18 shows the effect of ablation temperature on the ablation velocity for various values of f_L from 0.01 to 1.0. The general effect of an increase either in ablation temperature or in f_L is a decrease in ablation velocity as is also obvious from Eq. (17.83). Observe, however, that the larger the value of f_L, the less effect the ablation temperature has on the ablation velocity. For instance, whereas for $f_L = 0.1$ ablation, velocity drops from 4 mm/sec to 1.15 mm/sec over a temperature change from 100°C to 500°C for the same range, and for $f_L = 0.8$, the change is only from 1.20 mm/sec to 0.69 mm/sec. Therefore, if f_L should indicate the fraction of water in tissue, for example, a decrease in ablation rate should be of concern as a result of higher pyrolysis/ablation temperatures only for tissue with "low" water content (e.g., 30 percent or lower).

A cross section of families of curves in Figure 17.18 at $\Delta T_{ab} = 200$°C for more values of f_L is shown in Figure 17.19.

This figure shows that a change of f_L from 0 to 1 for $\Delta T_{ab} = 200$°C results in a change in the ablation velocity from 3.2 mm/sec to 0.86 mm/sec; that is, the ablation velocity can be

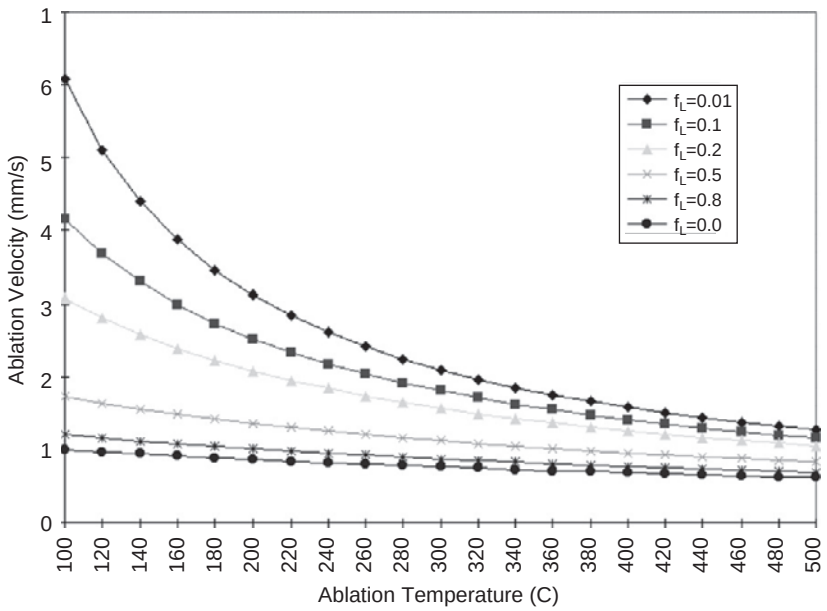


FIGURE 17.18 Effect of ablation temperature on ablation velocity for various water contents.

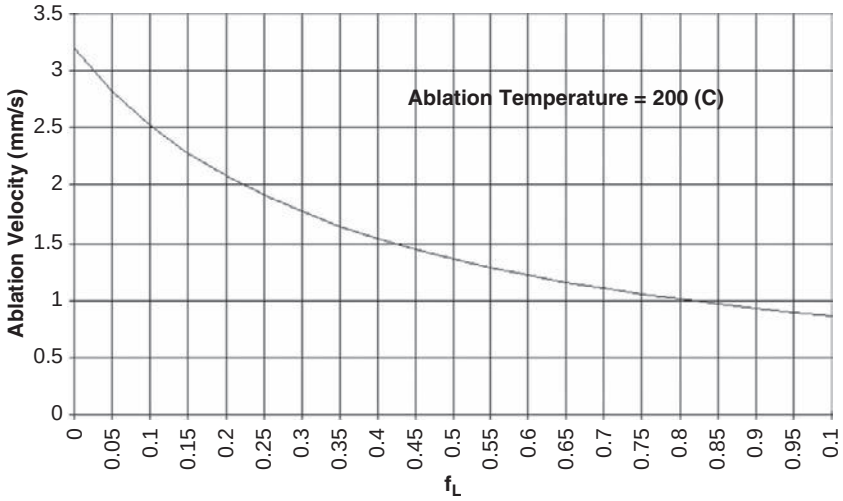


FIGURE 17.19 Effect of water content on ablation velocity.

over three times as large as its value for $f_L = 1.0$. The families of curves of Figure 17.18 are more closely packed at higher values of ablation temperature. This means that the effect of a change in f_L on the ablation velocity is less significant at higher ablation temperatures.

17.6 FIBER OPTICS AND WAVEGUIDES IN MEDICINE

Rigid tubes for the examination of body cavities had already been in use for thousands of years but in the 1800s illumination was added by means of a candle and a 45-degree mirror. The introduction in the early 1900s of multiple lenses to transmit images led to semiflexible tubes for insertion into the body. The use of fiber optic probes based on thin and transparent threads of glass dates back to the late 1920s but lay dormant for two decades until the idea was revived in the 1950s. The first medical instrument, a flexible fiber optic gastroscope, was developed and first used on patients in 1959. In the 1960s, the first lasers were developed, and in the early 1970s, there was a rapid development in the field of fiber optics for communications. All of these events have contributed to the modern fiber optic probes and endoscopes used today.

17.6.1 Principles of Fiber Optics and Waveguides

In Section 17.2.1, the interaction of light with a nonparticipating medium was described. In that section, light was described as rays, and the Fresnel equations for the interaction of the light rays at the boundaries of two media were derived. It was found that, depending on the index of refraction of two slab-types of materials and the angle of incident of the light, the amount of reflection and refraction could change. Although transmission of light through an optical fiber is a complex problem, the phenomenon

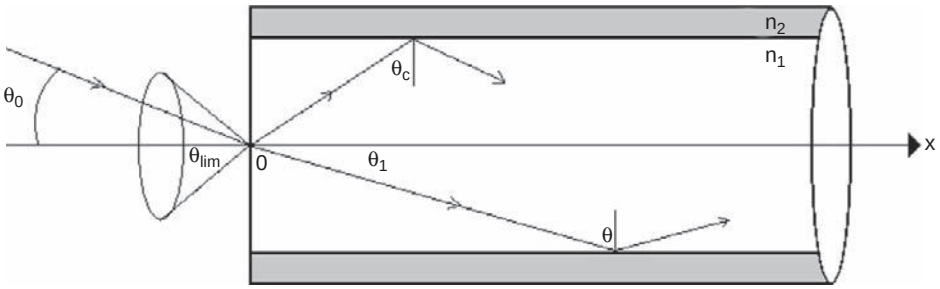


FIGURE 17.20 Basic model of an optical fiber with a cylindrical core with index of refraction (n_1) and cladding index (n_2), where ($n_2 < n_1$).

can be understood with a simple geometrical model. As shown in Figure 17.20, the fiber can be thought of as a long rod of transparent material in which the rod or core of the fiber has a higher index of refraction (n_1) than the surrounding cylindrical shell material or cladding (n_2). The propagation of the light occurs down the core because of the total internal reflection of the light from the core-cladding interface, which, from Snell's laws, occurs when the angle of reflection at this interface is greater than the critical angle. In order for the light to be internally reflected at the core-cladding interface, the light injected into the end of the fiber must be smaller than a cone with some limiting angle θ_o defined by

$$n_0 \sin(\theta_o) = (n_1^2 - n_2^2)^{1/2} = \text{NA} \tag{17.84}$$

where NA, which stands for "numerical aperture," is defined by this limiting value. In silica fibers the NA is generally between 0.2 and 0.4. As can be depicted in Eq. (17.84), increasing the difference in the index of refraction of the core and the cladding will increase the NA and also the acceptance angle. A large NA and acceptance angle produce a large number of rays with different reflection angles or, rather, different transmission modes. A different zigzag of the beam path is thus made for each mode. For a small core diameter, a single mode propagation can be generated in a fiber.

There are many different kinds of optical fibers that have been created with different structure, geometry, and materials, depending on the ultimate application. For instance, the tips of the fibers can be changed to produce side-firing beams for therapeutic applications such as coagulation of prostate tissue. The fiber tips can also be tapered for pinpoint application of the light beam or made as a diffuse tip for broad uniform application of the light. In general, optical fibers have been classified in terms of the refractive index profile of the core and whether there are single modes or multimodes propagating in the fiber. For instance, as shown in Figure 17.21, if the fiber core has a uniform or constant refractive index, it is called a step-index fiber; if it has a nonuniform, typically parabolic refractive index that decreases from the center to the cladding interface, it is known as a graded-index fiber; and if the fiber core is small with a low NA, only a single mode will propagate. The graded index fibers have been shown to reduce the modal dispersion by a factor of 100 times and increase the bandwidth over a comparably sized step-index fiber. These general classifications apply to most fibers, but special fiber geometries are often required, as

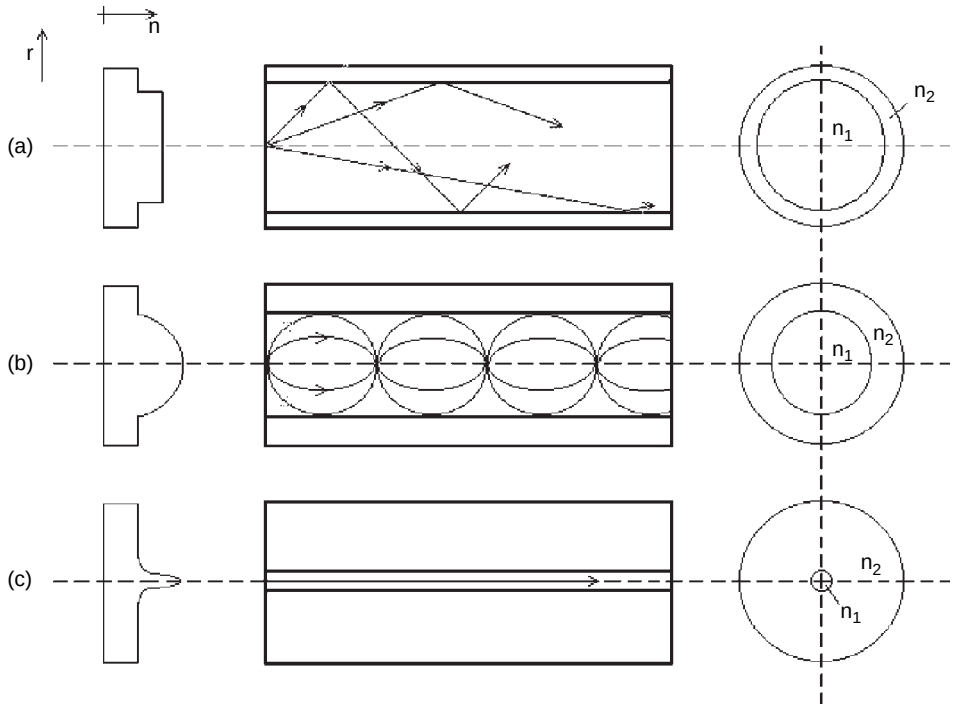


FIGURE 17.21 Typical geometry of the main types of optical fibers including a (a) step index, (b) graded index, and (c) single mode fiber.

discussed following, for endoscopic coherent fiber bundle fiber imaging, single fibers or noncoherent bundles for sensing and diagnostics, and optical fibers made for high-powered therapeutic applications.

In terms of the fiber material, three areas must be considered: the wavelength required, the power, and the biocompatibility. For therapeutic applications the amount of power, both for continuous lasers and pulsed lasers, needs to be considered. In particular, any impurities in the fiber may absorb the light and either decrease the amount of light reaching the probe for sensing or, for therapeutic applications, heat the fiber, which can then cause fiber damage. The fiber itself may also cause a pulsed laser to be broadened in time due to dispersion. In addition, besides impurities, the fiber material itself may work well for one wavelength and not work at all for others. For instance, the standard silica or glass fibers transmit well in the visible wavelength range, but beyond 2.0 micrometers these fibers absorb the light and thus cannot be used for infrared light transmission. Other fibers such as germanium, sapphire, barium fluoride, calcogenide, or hollow waveguide (i.e., gold coated) can be used for these wavelengths, but these fibers, in addition to impurities, may have problems for use in the body such as being too stiff, being hydroscopic (dissolves in water), or being toxic. Thus, the biocompatibility is a function of the material in that it needs to not be toxic to the patient and be able to function in the body without dissolving or breaking off.

17.6.2 Coherent Bundles for Imaging

Optical fibers can be bundled together in an aligned fashion such that the orders of the fibers at both ends are identical. Such a fiber bundle is called a coherent bundle or an ordered bundle. Note that the coherence here means the correlation between the spatial positions of the fibers at both ends and has nothing to do with the light coherence. The important property of coherent bundles is their capability to transmit images through a flexible channel. If an image is projected onto one end of a bundle, a replicate of the image is produced at the other end. Coherent bundles of optical fibers are the key components in endoscopes.

In endoscopic applications, an internal organ is imaged and viewed outside the body in a minimally invasive fashion. An incoherent (nonordered) bundle of optical fibers is used to illuminate the portion to be imaged inside the human body. A coherent (ordered) bundle of optical fibers is used to transmit an image of the target portion. A white light source is usually used for the illumination so an accurate color image of the tissue can be obtained.

The quality of image transmission is mainly determined by two factors: light collection and image resolution. The collection power of each individual optical fiber is limited by the diameter of its core and the numerical aperture (NA). A large NA and core diameter allow a good transmission of light from the illuminated object to the eye of the physician. The image resolution indicates how fine details can be seen and is limited by the core diameter, d, of the cores of the individual fibers. The resolution, in number of discernible lines per millimeter, is approximately $d/2$. The smaller the core diameter, the better the image resolution. Due to the limited resolution, a straight line in an object may appear zigzagged in the image.

Cladding of each individual optical fiber is required to minimize or avoid crosstalking among the fibers. Unclad fibers were used in early endoscopes and had poor imaging quality. Light in unclad fibers may leak from the core and cross into other fibers. The crosstalk causes an overlay of various portions of an image during transmission and results in a blurred image.

The requirements of high light collection and good image resolution have conflicts. A large core diameter allows good light transmission but gives poor resolution. A thick cladding layer avoids crosstalk but limits light collection and image resolution. A tradeoff has to be made. In practice, the core diameter is usually 10–20 μm, and the cladding thickness is of the order of 1.5–2.5 μm.

Several other factors may deteriorate the image obtained with a coherent bundle. Some stray light may transmit through the cladding layers into the cores. The stray light would add an undesirable background that reduces the image contrast. Defective fibers in a coherent bundle would cause a serious problem. If a defective fiber does not transmit any light, a static dark spot appears in the image. If the ordering of the fibers is not identical at both ends, image distortion will degrade the images.

Lenses may be added at both ends of an imaging coherent bundle to adjust the magnification. Although the distance between the objective lens and the fiber bundle can be adjusted in principle to adjust the focusing of the coherent bundle, a fixed focus with a large depth of focus is often used for simplicity. A digital image acquisition board can be used to capture the image in a computer for display, or a VCR can be used to record the images in real time while the images are displayed on a monitor.

Endoscopes may be specially built for imaging of various organs. The following is a list of commonly used endoscopes: angioscopes for veins and arteries, arthroscopes (or orthoscopes) for the joints, bronchoscopes for the bronchial tubes, choledoschoscopes for the bile duct, colonoscopes for the colon, colposcopes for the vagina, cystoscopes for the bladder, esophagoscopes for the esophagus, gastroscopes for the stomach and intestines, laparoscopes for the peritoneum, laryngoscopes for the larynx, and ventriculoscopes for the ventricles in the brain.

17.6.3 Diagnostic and Sensing Fiber Probes

In terms of sensors for minimally invasive measurements into the body, the use of fiber optic probes began to thrive beginning in the 1980s. To understand the potential capability of these probes, a sensor is first defined along with some of the requirements for a good sensor. A sensor is a device that transforms an input parameter or measurand into another parameter known as the signal. For instance, a displacement membrane on the tip of a fiber probe could transform a pressure signal into a light intensity change that is then depicted as a change in voltage from a light detector. The requirements for any good sensor are specificity or the ability to pick out one parameter without interference from the other parameters, sensitivity or the capability to measure small changes in a given measurand, accuracy or closeness to the true measurement, and low cost. As with most sensors, fiber optics have to trade off these parameters to within some limit. For example, you may be able to get 95 percent accuracy at a reasonable cost, but obtaining 99 percent accuracy might require a huge increase in cost, so one must trade off what might be clinically acceptable given the cost. Fiber optic probes offer the potential to meet the preceding sensor requirements, as well as provide for miniaturization; good biocompatibility for the visible and near-infrared wavelengths; speed, since light is used; and safety, since no electrical connections to the body are required.

All fiber optic probes transmit light into the body, and the light, directly or indirectly, interacts with the biological parameter of interest, be it a biological fluid, tissue, pressure, or body temperature. The interaction causes a change in the light beam or beams that travel back to the detection system. The returning signal can be physically separated from the input signal in a separate fiber or fibers, or it can be separated if the output beam is at a separate wavelength from the input beam, as is the case for fluorescence probes. Fiber optic sensors can be used for each of the preceding chemically and physically based measurements described in Sections 17.3 and 17.4. However, rather than try to discuss each fiber optic probe for each application, the rest of this section will be focused on separating the fiber probes into two classes: indirect and direct fiber optic sensors. Direct fiber optic sensors are defined as those probes in which the light interacts directly with the sample. For instance, absorption measurements can theoretically be made to determine such things as blood oxygenation or to quantify blood analytes with either two side firing fiber optic probes separated at a fixed distance, as shown in Figure 17.22a, or an evanscent wave fiber optic, which has the cladding stripped off so some of the light travelling down the core travels out of the fiber, interacts with the sample, and then returns to the fiber core, as shown in Figure 17.22b.

In Figure 17.22c, a fiber optic bundle is depicted that could be used to transmit and receive light to and from tissue to distinguish between normal and cancerous or

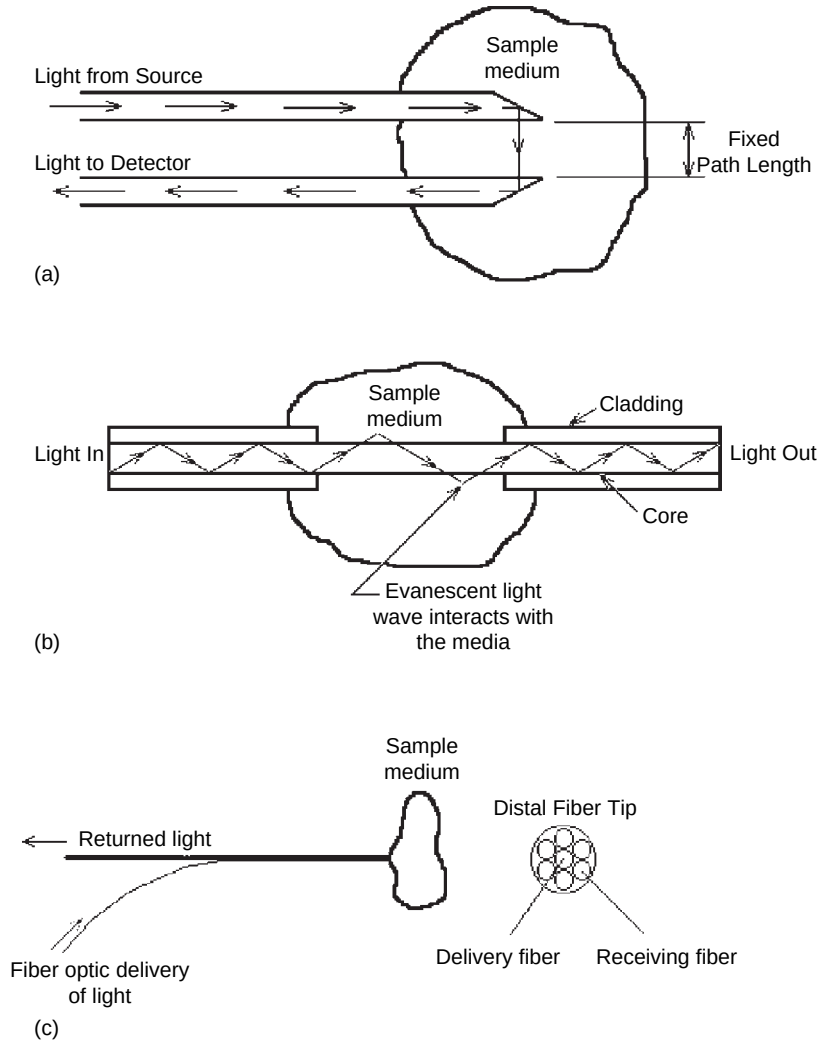


FIGURE 17.22 Direct fiber optic probe designs including (a) absorption probe using two side firing fibers, (b) an evanescent wave probe in which the light transmits from the core to the sample and back into the core, and (c) a multifiber design for use in distinguishing normal from cancerous tissue by directly measuring autofluorescence or Raman spectra.

precancerous lesions using direct measurement of tissue autofluorescence or Raman spectrum. Interferometers, such as the Fabrey-Perot type described in Section 17.3.4, have been designed using partially reflecting mirrors built into a fiber optic, and any physical change imparted on the fiber, such as that due to fluctuations in body temperature or pressure, can be directly measured by the fiber in the form of a change in the interference pattern of the light. Finally, the fiber optic probe for measurement of tissue optical properties, described in Section 17.2.4, is another example of a direct probe that measures reflections from the

sample. It should be noted that each of these descriptions is somewhat oversimplified in that a great deal of design needs to be done to gain the specificity, accuracy, and sensitivity required to measure these various low-level parameters in the noisy environment of the body. For indirect fiber optic measurements, a miniaturized transducer (sometimes referred to as on optode) is attached to the distal end of the fiber so the light interacts with this transducer and the transducer interacts with the sample. A displacement optode such as that shown in Figure 17.23a can be used to monitor pressure or temperature changes by simply causing less light to be specularly reflected into the fiber as the tip is moved. In addition, fluorescent chemistry can be placed within a membrane on the distal tip of the fiber, as shown in Figure 17.23b, in which the amount of fluorescent light produced is a measure

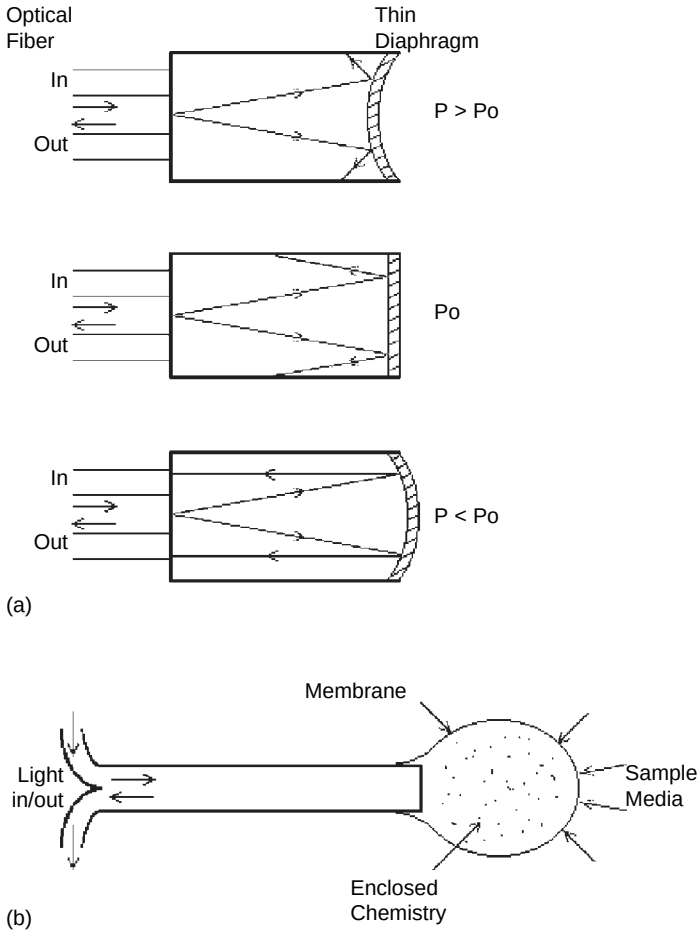


FIGURE 17.23 Indirect fiber optic probe designs including (a) a thin reflectance diaphragm for pressure or temperature measurements, and (b) an optode with some chemistry enclosed within a membrane at the distal tip of the fiber, which can be used, for instance, to determine a particular analyte concentration using fluorescent chemistry specific to the analyte of interest.

of the concentration of a particular blood analyte. These indirect methods of measurement are also simplified, and the actual final design to obtain an accurate and sensitive signal can be quite complicated. Overall, the indirect fiber optic measurements typically have higher specificity over direct measurement approaches but at the expense of requiring a more complicated probe.

17.7 BIOMEDICAL OPTICAL IMAGING

Medical imaging has revolutionized the practice of medicine in the past century. Physicians are empowered to "see" through the human body for abnormalities noninvasively and to make diagnostic decisions rapidly, which has impacted the therapeutic outcomes of the detected diseases.

Medical imaging dates back to 1895 when x-rays were discovered serendipitously by German physicist W. C. Roentgen, who received the first Nobel Prize in physics in 1901 for this important work. Human anatomy can be easily imaged by simple x-ray projections, such as chest x-rays, that are still used today. Contemporary medical imaging began with the invention of computerized tomography (CT) in the 1970s. G. N. Hounsfield in England produced the first computer reconstructed images experimentally, and A. M. Cormack in the United States laid the theoretical foundation. Both of them were awarded the Nobel Prize in medicine in 1979. The essence of CT is that if an object is viewed from a number of different angles, then a cross-sectional image of it can be computed—that is, "reconstructed."

The advent of CT has inspired other new tomographic and even 3D imaging techniques. The application of reconstruction to conventional nuclear medicine imaging led to positron emission tomography (PET) and single photon emission computed tomography (SPECT). A similar application to the technique of nuclear magnetic resonance led to magnetic resonance imaging (MRI).

Different imaging modalities are used to detect different aspects of biological tissues through a variety of contrast mechanisms. X-ray imaging senses primarily electron density and atomic number. In MRI, proton density and its associated relaxation properties are detected. Ultrasonography images use acoustic impedance mismatches for contrast. In nuclear imaging, nuclear radiation emitted from the body is detected after introducing a radiopharmaceutical inside the body to tag a specific biochemical function.

The advantages and disadvantages of different imaging modalities may be illustrated using breast cancer detection. Breast cancer is the most common malignant neoplasm and the leading cause of cancer deaths in women in the United States. A means for prevention of breast cancer has not been found, and early detection and treatment are the best solutions to improving the cure rate. At present, x-ray mammography and ultrasonography are clinically used for breast cancer detection. Mammography is currently the only reliable means of detecting nonpalpable breast cancers. As a supplementary tool, ultrasound is used to evaluate the internal matrix of circumscribed masses found using mammography or of palpable masses that are obscured by radiographically dense parenchyma using mammography. However, x-ray mammography is ionizing radiation, and imaging of radiographically dense breasts is difficult. Ultrasonography cannot detect many of the nonpalpable cancers that are not visible on mammograms of good quality.

Several other techniques are under investigation for breast cancer imaging. Magnetic resonance imaging (MRI) offers great promise for imaging of the radiographically dense breast. Breast MRI is superior to mammography in differentiating solid from cystic lesions and is equivalent to mammography in providing information regarding different parenchymal patterns. Injection of intravenous contrast material with MRI increases cancer detectability even though breast cancer and glandular tissues have similar magnetic resonance tissue characteristics. However, breast MRI is expensive, has inferior spatial resolution to mammography, and cannot image microcalcifications. Breast CT has been investigated for the differentiation of benign from malignant solid masses. Because breast CT involves the use of intravenous injection of iodinated contrast material and is expensive, it is not suited for routine breast cancer screening.

17.7.1 Optical Tomographic Imaging

Nonionizing optical tomography is a new and active research field, although projection light imaging was investigated as early as 1929. The optical properties of normal and diseased tissues are usually different despite the large variation of values in optical properties of the normal tissues alone. Therefore, it is possible to detect some breast cancers based on measurements of optical properties.

The optical difference is not surprising because cancerous tissues manifest significant architectural changes at the cellular and subcellular levels, and the cellular components that cause elastic scattering have dimensions typically on the order of visible to near-IR wavelengths. Some tumors are associated with vascularization, where blood causes increased light absorption. The use of optical contrast agents may also be exploited to enhance the optical contrast between normal and abnormal tissues. Because the optical information is determined by the molecular conformations of biological tissues, optical imaging is expected to provide sensitive signatures for early cancer detection and monitoring.

Because tissues are optically turbid media that are highly scattering, light is quickly diffused inside tissues as a result of frequent scattering. The strong scattering has made optical detection of biological tissues challenging. A typical scattering coefficient for visible light in biological tissues is 100 cm^{-1} in comparison with 0.2 cm^{-1} for x-rays used in medical diagnostics. Light transmitted through tissues is classified into three categories: ballistic light, quasi-ballistic light, and diffuse light. Ballistic light experiences no scattering by tissue and thus travels straight through the tissue. Ballistic light carries direct imaging information just as x-ray radiation does. Quasi-ballistic light experiences minimal scattering and carries some imaging information. Multiply-scattered diffuse light carries little direct imaging information and overshadows ballistic or quasi-ballistic light in thick biological tissue.

One of the techniques used for optical tomography is called "early-photon imaging." If diffuse light is rejected, and ballistic or quasi-ballistic light is collected, buried objects can be detected much like x-ray projection. This technique uses a short-pulse laser (<1 ps pulse width) to illuminate the tissue. Only the initial portion of transmitted light is allowed to pass to a light detector, and the late-arriving light is gated off by a fast optical gate. Because the ballistic or quasi-ballistic photons travel along the shortest path length, they arrive at the detector sooner than diffuse photons. If only ballistic light is detected, the technique is called ballistic imaging. It has been shown that ballistic imaging is feasible only for tissue

of thickness less than 1.4 mm or 42 mean free paths. Most ballistic imaging techniques reported in the literature have achieved approximately 30 mfp. Therefore, this approach is suitable for thin tissue samples but suffers loss of signal and resolution for thick tissues as a result of the strong scattering of light by the tissue.

EXAMPLE PROBLEM 17.10

Calculate the decay of ballistic light after penetrating a biological tissue 30 mean free paths thick. If the scattering coefficient of the tissue is 100 cm^{-1}, calculate the corresponding thickness in cm.

Solution

Based on Beer's law, the decay is $\exp(-30) = 9.4 \times 10^{-14}$. The thickness is $30/100 = 0.3$ cm.

For tissue of clinically useful thickness (5 to 10 cm), scattered light must be used to image breast cancers. It has been shown that for a 5-cm-thick breast tissue with an assumed absorption coefficient of 0.1 cm^{-1}, and reduced scattering coefficient of 10 cm^{-1}, the detector must collect transmitted light that has experienced at least 1,100 scattering events in the tissue to yield enough signal. Therefore, ballistic light or even quasi-ballistic light does not exist for practical purposes. However, if a 10 mW visible or near-infrared laser is incident on one side of the 5-cm-thick breast tissue, it has been estimated, using diffusion theory, that the diffuse transmittance is on the order of 10 nW/cm^2 or 1,010 photons/(s cm^2), which is detectable using a photomultiplier tube. Similarly, the diffuse transmittance through a 10-cm-thick breast tissue would be on the order of 1 pW/cm^2 or 106 photons/(s cm^2). The significant transmission of light is due to the low absorption coefficient despite the high scattering coefficient.

Imaging resolution of pure laser imaging degrades with increased tissue thickness. The temporal profiles of the scattered light may be detected using a streak camera. The early portion of the profiles was integrated to construct the images of buried objects in a turbid medium. This time-domain technique requires expensive short-pulse lasers and fast light detectors.

Optical-coherence tomography (OCT) has emerged as a useful clinical tool. This technique is based on the Michelson interferometer (see Figure 17.9) with a short-coherence length light source. One arm of the interferometer leads to the sample of interest, and the other leads to a reference mirror. The reflected optical beams are detected at the photodetector. The two beams interfere only when the sample and the reference path lengths are equal to within the source coherence length. Heterodyne detection is performed by taking advantage of the direct Doppler frequency shift that results from the uniform high-speed scan of the reference path length. Recording the interference signal magnitude as a function of the reference mirror position profiles the reflectance of the sample, which produces an image similar to an ultrasonic A scan. OCT has achieved less than 10 μm resolution with a penetration depth of approximately 1 mm.

OCT was extended to image blood flow in superficial vessels based on Doppler shift. The blood flow causes a Doppler shift on the frequency of the light. Frequency analysis with Fourier transformation of the optical interference signal yields the Doppler shift, which is used to calculate the velocity of the blood flow.

A technique for optical imaging of thick tissue is the frequency-domain technique, which is based on photon-density waves. The governing equation of photon-density waves is the diffusion equation

$$\partial\Phi(r,t)/c\partial t + \mu_a\Phi(r,t) - D\nabla^2\Phi(r,t) = S(r,t) \tag{17.85}$$

where $\Phi(\boldsymbol{r},t)$ is the light fluence rate at point $\boldsymbol{r}$ and time t [W/cm^2], c is the speed of light in the medium [cm/s], μ_a is the absorption coefficient [cm^{-1}], and $S(\boldsymbol{r},t)$ is the source intensity [W/cm^{-3}]. The frequency-domain imaging technique requires the use of inverse algorithms for image reconstruction. Amplitude-modulated (at approximately 100 MHz) laser light is used to illuminate the tissue at multiple sites. At each illumination site, an optical detector measures the amplitude and phase of diffuse light at multiple locations around the tissue. The measured diffuse light can be estimated by the diffusion equation, Eq. (17.85), if the optical properties of the tissue sample are known. Conversely, the optical properties can be calculated by use of the measured diffuse light, which is the image reconstruction. A sample reconstructed image is shown in Figure 17.24, which was based on a theoretically generated data set of diffuse light.

17.7.2 Hybrid Optical Imaging

Several emerging imaging techniques are being developed by combining relatively transparent acoustic energy with strongly scattering light, which is called hybrid optical imaging. Ultrasound-modulated optical tomography, photoacoustic tomography, and sonoluminescence are briefly discussed.

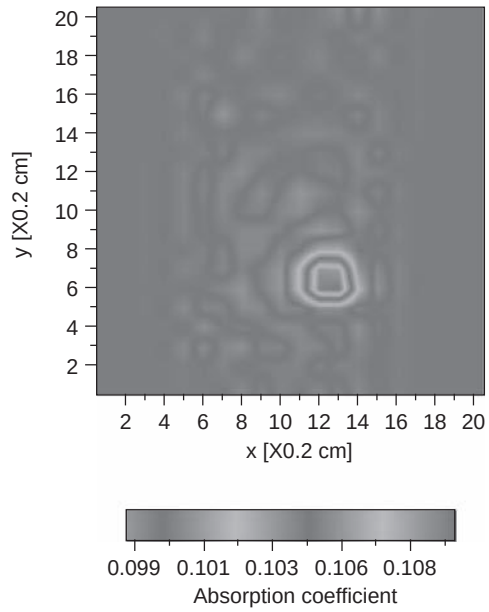


FIGURE 17.24 Sample reconstructed image using photon-density waves. A simulated tumor with elevated absorption coefficient is visible in the lower right quadrant.

In ultrasound-modulated optical tomography, an ultrasonic wave is focused into a scattering medium to modulate the laser light passing through the medium containing buried objects. The modulated laser light collected by a photodetector is related to the local mechanical and optical properties in the zone of ultrasonic modulation. If the buried objects have optical properties that are different from those of the background scattering medium, an image can be obtained by raster-scanning the device.

In photoacoustic tomography, a short-pulse light beam illuminates the scattering medium. The light is diffused in the medium and partially absorbed. If there is a strong optical absorber such as a tumor in the middle of the medium, more light will be absorbed by this optical absorber than by its neighboring background. The absorbed optical energy is converted into heat. Due to thermal elastic expansion, an acoustic wave is generated. Stronger heat generation will produce a stronger acoustic wave. Therefore, a strong optical absorber emanates a strong acoustic wave. If multiple acoustic transducers are used to measure the acoustic signal around the medium, the absorber can be located based on the temporal distribution of the acoustic signals and thus produce an image of the medium. Functional and molecular imaging has been achieved.

The inverse of photoacoustic tomography is sonoluminescent tomography (SLT). The ultrasonic generation of light known as sonoluminescence (SL) was first reported in 1934, which was multiple-bubble sonoluminescence (MBSL). SL has attracted an extraordinary amount of attention in this decade, since single-bubble sonoluminescence (SBSL) was reported in 1990. Although the full explanation of SL is still in development, it is well known that light is emitted when tiny bubbles driven by ultrasound collapse. The bubbles start out with a radius of several microns and expand to about 50 microns due to a decrease in acoustic pressure in the negative half of a sinusoidal period. After the sound wave reaches the positive half of the period, the situation rapidly changes. The resulting pressure difference leads to a rapid collapse of the bubbles accompanied by the emission of light. The flash time of SL has been measured to be in the tens of picoseconds. SBSL is so bright that it can be seen by the naked eye even in a lighted room, whereas MBSL is also visible in a darkened room. Researchers have envisioned possible applications of SL in sonofusion, sonochemistry, and building ultrafast lasers using the ultrafast flash of light in SL.

SLT is an application of SL that has been developed for cross-sectional imaging of strongly scattering media noninvasively. Sonoluminescence, which is generated internally in the medium by exposure of the medium to external ultrasound, is used to produce images of a scattering medium by raster-scanning the medium. The spatial resolution is limited by the focal spot size of the ultrasound and can be improved by tightening the focus.

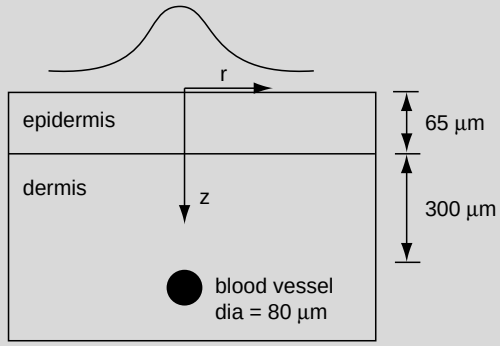
Optical imaging techniques may also measure the optical spectra of biological tissues noninvasively. Besides the absorption spectra that are usually measured for other materials, the scattering spectra may also be measured. The optical spectra may be used to quantify important physiological parameters such as the saturation of hemoglobin oxygen.

In summary, optical imaging and biomedical optics techniques in general have the advantages of (1) the use of nonionizing radiation, (2) the capability of measuring functional (physiological) parameters, (3) the potential high sensitivity to the pathologic state of biological tissues, and (4) low cost. However, biomedical optics is a challenging research field in its infancy, and as it grows, it will continue to require the participation of many diverse talented physicians, scientists, and engineers.

17.8 EXERCISES

1. (a) Name at least three reasons why it might be desirable to use optic or fiber optic sensors for biomedical applications. (b) Name three potential drawbacks to using optic or fiber optic biomedical sensors.
2. In one paragraph, or using bullets, explain the similarities and differences among polarization, luminescence, absorption, scatter, and Raman scattering of light.
3. Given an interface between clear tissue (i.e., water; $n_t = 1.33$) and glass ($n_g = 1.5$), compute the transmission angle for a beam incident in the water at 45°. If the transmitted beam is reversed so it impinges on the interface from the glass side, calculate what the transmitted angle will be through the tissue. Does this make sense? Explain.
4. Using MATLAB, or some other software, plot ($R_\perp = r_\perp{}^2$) and ($R_p = r_p{}^2$) as a function of the input angle θ_i for a glass ($n_i = 1.5$) and air ($n_t = 1.0$) interface. Explain what it means.
5. An Nd:YAG laser operating at 1,064 nm with a Gaussian beam is to be used for performing a port-wine stain treatment. Assume a heuristic model for light distribution in a homogeneous material such that $I(r,z) = I_o \exp(-\mu z) \exp(-2(r/\omega_b)^2)$, where $\mu = \mu_a + \mu'_s$, μ_a and μ'_s being the absorption and reduced scattering coefficients. For the following multilayer case at r = 0, give the expressions and graph intensity I and the rate of heat generation Q in the tissue by the laser, $I(r = 0,z)$ and $Q(r = 0,z)$. Assume $I_o = 1$. The absorption and scattering coefficients of each skin layer and blood vessel for this wavelength are given below. Discuss how selective photothermolysis of the blood layer is achieved.

	μ_a (cm^{-1})	μ'_s (cm^{-1})
Blood	20	20
Epidermis	0.1	10
Dermis	0.5	10



6. Consider a turbid biological tissue with an absorption coefficient of $\mu_a = 0.1\ \text{cm}^{-1}$. Using the diffusion approximation, graph on the same plane the logarithm of relative intensity versus radius for four different values of reduced scattering coefficient—$\mu'_s = 1$, 10, 50, and 100 cm^{-1}—r varying from 0.1 cm to 2 cm. Describe the results in terms of the effect of scattering on the intensity profile.
7. For liver tissue irradiated by the pulsed Ho:YAG laser, determine the exposure duration during which the equation for heat conduction may be ignored. Assume that the absorption coefficient at the Ho:YAG laser wavelength is 30 cm^{-1}, and assume the thermal diffusivity of the tissue is the same as that of water. If the threshold temperature of ablation is 150°C, determine the threshold fluence (J/cm^2) for ablation.

8. Given a Helium-Neon laser-based fiber optic Doppler probe at an angle of 60° with a blood vessel that registers a frequency shift of 63 KHz, what is the velocity of the blood? Is this a reasonable number from a physiologic point of view when compared to the average velocity in the major and minor blood vessels in the human body?
9. Assume you want to measure the thickness of a piece of tissue ($n_t = 1.33$) bounded by air ($n_a = 1.0$) using either the Michelson or Fabry-Perot interferometric approach. Your light source has a wavelength of 780 nm, and your instrument can count 20 fringes. The tissue is ablated (or cut) with a high-powered pulsed laser that removes roughly 2 micrometers of tissue per pulse. The maximum tissue thickness before it is all removed is 20 micrometers. (a) Calculate to see if each of the interferometer systems will have a long enough dynamic range, without moving the system, to be used with this piece of tissue. Explain your result. (b) Assume you want to use either of your interferometric systems as a feedback unit for the laser removal of tissue. What is the range in the number of fringes (m_{min} and m_{max}) for each of your systems in order to measure both the minimum slice thickness and the maximum tissue thickness? Explain.
10. In general, when is a Fabry-Perot interferometer preferred over a standard Michelson interferometer? Can you make a Fabry-Perot interferometer that performs worse than a Michelson interferometer? Explain.
11. For Raman spectroscopy, what are the Stokes and anti-Stokes bands? Which bands are typically used for sensing? What are the main challenges to overcome for Raman spectroscopy to be used for biomedical sensing?
12. Assume you want to measure glucose through the anterior chamber of the eye as a means of noninvasively quantifying blood glucose. Given that glucose has a specific rotation of 41.89 degrees/(dm g/ml) at a wavelength of 656 nm, and the anterior chamber of the eye has a path length of roughly 0.8 cm, calculate the concentration of glucose for a rotation of 15 millidegrees. Is this a reasonable value from a physiologic point of view? Would the patient be considered normal or diabetic? Explain your result.
13. In the near-infrared region of the optical spectrum between 600 and 1,100 nanometers, there are well-known absorbance peaks for oxygenated and deoxygenated hemoglobin. A big assumption can be made for the moment that in this region the dominant optical signal is due to absorption (which is not generally the case). Given that you propagate through a 2-centimeter sample of tissue, you need to calculate three parameters: the concentration of oxyhemoglobin, deoxyhemoglobin, and a background blood absorbance. At three wavelengths (758 nm, 798 nm, and 898 nm), you can measure the extinction coefficients for oxygenated (1.612, 2.072, and 2.763 $mM^{-1}\,cm^{-1}$) and deoxygenated (3.914, 2.072, and 2.072 $mM^{-1}\,cm^{-1}$) hemoglobin, respectively. The total absorption coefficients at these three wavelengths in the tissue can be measured using a time-resolved system as 0.293 cm^{-1}, 0.1704 cm^{-1}, and 0.1903 cm^{-1}, respectively. Calculate the oxygenated and deoxygenated hemoglobin levels given these parameters. Is this reasonable from a physiologic point of view? Explain.
14. Using the parameters of the equation for problem 7, give a graph of temperature versus time at $z = 0$, for t varying from 0 to 1 ms, using the analytical solution of the heat conduction

Continued

equation given in the chapter. On the same graph, plot the temperature, ignoring the effect of conduction, which assumes linear variation with time. Comment on your findings in relation to the diffusion time found in problem 7.

15. Consider skin tissue that is exposed to linear heating—for example, as in pulsed laser coagulation—such that $T = T_o + \mathrm{mt}$, when T_o is the initial temperature (e.g., 37°C) and m is the rate of heating (e.g., 10^3 °C/s). Describe how you would find the critical temperature for thermal damage. Do you expect that to be lower or higher than the constant temperature case of the same exposure duration?

16. For the study of ablation onset time, plot A graph of $\mathrm{B}\sqrt{\tau_{\mathrm{ab}}}$ versus $\mathrm{B}\lambda$. Measure the slope for "large" values of $B\ \lambda$; it should be $\frac{2}{\sqrt{\pi}}$. For absorption coefficient 100/cm and a laser intensity 1,000 W/cm^2, find the time for the onset of ablation using water thermal properties, assuming ablation initiates at 150°C.

17. In the visible optical region, tissue scattering dominates absorption. Assume that the scattering coefficient of a 5-cm-thick tissue is 100 cm^{-1} and the wavelength of light is 0.5 μm. In order to detect on average a single ballistic photon transmitted through the tissue, what is the energy in Joules that is required for the incident light? Compare the energy of the incident light with the rest energy of the earth using Einstein's mass-energy equivalence equation $E = mc^2$, where the mass of the earth is 6×10^{24} kg.

18. Use the diffusion theory to estimate light penetration in biological tissues. A 10 mW isotropic point source is buried in an infinite turbid medium. Assume that the absorption coefficient is 0.1 cm^{-1}, the scattering coefficient of a 5-cm-thick tissue is 100 cm^{-1}, the scattering anisotropy is 0.9, and the wavelength of light is 0.5 μm. Calculate the light fluence rate 5 cm from the source.

19. Use diffusion theory to estimate sonoluminescence light transmission in biological tissues. A 1 mW isotropic point source is buried in an infinite turbid medium. Assume that the absorption coefficient is 0.1 cm^{-1}, the scattering coefficient is 100 cm^{-1}, and the scattering anisotropy is 0.9. Calculate the light fluence rate integrated over a sphere of a 5 cm radius centered at the source.

20. What additional information does the phase measurement in a frequency-domain imaging technique provide compared with the continuous wave technique that measures only the amplitude of the diffuse light?

Suggested Readings

G. Boisdé, A. Harmer, Chemical and Biochemical Sensing with Optical Fibers and Waveguides, Artech House, Boston and London, 1996.

M. Born, E. Wolf, Principles of Optics, seventh expanded ed., Cambridge University Press, 1999.

B. Culshaw, J. Dakin, Optical Fiber Sensors: Vols. I and II, Artech House, Boston and London, 1989.

F.A. Duck, Physical Properties of Tissue, Academic, London, 1990.

E. Hecht, Optics, Addison-Wesley, Reading, MA, 1987.

A. Katzir, Lasers and Optical Fibers in Medicine, Academic Press Inc., San Diego, CA, 1993.

J. Lakowicz, Principles of Fluorescence Spectroscopy, second ed., Kluwer Academic/Plenum Publishers, 1999.

F.L. Pedrotti, L.S. Pedrotti, Introduction to Optics, second ed., Prentice Hall, Upper Saddle River, NJ, 1993.

V. Tuchin, Tissue Optics: Light Scattering Methods and Instruments for Medical Diagnosis, Society of Photo-Optical Instrumentation Engineers Press, Bellingham, Washington, 2000.

T. Vo-Dinh, Biomedical Photonics Handbook, CRC Press, January 24, 2003.
L.V. Wang, H.-I. Wu, Biomedical Optics: Principles and Imaging, Wiley, 2007.
A.J. Welch, M.J.C. van Gemert, Optical-Thermal Response of Laser Irradiated Tissue, Plenum Press, New York and London, 1995.
H.H. Willard, L.L. Merritt., J.A. Dean, F.A. Settle, Instrumental Methods of Analysis, seventh ed., Wadsworth Publishing Inc., Belmont, CA, 1988.

This page intentionally left blank

Appendix

John D. Enderle, PhD

A.1 MATLAB

MATLAB is a high-level computer program that performs technical computing and plotting based on matrices. It is a programming language that has many of the same capabilities as FORTRAN, C and other programming languages, but is much easier to use. This section introduces the reader to MATLAB in sufficient detail so that the program can be used to solve relevant problems in this book. Other MATLAB features are introduced in later chapters as needed.

A.1.1 Matrix Basics

MATLAB was originally developed as MATrix LABoratory, with most commands stated in terms of matrices. A matrix is a rectangular array consisting of n rows and m columns of elements. The array $\mathbf{A}$, denoted with a bold uppercase letter, is denoted as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{A.1}$$

where each element of the array, a_{ij}, is a constant or a function. The order of array $\mathbf{A}$ is referred to as $n \times m$. If $n = m$, $\mathbf{A}$ is called a square matrix of order n. For our purposes, the matrix $\mathbf{A}$ usually stores the parameters for a system, such as values for resistors, capacitors and inductors.

An array of one column is referred to as a column vector, that is, an array of order $n \times 1$. We typically denote a column vector with a bold lowercase letter, such as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{A.2}$$

where each element x_i is a variable, constant or a function of time. The column vector is useful for representing variables and inputs for a system. An array of one row is referred to as a row vector, that is, an array of order $1 \times n$. The row vector is useful for representing a polynomial of the characteristic function (defined later) for a system. A matrix of order 1×1 is a scalar.

Addition and subtraction of matrices are valid only for matrices of the same order. Either operation is applied to corresponding elements term by term, that is,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix}$$
$$= \begin{bmatrix} c_{11} = a_{11} + b_{11} & c_{12} = a_{12} + b_{12} & \cdots & c_{1m} = a_{1m} + b_{1m} \\ c_{21} = a_{21} + b_{21} & c_{22} = a_{22} + b_{22} & \cdots & c_{2m} = a_{2m} + b_{2m} \\ \vdots & \vdots & & \vdots \\ c_{n1} = a_{n1} + b_{n1} & c_{n2} = a_{n2} + b_{n2} & \cdots & c_{nm} = a_{nm} + b_{nm} \end{bmatrix} \tag{A.3}$$

where, in general, $c_{ij} = a_{ij} + b_{ij}$. Subtraction follows similarly with $c_{ij} = a_{ij} - b_{ij}$.

Matrix multiplication is valid only when the number of columns in the first matrix is equal to the number of rows in the second matrix. If **A**, order $n \times m$, is multiplied by **B**, order $m \times n$, then the order of the resulting matrix **C** is $n \times n$. Each element of **C**, c_{ij}, is equal to the sum of products of the i^{th} row of **A** with the j^{th} column of **B**, that is

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{A.4}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. In general, matrix multiplication for matrices of the same order is not commutative, that is, $\mathbf{AB} \neq \mathbf{BA}$. The multiplication of a matrix by a scalar α equals the product of each element of the matrix by α.

The identity matrix **I** is a square matrix whose nondiagonal elements are zero and whose diagonal elements are one, that is

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{A.5}$$

The null matrix, **0**, has 0 for all elements in the matrix.

Matrix operations of addition, subtraction, multiplication, and division follow much the same processes that these operations do with real numbers. For arbitrary and appropriately defined matrices **A**, **B**, and **C**, we have

Commutative Property: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

Associative Property: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$

Distributive Property: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

Identities Involving I and 0:

$$\mathbf{AI} = \mathbf{A}$$
$$\mathbf{0A} = \mathbf{0}$$
$$\mathbf{A0} = \mathbf{0}$$
$$\mathbf{A} + \mathbf{0} = \mathbf{A}$$

The inverse of a square matrix $\mathbf{A}$, denoted, $\mathbf{A}^{-1}$, is defined by

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \tag{A.6}$$

The use of the matrix inverse is important in solving a set of simultaneous equations that describe a system.

The transpose of matrix $\mathbf{A}$ is denoted as either $\mathbf{A}'$ or $\mathbf{A}^T$, where we interchange rows and columns of $\mathbf{A}$ to form $\mathbf{A}^T$ as follows from Equation (A.1).

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$$

SIMULTANEOUS EQUATIONS AND MATRICES

A set of equations with unknown variables often results in the process of analyzing a system, after applying interconnection laws or other techniques. At other times, a set of equations with unknown parameters results in solving the coefficients of a differential equation with initial conditions.

Suppose the following set of equations describes the currents in an electric circuit, i_i, i_2, and i_3, after applying Kirchhoff's voltage law.

$$\begin{aligned} 2i_1 - i_2 + 0i_3 &= 6 \\ -2i_1 + 3i_2 + i_3 &= 0 \\ -i_1 + 5i_2 - 4i_3 &= 0 \end{aligned} \tag{A.7}$$

The constants multiplying the currents involve the resistors in the circuit, with the value of 6 due to the input to the circuit. In writing each equation in (A.7) we have included all variables, written in order, even if a variable is multiplied by 0.

Equation (A.7) is written in matrix form as

$$\begin{bmatrix} 2 & -1 & 0 \\ -2 & 3 & 1 \\ -1 & 5 & -4 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \tag{A.8}$$

or

$$\mathbf{A}\mathbf{i} = \mathbf{F} \tag{A.9}$$

with appropriately defined matrices

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -2 & 3 & 1 \\ -1 & 5 & -4 \end{bmatrix}, \mathbf{i} = \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix}, \text{ and } \mathbf{F} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$$

To solve Equation (A.8) for the current vector $\mathbf{i}$, we premultiply both sides of the equation by the inverse $\mathbf{A}^{-1}$, that is

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{i} = \mathbf{A}^{-1}\mathbf{F} \tag{A.10}$$

or

$$\mathbf{i} = \mathbf{A}^{-1}\mathbf{F} \tag{A.11}$$

A.1.2 Using MATLAB

MATLAB is a program that evaluates mathematical expressions, plots results, and comes with a wide assortment of toolboxes (applications) for various engineering fields. The toolboxes include neural networks, imaging, digital signal processing, control theory, linear algebra, signals and systems, numerical methods, symbolic math, and SIMULINK. SIMULINK provides an iterative solution (i.e., an approximation) to very complex differential-integral equations. The plotting capabilities of MATLAB are sufficient for most work. For more sophisticated or professional plotting, output from MATLAB can be exported into Microsoft® EXCEL.

The material presented here is a very brief introduction to MATLAB. Readers interested in learning more about the program should look in the library and bookstores to find many books on this powerful program.

GETTING STARTED

To begin a MATLAB session in Microsoft Windows, click "Start", "Programs", the folder "MATLAB", and then the program "MATLAB". A command window opens with three parts as shown in Figure A.1.

Our work at this stage is done in the MATLAB Command Window on the right side of Figure A.1. To begin, we start typing in the Command Window after the ">>". Text can be typed interactively with MATLAB executing the command after a carriage return (called CR hereafter) by clicking the "Enter" key. Keep in mind that MATLAB differentiates

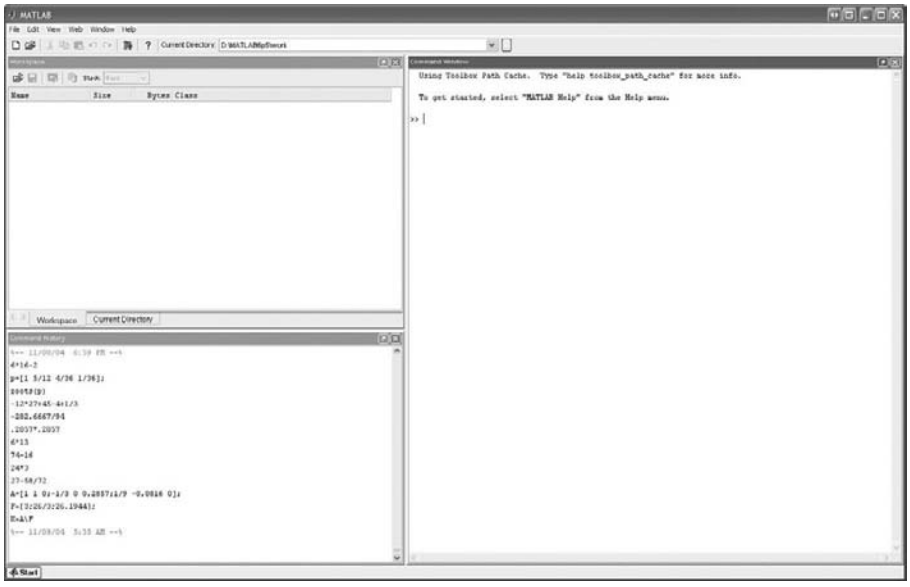


FIGURE A.1 MATLAB program with three windows. (1) Launch Pad Window, (2) Command Window, and (3) Command History.

between lower and uppercase in executing commands (so "a" is different from "A"), and it does not use italic or bold characters as defined in Section A.1.

MATLAB evaluates scalars as well as matrices. For example, if we enter

```
>> A = 10/5
```

MATLAB returns with

```
A =
     2
```

To enter a matrix, we start with the matrix name, then an equal sign followed by a left bracket, the elements of the array, and then a right bracket. Each element in a row is entered, beginning with the first row, separated by a blank space or a comma. To move from one row to the next row, a semicolon or a CR is entered. For example, matrix **A** is entered in MATLAB by writing

```
>> A = [5 9 6; 3 2 3; 5 9 1]
```

MATLAB returns with

```
A =
     5   9   6
     3   2   3
     5   9   1
```

Matrix elements can also be entered as a function, such as

```
>> A = [- 5/2, sqrt(4), 3*(6 - 4)]
```

MATLAB returns with

```
A =
    -2.5000    2.0000    6.0000
```

Note that a ";" at the end of any line suppresses printing of that line. A line can also be extended by typing "..." and a CR at the end of any line so that the next line is just a continuation of the previous line.

For example,

```
>> C = [1 ...
2 3
4 5 6
7 8 9]
C =
     1   2   3
     4   5   6
     7   8   9
```

Here the first row is written on two lines with "..." at the end of the first line and a CR rather than ";" separating the rows.

To invert matrix **A**, given by

```
>> A = [5 9 6; 3 2 3; 5 9 1]
```

type

```
>> B = inv(A)
```

MATLAB returns with

```
B =
   −0.2941    0.5294    0.1765
    0.1412   −0.2941    0.0353
    0.2000    0        −0.2000
```

There are many special functions within MATLAB like the exponential and π. For example, to use π in a calculation, we use the word "pi"; enter the following and note the response.

```
>> pi
ans =
    3.1416
```

MATLAB returns a numerical result for each command entered. The exponential function uses the syntax "exp (− 0.5)", which, when entered, gives 0.6065. Most standard elementary functions in a good scientific calculator, including built-in functions, are included in MATLAB, like the sine and cosine functions. The entire list of MATLAB functions is given from the "HELP" menu: select "MATLAB HELP", then "Functions – Alphabetical List".

A.1.3 Arithmetic Expressions

Arithmetic operations are carried out in MATLAB with the following symbols:

$$\wedge \quad * \quad / \quad \backslash \quad + \quad -$$

for power operator, multiplication, right division, left division, addition, and subtraction, respectively. Note that 1/4 and 4\1 both equal 0.25, and represent 1 divided by 4 and 4 divided into 1, respectively. We often use the "\" operation when solving simultaneous equations as in Example A.6.

MATLAB performs operations in the order previously listed, that is, ^ first, then * or / or \ calculated in order from left to right, and then + or − calculated in order from left to right. For example,

```
>> 5 + 6^2*3/7*2
```

gives

```
ans =
    35.8571
```

This operation is done by first calculating 6 ^ 2 = 36, then multiplying 36 by 3, followed by dividing this result by 7, then multiplying the entire result by 2, and then finally adding 5. The sequence of like math operations works from left to right, with parentheses used to change the order of calculation. For example,

```
>> 3*(6 - 4)
```

gives

```
ans =
     6
```

The first calculation is $6-4$ and then this result is multiplied by 3.

Trigonometric functions are calculated in MATLAB with the angle in radians. For example, to evaluate sin (45°), we enter

```
>> sin (pi/4)
```

which gives

```
ans =
     0.7071
```

EXAMPLE PROBLEM A.1

Solve the following set of simultaneous equations for $i_1, \ldots, i_5$.

$$\begin{aligned} 4i_1 - i_2 - i_3 - i_4 - i_5 &= 240 \\ -i_1 + 8i_2 - i_3 + 0 \times i_4 + 0 \times i_5 &= 0 \\ -i_1 - i_2 + 5i_3 - i_4 + 0 \times i_5 &= 0 \\ -i_1 + 0 \times i_2 - i_3 + 5i_4 - i_5 &= 0 \\ -i_1 + 0 \times i_2 + 0 \times i_3 - i_4 + 8i_5 &= 0 \end{aligned}$$

Solution

The simultaneous equations are written in matrix format as

$$\mathbf{Ai} = \mathbf{F}$$

with

$$\mathbf{A} = \begin{bmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 8 & -1 & 0 & 0 \\ -1 & -1 & 5 & -1 & 0 \\ -1 & 0 & -1 & 5 & -1 \\ -1 & 0 & 0 & -1 & 8 \end{bmatrix}; \quad \mathbf{i} = \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \end{bmatrix}; \quad \mathbf{F} = \begin{bmatrix} 240 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Column vector ***i*** is easily solved by using the reverse division operation on the previous equation, that is,

$$\mathbf{A} \backslash \mathbf{Ai} = \mathbf{i} = \mathbf{A} \backslash \mathbf{F}$$

The MATLAB commands to solve for ***i*** are

```
>> A = [4 -1 -1 -1 -1; -18 -100; -1 -15 -10; -10 -15 -1; -100 -18];
>> F = [240;0;0;0;0];
>> i = A \ F
```

Continued

```
i =
    77.5000
    12.5000
    22.5000
    22.5000
    12.5000
```

A.1.4 Vectors

MATLAB stores vectors as a row vector with the first element as the **i** component, the second the **j** component, and the third the **k** component, all written within brackets. For example, the vector $\mathbf{F} = 3\mathbf{i} - 4\mathbf{j} + 2\mathbf{k}$ is entered in MATLAB as

```
>> F = [3   −4   2]
```

which gives

```
F =
    3   −4   2
```

To calculate the magnitude of the vector, the MATLAB command "norm" is used. For example, the magnitude of the vector F is calculated as

```
>> Fmagnitude = norm(F)

Fmagnitude =
     5.3852
```

MATLAB also computes the dot product of two vectors with the command "dot(A,B)". For example, if $\mathbf{A} = 5\mathbf{i} - 8\mathbf{j} - 2\mathbf{k}$ and $\mathbf{B} = 6\mathbf{i} + 4\mathbf{j} - 3\mathbf{k}$, the dot product is computed in MATLAB as

```
>> A = [5 −8 −2];
>> B = [6 4 −3];
>> ABdot = dot(A, B)
ABdot =
         4
```

The cross product is also a built-in function in MATLAB using the command "cross(A,B)". For example, $\mathbf{A} \times \mathbf{B}$ is written as

```
>> A = [5 −8 −2];
>> B = [6 4 −3];
>> ABcross = cross(A, B)
ABcross =
   32   3   68
```

A.1.5 Complex Numbers

MATLAB stores all numbers as complex numbers and uses either i or j to represent the imaginary component of complex numbers. In MATLAB $z = 3 + i4$ and $z = 3 + j4$ are equivalent, that is

```
>> z = 3 + 4j
```

gives

```
z =
   3.0000 + 4.0000i
```

and

```
>> z = 3 + 4i
```

gives

```
z =
   3.0000 + 4.0000i
```

As mentioned previously, here we use the symbol i for current (as in Example A.1) and the symbol j for the complex numbers. MATLAB, however, defaults to the symbol i. The polar form for complex numbers is also valid in MATLAB; that is, for any arbitrary real value x and y,

$$z = x + jy = |z|e^{j\theta}$$

where

$$|z| = \sqrt{x^2 + y^2}$$
$$\theta = \tan^{-1}\frac{y}{x}$$

MATLAB has many built-in functions for working with complex numbers, with a few essential ones given in Table A.1.

TABLE A.1 MATLAB Commands for Complex Numbers with $z = x + jy$

Value	Syntax
Re(z)	real(z)
Im(z)	imag(z)
$\|z\| = \sqrt{x^2 + y^2}$	abs(z)
$\theta = \tan^{-1}\frac{y}{x}$	angle(z)

EXAMPLE PROBLEM A.2

Use MATLAB to determine $z_3 = \frac{z_1}{z_2}$ if $z_1 = 3 + j2$ and $z_2 = 2 - j$.

Solution

In MATLAB we write

```
>> z1 = 3 + j * 2
z1 =
   3.0000 + 2.0000i
>> z2 = 2-j
z2 =
   2.0000 - 1.0000i
>> z3 = z1/z2
z3 =
   0.8000 + 1.4000i
```

EXAMPLE PROBLEM A.3

Evaluate

$$z = (60 + j30) + 100e^{-j0.4887}$$

Solution

To evaluate z, we use the following MATLAB commands, mixing the rectangular and polar forms of a complex number.

```
>> z = 60 + j*30 + 100*exp(-j*0.4887)
z =
   1.4829e + 002 - 1.6948e + 001i
```

EXAMPLE PROBLEM A.4

Evaluate the magnitude and angle of the following complex number:

$$z = (5 - j4)^3$$

Solution

```
>> z = (5 - j*4)^3
z =
   -1.1500e + 002 - 2.3600e + 002i
>> r = abs(z)
r =
   262.5281
>> theta = angle(z)
theta =
   -2.0242
```

A.1.6 Polynomials and Roots

MATLAB works with polynomials using a row vector and can calculate the roots of the polynomial using a built-in function. A polynomial is written as a row vector of the polynomial coefficients in descending order, starting with the highest order. Consider the polynomial

$$x^4 - 12x^3 + 0x^2 + 25x + 116 = 0 \tag{A.12}$$

which is entered in MATLAB as

```
>> p = [1   −12   0   25   116]
```

returning

```
p =
   1   −12   0   25   116
```

It is important to enter every term, even terms with a 0 coefficient as in Equation (A.12). To find the roots of the polynomial, we use the built-in function "roots". To find the roots of Equation (A.12), we write

```
>> r = roots(p)
```

and MATLAB returns with two real roots and a pair of complex conjugate roots:

```
r =
   11.7473
   2.7028
   −1.2251 + 1.4672i
   −1.2251 − 1.4672i
```

The convention for MATLAB is that polynomials are row vectors and roots are column vectors. One can work in the opposite direction and find the polynomial given as a set of roots via the MATLAB "poly" command. Continuing from the last MATLAB calculations to find the roots of a polynomial, use the "poly" command to restore the polynomial

```
>> pp = poly(r)
```

which returns

```
pp =
   1.0000   −12.0000   −0.0000   25.0000   116.0000
```

Polynomial multiplication is carried out using the MATLAB function "conv", which performs the convolution of the two polynomials. Consider multiplying the following two polynomials using MATLAB.

$$A(x) = x^3 + 2x^2 + 3x + 4 \tag{A.13}$$

$$B(x) = x^3 + 4x^2 + 9x + 16 \tag{A.14}$$

In MATLAB we type

```
>> A = [1 2 3 4];
>> B = [1 4 9 16];
>> AB = conv(A, B)
```

which returns

```
AB =
   1   6   20   50   75   84   64
```

Note that the semicolon after the matrix A and B suppresses the echo of the command. Symbolically, the polynomial product of Equation (A.13) and Equation (A.14) is given by

$$AB(x) = x^6 + 6x^5 + 20x^4 + 50x^3 + 75x^2 + 84x + 64 \tag{A.15}$$

To add polynomials together, the polynomials need to be of the same order. For example, consider adding the following two polynomials:

$$S(x) = x^5 + 4x^4 + 5x^3 + 6x^2 + 7x + 2 \tag{A.16}$$

and

$$T(x) = 2x^4 + 9x^3 + 8x^2 - 4x + 5 \tag{A.17}$$

Since the polynomial in Equation (A.17) is 4^{th} order, we pad the row vector with a zero in place of a x^5 coefficient. Thus the polynomials are entered in MATLAB as

```
>> S = [1 4 5 6 7 2];
>> T = [0 2 9 8 −4 5];
>> U = S + T
```

which returns

```
U =
   1   6   14   14   3   7
```

or symbolically

$$x^5 + 6x^4 + 14x^3 + 14x^2 + 3x + 7 \tag{A.18}$$

To divide one polynomial by another, the MATLAB command "deconv(A,B)" is used, which represents B divided into A, or equivalently A divided by B. As there is usually a remainder from polynomial division, the syntax used is [Q, V] = deconv(A,B), where Q is the quotient polynomial and V is the remainder. Using MATLAB, the ratio of polynomials given by Equations (A.16) and (A.17), that is, $x^5 + 4x^4 + 5x^3 + 6x^2 + 7x + 2$ divided by $2x^4 + 9x^3 + 8x^2 - 4x + 5$, is written

```
>> S = [1 4 5 6 7 2];
>> T = [2 9 8 −4 5];
>> [Q, R] = deconv(S, T)
Q =
   0.5000 − 0.2500
```

```
R =
   0   0   3.2500   10.0000   3.5000   3.2500
```

Symbolically, this result is written as

$$\frac{1}{2}x - \frac{1}{4} + \frac{3.25x^3 + 10x^2 + 3.5x + 3.25}{x^5 + 2x^4 + 9x^3 + 8x^2 - 4x + 5}$$

To illustrate a simpler example with no remainder, divide $x^2 - 2x + 1$ by $x - 1$, giving

```
>> S = [1, −2, 1];
>> T = [1, −1];
>> [Q, V] = deconv(S, T)
Q =
   1 −1
V =
   0   0   0
```

The result is $x - 1$.

EXAMPLE PROBLEM A.5

Consider the following equation:

$$\frac{(1600 - 3x)^3}{27x^3} = \frac{x}{500 - x}$$

Find x given that it must be between 0 and 500.

Solution

We use "format long" in this solution so that enough significant digits are carried through. This is necessary because ordinarily MATLAB evaluates expressions with "format short". First evaluate $(1600 - 3x)^3$ as follows:

```
>> format long
>> y = [− 3 1600];
>> r = conv(y, y);
>> w = conv(y, r)
```

which returns

```
w =
   1.0e + 009*
−0.00000002700000   0.00004320000000   −0.02304000000000   4.09600000000000
```

where "w" is the expression $(1600 - 3x)^3$. Notice that "1.0e + 009" multiplies all of the coefficients in the previous polynomial.

We cross-multiply to eliminate the denominator term on the right side as follows:

```
>> z = [− 1 500];
```

and multiplying w*z is done by another "conv"

Continued

```
>> s = conv(z, w)
resulting in
s =
   1.0e + 012*

Columns 1 through 4
0.00000000002700   −0.00000005670000   0.00004464000000   −0.01561600000000
Column 5
2.04800000000000
```

The left-side denominator term $27x^3$ multiplied by the right-side numerator term x is ($27x^4$), written in MATLAB as

```
>> q = [27 0 0 0 0]

giving
q =
  27   0   0   0   0
```

Next, we subtract the two polynomials, using the MATLAB command

```
>> t = s − q
resulting in
t =
  1.0e + 012*
Columns 1 through 4
        0 − 0.00000005670000   0.00004464000000   −0.01561600000000
Column 5
2.04800000000000
```

The polynomial evaluated by the previous steps can be written as

$$-5.67 \times 10^4 x^3 + 4.464 \times 10^7 x^2 + 1.5616 \times 10^{10} x + 2.048 \times 10^{12} = 0$$

The final step in the problem is to evaluate the roots of the resulting polynomial:

```
>> v = roots(t)

which yields
v =
   1.0e + 002*
   2.62504409150663 + 2.62297313283571i
   2.62504409150663 −2.62297313283571i
   2.62292769000261
```

Two of the roots are complex and are eliminated because they fall outside the problem constraints, giving a solution of

$$x = 262.3$$

Table A.2 summarizes some MATLAB commands used for evaluating polynomials.

TABLE A.2 MATLAB Commands for Polynomials

Operation	Symbolic Expression	Syntax
Polynomial	$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$	P = [$a_n a_{n-1} \cdots a_1 a_0$]
Roots of a polynomial	$(x - r_1)(x - r_2) \cdots (x - r_n)$	r = roots(P)
Write a polynomial from the roots	$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$	poly(r)
Multiply two polynomials $S(x) = A(x) \times B(x)$	$S(x) = (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0) \times (b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0)$	S = conv(A, B)
Add two polynomials $S(x) = A(x) + B(x)$; note that the order of the polynomials must be the same	$S(x) = (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0) + (0x^n + \cdots + b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0)$	S = A + B
Divide two polynomials $S(x) = \frac{A(x)}{B(x)} = Q(x) + V(x)$;	$S(x) = \frac{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0} = q_p x^p + \cdots + q_1 x + q_0 + \frac{v_l x^l + \cdots + v_1 x + v_0}{b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0}$	[Q, V] = deconv(A, B)

A.1.7 Plotting with MATLAB

MATLAB has extensive plotting capabilities for two- and three-dimensional graphing of vectors and matrices for data visualization and presentation graphics. The appearance of the graphs, including line widths, color, annotations, and labeling can be customized to meet the needs of the user.

fplot

The basic plotting command in MATLAB is "fplot('f', [T_1, T_2])", which plots the function f over the interval T_1 to T_2. Note that the expression "f" is placed within apostrophes so that MATLAB associates this as a string expression rather than a function to be evaluated at time "t". Consider the plotting $f = 2e^{-t}$ in the interval 0 to 5 using MATLAB. We enter the following command:

```
>> fplot('2*exp(-t)', [0, 5])
```

The command "fplot" graphs the function f in the time interval from 0 to 5 in a separate window as shown in Figure A.2. To edit the properties of the plot shown in Figure A.2, select the arrow icon (next to the print icon) and then double-click on the graph. Another command window opens whereby one can change the color of the line, line styles, fonts, axes labels, axes, etc.

Graphs are easily moved from MATLAB to other applications by copying the graph to the clipboard (Edit, Copy Fig.) and then pasting it into another application such as Microsoft® Word. One has control of the graph copied to the clipboard by editing the Copy Options from the Edit menu. The "Metafile" format copies the graph without background color. The "Bitmap" option copies an exact replica of the graph including the background color.

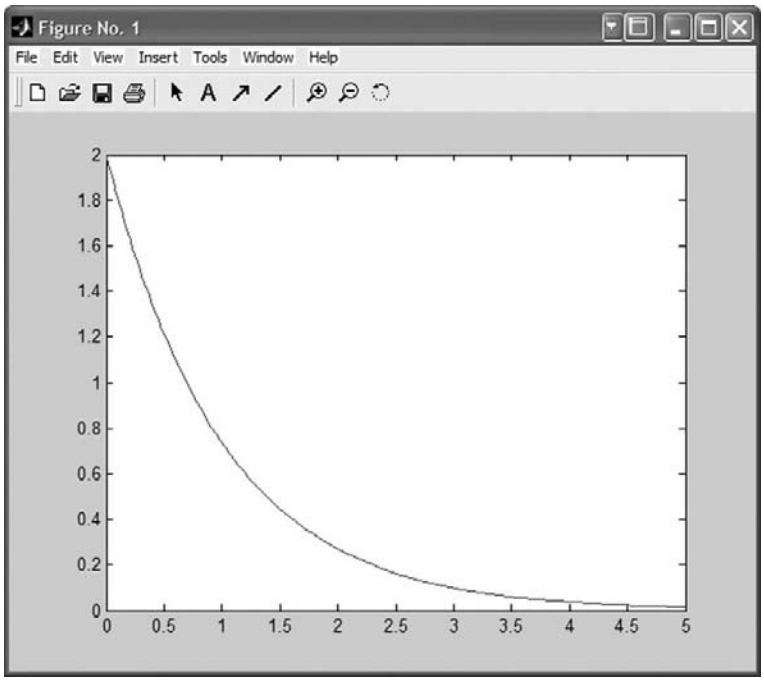


FIGURE A.2 MATLAB plot of the function $f = 2e^{-t}$ plotted using the command fplot.

plot

Another plotting command is "plot(t,y)", which plots y versus t with points connected by straight lines. The independent variable t is easily defined using the "linspace" command with the syntax "t = linspace(a,b,n)" that generates a row vector of n linearly spaced elements over the timespan a to b.

For example, we can plot the function $y = 210\sqrt{2}e^{-2t} \sin(\sqrt{2}t)$ using the following commands:

```
>> t = linspace(0, 8, 1000);
>> y = 297*sin(1.414*t).*exp (−2*t);
>> plot(t, y)
```

The first command creates a 1000-element row vector t, with values ranging from 0 to 8. The second command creates a 1000-element row vector y by multiplying sin (1.414*t) by exp (−2*t) with the ".*" operation. The ".*" operation is the dot multiplication function that performs an element-by-element multiplication. MATLAB generates an error if one uses "*" instead of ".*" in the previous calculation, because it associates t as a 1×1000 row vector, which forces sin (1.414*t) as a 1×1000 row vector and exp (−2*t) as a 1×1000 row vector. In this case, the matrix multiplication is not valid because the order of the matrices does not agree. The command "plot" graphs the function y in the time interval from 0 to 8 with 1000 connected points as shown in Figure A.3.

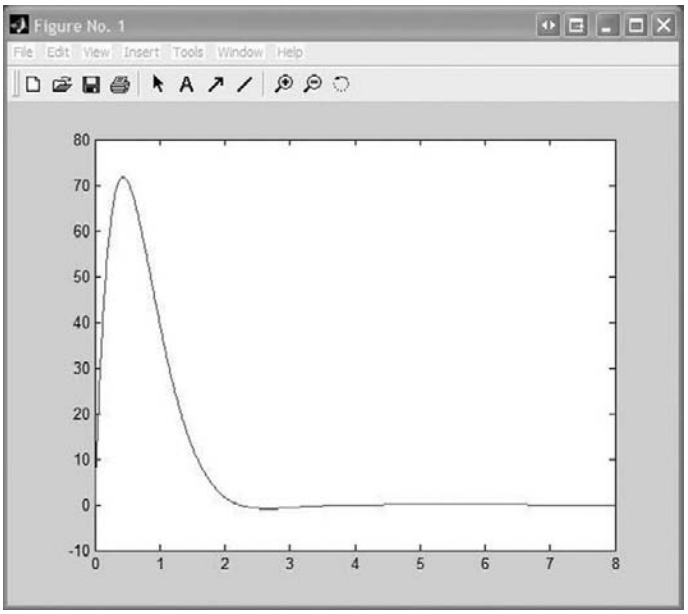


FIGURE A.3 The function $y = 210\sqrt{2}e^{-2t}\sin(\sqrt{2}t)$ graphed using the MATLAB "plot" command.

It is also possible to plot two functions on a graph using the "plot" command. To plot $f = 100e^{-t}$ and $y = 210\sqrt{2}e^{-2t}\sin(\sqrt{2}t)$ on the same graph, enter the following:

```
>> t = linspace(0, 8, 1000);
>> y = 297*sin (1.414*t).*exp(−2*t);
>> f = 100*exp(−t);
>> plot(t, y, t, f)
```

which generates the graph in Figure A.4. As before, the graph is fully customizable, with the line color, line styles, fonts, axes labels, axes, etc., changed to the user's needs.

There are many more plotting functions available within MATLAB, including three-dimensional graphs, histograms, pie charts, etc. The reader is directed to the Help function in MATLAB to discover all of the plotting capabilities.

Plotting with Microsoft Excel

While MATLAB provides very good plotting capabilities, exporting the data to other applications such as Microsoft® Excel gives graphics a more professional appearance. Before plotting the data in Excel, we need to first export the data from MATLAB and then import it into Excel. All data in MATLAB is stored under variable names in a common workspace accessible to all toolboxes, including SIMULINK. All variables used in the MATLAB workspace are given by the "who" or "whos" command. The "who" command simply lists the variables and the "whos" command lists the variables along with their sizes, number of bytes used, and class (i.e., logical character, integer array, floating point number array).

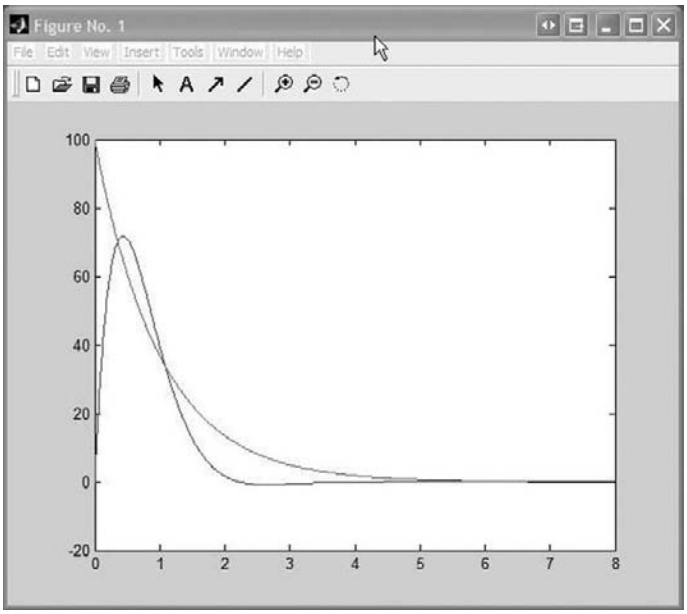


FIGURE A.4 The functions $f = 100e^{-t}$ and $y = 210\sqrt{2}e^{-2t}\sin(\sqrt{2}t)$ graphed using the MATLAB "plot" command.

Before exporting the data, it is easiest to collect the data in a single matrix with the independent variable listed first, along with the dependent variables—each evaluated at the points listed for the independent variable. Consider exporting the function shown in Figure A.3, where t and y are given by

```
>> t = linspace(0, 8, 1000);
>> y = 297*sin (1.414*t).*exp (−2*t);
```

Keep in mind that t and y are stored as 1×1000 row vectors, so we must take the transpose of each row vector when concatenating to form the new matrix *data* by

```
>> data = [t', y'];
```

Data is stored in the matrix *data* with order 1000×2. If there are additional vectors to be exported as in Figure A.4, we concatenate all the data with the command "data = [t′, y′, f′];".

To export the data to an arbitrary ASCII file "output.txt", with tabs separating each variable, we use the following command:

```
>> save output.txt data – ascii – tabs
```

This file is stored in the MATLAB subfolder "work" by default. Next, open the file "output.txt" in Excel. At this time, the data are stored in the first two columns in the workspace. From here one can use the step-by-step "chart wizard" to create a "scatter chart" like that shown in Figure A.5.

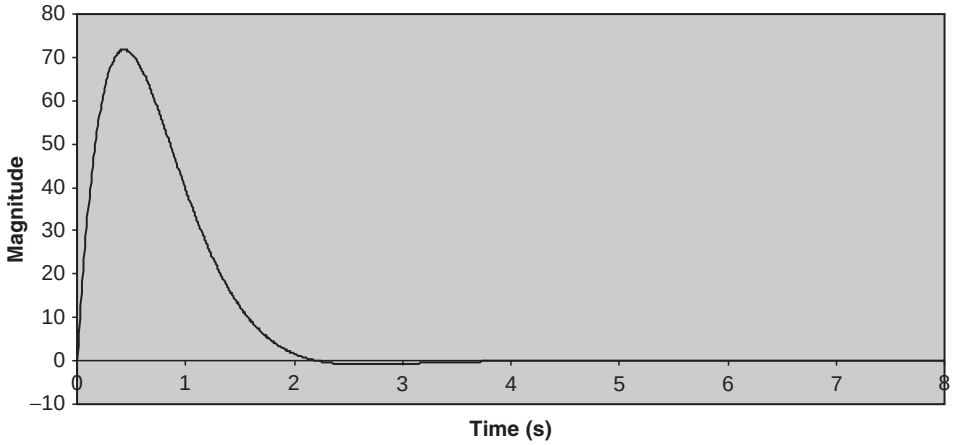


FIGURE A.5 The function $y = 210\sqrt{2}e^{-2t}\sin(\sqrt{2}t)$ drawn using Microsoft Excel.

A.1.8 Loading, Quitting and Saving the Workspace

Most problems can be solved interactively using MATLAB by entering information on the command line, pressing Enter, and having MATLAB evaluate the expression. For more challenging and lengthy projects, one can use m-files to store work to be retrieved later for further analysis. As a user becomes more proficient in MATLAB, m-files are used more frequently to store MATLAB commands, data, and input-output. The default extension for MATLAB is .mat. To create an m-file, one can save the workspace, or write the commands in Windows Notepad and load the commands in MATLAB with the load command. All of the expressions in the m-file are computed when loaded.

To open an m-file, type

>> load "filename.mat"

The command "save" is used to save the workspace to an m-file. If no filename is given, MATLAB defaults to "matlab.mat."

A.2 SOLVING DIFFERENTIAL EQUATIONS USING MATLAB

MATLAB provides the Symbolic Math Toolbox for symbolically solving many calculus and algebraic problems. The Symbolic Math Toolbox is a collection of commands that operate on functions based on a program called Maple® developed at the University of Waterloo in Ontario, Canada. Unlike previous exercises in which the answers MATLAB provided were numbers, using this toolbox results in answers that are symbolic functions. Examples of symbolic functions are

$$\tan(t^2), \quad \frac{d}{dt}(4y^2 + 2y + 4), \quad \int \sin(y)e^{-2y}dy, \quad \text{and} \quad 4\ddot{y} + 2\dot{y} + 6 = e^{-2t}.$$

MATLAB has commands among its features in this toolbox that determine the exact symbolic solution for a derivative or an integral of a function and for differential equations. For example, if we want to compute the derivative of tan (t^2), MATLAB provides the answer $2t(1 + \tan(t^2)^2)$. In this section, we introduce a few of these features, which are appropriate for solving problems in this chapter.

Before performing symbolic computations, we need to declare all of the symbolic variables. To write a symbolic variable we enclose it between single quotation marks using the "sym" MATLAB command or use the "syms" command. For example, the command

```
>> y = sym('y')
```

creates a symbolic variable "y". Instead of defining a symbolic variable using the "sym" command, we can write it directly by writing the function within single quotes as follows:

```
>> f = 'tan(t^2)'
f =
  tan(t^2)
```

which creates a symbolic variable "f" that is stored in memory as "tan (t^2)". To declare more than one symbolic variable, use the "syms" command. For example,

```
>> syms x y w z
```

or as

```
>> syms('w', 'x', 'y', 'z')
```

declares two four symbolic variables. This command is equivalent to the "sym" command for each variable separately. Whenever a symbolic variable is used after it is declared as a symbolic variable, any function using it is also a symbolic function. For example,

```
>> f = tan(y^2)
        f =
          tan (y^2)
```

Notice that in the previous MATLAB command, we did not need to write f = 'tan(t^2)' (with single quotes) because "y" is already declared as a symbolic variable. These variables are now available to be used with any MATLAB command to create symbolic results. With "w", "x", "y", and "z" declared as symbolic variables, we can create a symbolic determinant by using a symbolic matrix using the MATLAB "det" command as follows:

```
>> m = [w, x;y, z]
          m =
             [w, x]
             [y, z]
>> det(m)
          ans =
             w*z - x*y
```

The MATLAB commands "diff" and "int" are the symbolic derivative and the indefinite integral functions, respectively. The following illustrates these two symbolic functions with some common expressions.

```
>> x = diff('cos(y)')
        x =
            -sin(y)
```

and

```
>> x = int('sin(y)')
        x =
            -cos(y)
```

Sometimes the result of a MATLAB operation gives an answer that is not easily recognizable. The MATLAB command "simplify" algebraically simplifies the result to a form that is more readily recognized. For example,

```
>> simplify((x^2 + 7*x + 12)/(x + 4) )
  =
    x + 3
```

As we will see, this command is very useful during integration and differentiating operations.

MATLAB also calculates definite integrals by including the limits of integration as arguments in the "int" command. For example, to calculate $\int_0^{\pi} \cos(y)dy$, we use the following MATLAB command

```
>> x = int('sin(y)', '0', 'pi')
        x =
          2
```

The limits of integration do not have to be numbers, but can be symbolic variables. For example,

```
>> x = int('sin(y)', 'a', 'b')
        x =
          -cos(b) + cos(a)
```

The "solve" command is convenient for solving symbolic algebraic expressions. For example, to solve for the roots of the polynomial $x^4 + 14x^3 + 71x^2 + 154x + 120 = 0$, we use

```
>> solve('x^4 + 14*x^3 + 71*x^2 + 154*x + 120 = 0')
ans =
  [-5]
  [-4]
  [-3]
  [-2]
```

Recall from the previous chapter that we could calculate the roots of a polynomial using the "roots" command, that is

```
>> p = [1 14 71 154 120];
>> r = roots(p)
        r =
  −5.0000
  −4.0000
  −3.0000
  −2.0000
```

The difference between the two commands is that "solve" allows us to symbolically solve for the roots without specifying the coefficients of the polynomial such as the well-known quadratic equation:

```
>> syms a b c y x
>> solve('a*x^2 + b*x + c = 0')
ans =
  [1/2/a*(−b + (b^2 − 4*a*c)^(1/2))]
  [1/2/a*(−b − (b^2 − 4*a*c)^(1/2))]
```

In solving for the response in many dynamic systems we need to solve a system of equations for several unknown variables such as the coefficients in the natural response. Suppose we needed to solve $K_1 + K_2 = 3$ and $-5K_1 - 3K_2 = 299$, which we did using a matrix approach. We can also solve a system of equations using the "solve" command by typing

```
>> syms k1 k2
>> [k1, k2] = solve('k1 + k2 = 3', '−5*k1 −3*k2 = 299')
      k1 =
           −154
      k2 =
            157
```

There are many other syntaxes available using the "solve" command; use the MATLAB help command to learn about them or consult other references.

The MATLAB command for solving ordinary differential equations is "dsolve". The syntax involves using the capital letter "D" to denote a derivative, "D2" to denote the second derivative, "D3" to denote the third derivative, and so on. The syntax for the argument of this command involves writing the given differential equation and the initial conditions, each separated by a comma and enclosed in a single quote. For example, to solve the following differential equation for $t \geq 0$

$$\ddot{y} + 4\dot{y} + 3y = 0$$

with initial conditions $y(0) = 1$ and $\dot{y}(0) = 0$ using MATLAB's "dsolve", we have

```
>> dsolve('D2y + 4*Dy + 3*y = 0', 'Dy(0) = 0', 'y(0) = 1')
  ans =
  −1/2*exp(−3*t) + 3/2*exp(−t)
```

To plot these results, we use the MATLAB plotting function called "ezplot", an easy way to plot functions defined by symbolic functions. To plot the previous result, we execute

```
>> ezplot(y, [0, 5])
```

which gives the results shown in Figure A.6.

The argument enclosed in the brackets of the "ezplot" command is the initial and final points.

If the initial conditions are not entered in the "dsolve" command, the solution is calculated in terms of unknown coefficients. To illustrate, consider entering the command "dsolve" without initial conditions,

```
>> dsolve('D2y + 4*Dy + 3*y = 0')
```

which gives us

$$C1 * \exp(-t) + C2 * \exp(-3 * t)$$

The values for C1 and C2 are determined from the initial conditions.

To solve

$$\ddot{y} + 16,000\dot{y} + 10^8 y = 5 \times 10^8 t$$

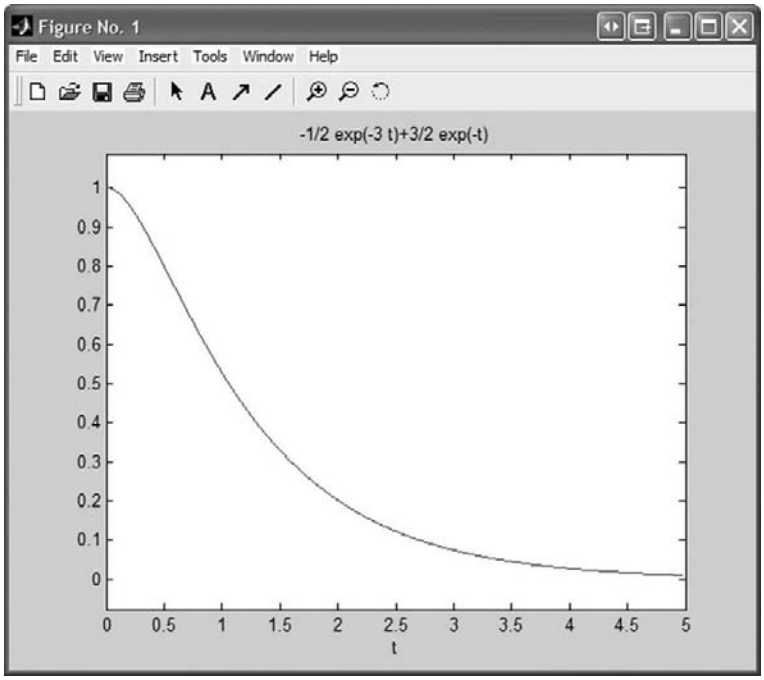


FIGURE A.6 An illustration of the MATLAB plotting function, "ezplot."

with $\dot{y}(0) = 300$ and $y(0) = -45 \times 10^{-3}$ using MATLAB, we have

```
>> dsolve('D2y + 16000*Dy + 10^8*y = 5*10^8*t', 'Dy(0) = 300', 'y(0) = -45*10^(-3)')
ans =
```

$$-1/1250 + 5*t - 293/30000 * \exp(-8000*t) * \sin(6000*t) - 221/5000 * \exp(-8000*t) * \cos(6000*t)$$

To solve a set of simultaneous differential equations, we use the command "dsolve" with the argument syntax consisting of the given differential equations and initial conditions, each separated by a comma and enclosed in a single quote. For example, suppose we wish to solve:

$$5\dot{y}_2 + 10\dot{y}_1 + 60y_1 = 300u(t)$$
$$5\dot{y}_2 + 40y_2 + 5\dot{y}_1 = 0$$

with zero initial conditions at $t = 0$. Using MATLAB, we write

```
>> [x, y] = dsolve('5*Dy + 10*Dx + 60*y = 300',
   '5*Dy + 40*y + 5*Dx = 0', 'x(0) = 0', 'y(0) = 0')
x =
   -45 + 45*exp(12/17*t)
y =
   -5*exp(12/17*t) + 5
```

On occasion, MATLAB produces incorrect or misleading results to the solution of differential equations, so care should be taken in verifying the solution by plotting, checking the initial conditions and final conditions, or simulating the solution with SIMULINK.

A.3 BLOCK DIAGRAMS AND SIMULINK

As discussed earlier, solving a differential equation or a set of simultaneous differential equations involves a considerable amount of work. In the previous section, MATLAB's symbolic math toolbox provided an easier solution method when it worked, but it sometimes failed to produce the correct solution. In this section, we use another MATLAB toolbox called SIMULINK to numerically calculate the solution for linear and nonlinear differential equations. For nonlinear differential equations, SIMULINK is usually the only solution option. To implement a numerical solution in SIMULINK, we draw the system of equations in a block diagram.

A.3.1 Block Diagrams

A block diagram is a graphical representation of the system's differential equations using basic mathematical operations such as constants, gains, summing junctions or summers, and integrators. To deal with nonlinear elements, we can create special blocks that contain mathematical functions or standard functions such as the square root, exponential, and

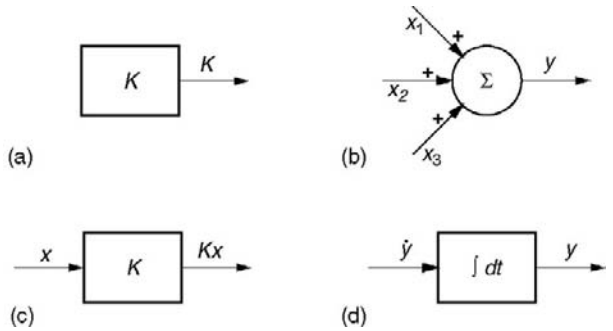


FIGURE A.7 Basic drawing elements in a block diagram: (a) constant, (b) summer, (c) gain, and (d) integrator.

logarithmic functions. The basic drawing elements for the block diagram are shown in Figure A.7.

To create a block diagram, we first take the highest derivative term output variable and put it on the left-hand side of the equal sign and the other terms on the right-hand side. Next, the block diagram is formed by connecting elements on the right-hand side to a summing junction representing the equation side, with the output being the left-hand side of the equation. From the summer, we include as many integrator blocks as necessary to reduce the highest order derivative to the output variable (i.e., $\frac{d^3y}{dt^3}$ requires three integrators). Finally, we use a gain block and connect the output variable and lower-order derivatives to appropriate gain blocks back to the summer. If there are any inputs, these are added to the summer using a constant block for a unit step function, or an appropriate function block such as t, or e^{-t}.

EXAMPLE PROBLEM A.6

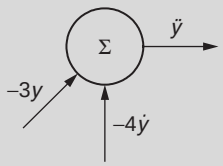
Draw a block diagram for $\ddot{y} + 4\dot{y} + 3y = 0$.

Solution

To begin the process of drawing a block diagram, the differential equation is rearranged as

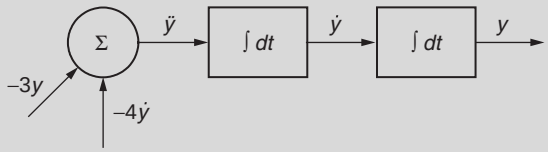
$$\ddot{y} = -4\dot{y} - 3y$$

and the block diagram is started by drawing the summer as shown in the following figure.

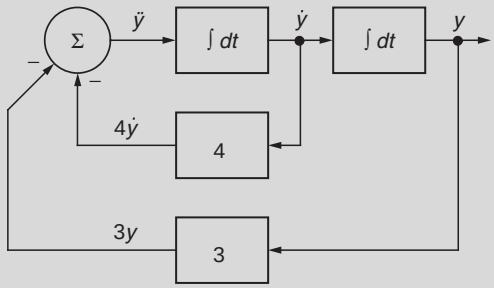


Next, we insert two integrator blocks to the output of the summer to obtain $\dot{y}$ and y as illustrated in the following diagram.

Continued



Finally, the block diagram is completed by connecting $\dot{y}$ and y to the summer via a gain block as shown in the following figure.



A block diagram can also be drawn for a set of simultaneous differential equations using the same approach as in Example A.6, drawing a block diagram for each equation and connecting the output variables and its derivatives to the appropriate summer.

EXAMPLE PROBLEM A.7

Draw a block diagram for the following set of simultaneous differential equations.

$$3\ddot{y}_1 + 10\dot{y}_1 + 60y_1 + 5\dot{y}_2 = f(t)$$
$$5y_1 + 4\ddot{y}_2 + 5\dot{y}_2 + 40y_2 = 0$$

where $f(t)$ is an unspecified input function.

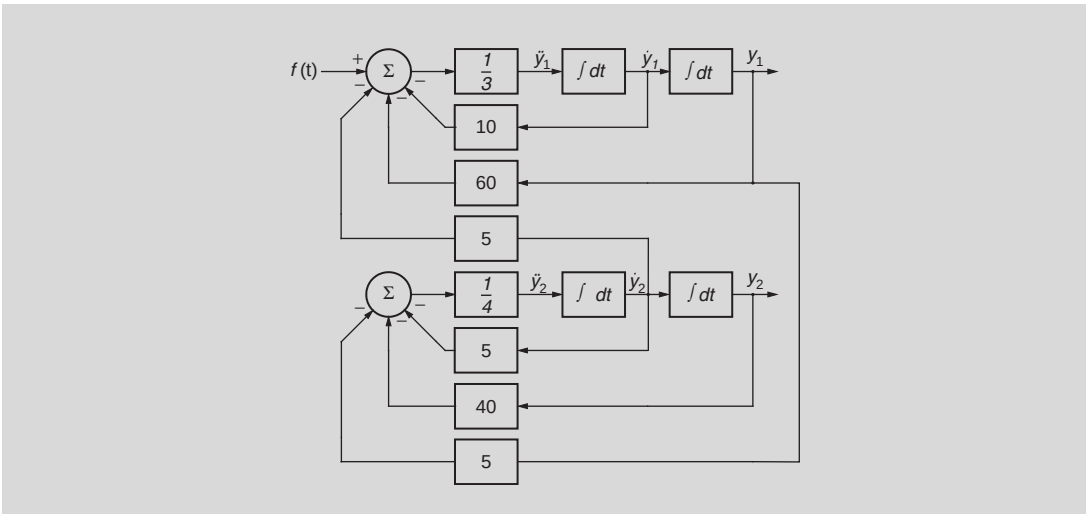
Solution

We begin by solving for the highest order derivative in each equation.

$$\ddot{y}_1 = \frac{1}{3}\left(-10\dot{y}_1 - 60y_1 - 5\dot{y}_2 + f(t)\right)$$

$$\ddot{y}_2 = \frac{1}{4}\left(-5y_1 - 5\dot{y}_2 - 40y_2\right)$$

The block diagram is drawn by first using two summers, with $\ddot{y}_1$ and $\ddot{y}_2$ as the output of each summer, followed by two integrators for each output variable. Next, we complete the diagram by including the terms on the right-hand side of the previous equations as the input to each summer with appropriate gain elements, giving us the following figure.



A.4 SIMULINK

SIMULINK is a toolbox in MATLAB that simulates linear and nonlinear continuous and discrete dynamic systems. We program a system from the block diagram as described in the previous section using SIMULINK by dragging and dropping blocks and connecting the blocks together according to the system equation or equations. After building the model we run the program to plot results or export the results to MATLAB.

To start SIMULINK, start MATLAB as before, then type "simulink" in lowercase letters in the command window. This brings us to another window containing the icons for implementing a SIMULINK simulation as shown in Figure A.8. Next, select "File" on the menu bar, then "New" and "Model". This brings up another window, a workspace, that is used to create the SIMULINK model as shown in Figure A.9. The icons shown make up the basic subsystem of blocks used to create a simulation, with each providing a set of blocks arranged into libraries. The libraries are a collection of blocks grouped into similar functions.

Double-clicking on any of the icons opens up the library of blocks. For example, double-clicking on the "Math Operations" library opens the window shown in Figure A.10 with common operations such as absolute value (Abs), gain, math function, etc. The library "Continuous" contains the blocks for the derivative and integrator among other blocks. The library "Sources" contains the blocks for a constant, clock, step, etc., all blocks that produce an output without an input. The block "clock" outputs the current simulation time as a variable, useful in decision blocks and outputting variables to MATLAB. The library "Sinks" contains the blocks associated that have an input but no output, such as the scope that displays a variable as a function of time as if on an oscilloscope. Another important block in this library is the block "To Workspace", which allows exporting a variable to MATAB for further analysis and plotting.

Blocks are copied from the "SIMULINK Library Browser" window to the workspace window in Figure A.9 by simply dragging them across from one to the other. Each block

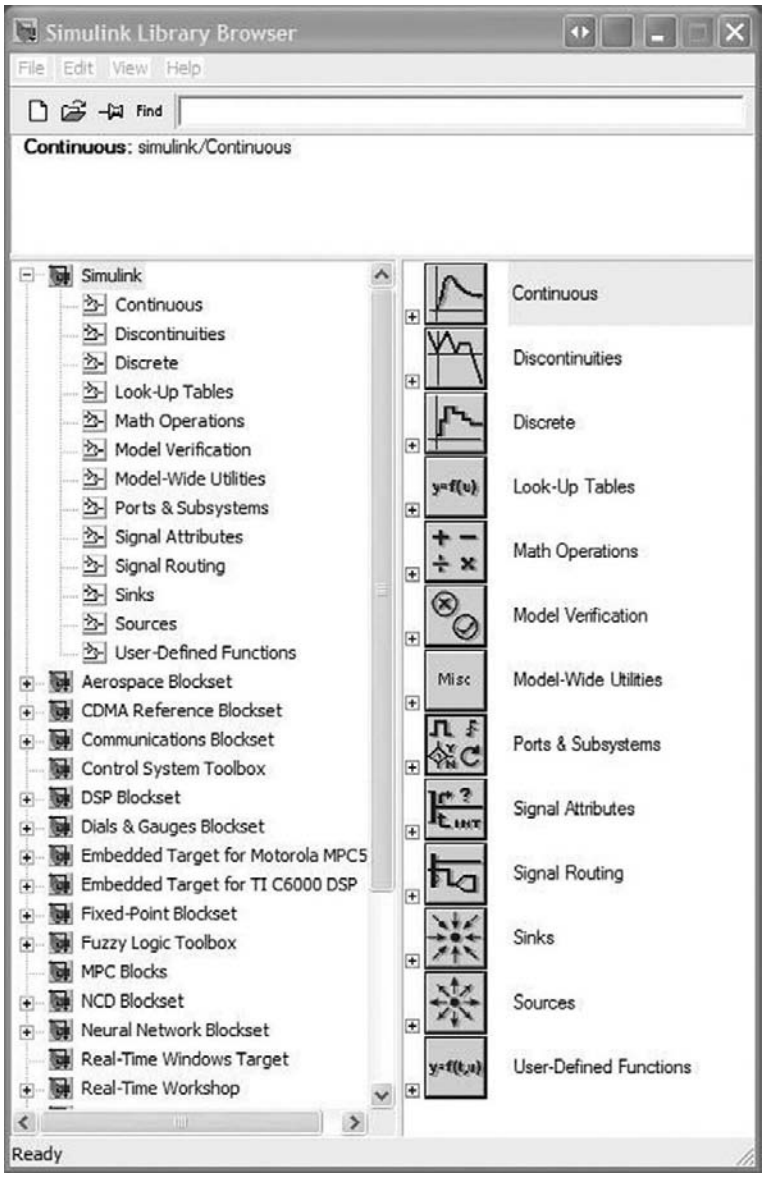


FIGURE A.8 SIMULINK Library of blocks on the right. Other MATLAB toolboxes are shown on the left.

is given a unique name below the block once it is placed in the workspace. If a block is repeated, its name is the generic block name followed by an integer. There is also a block identifier placed inside the block. The block name can be edited by placing the mouse pointer on the name, clicking the left mouse button, and then editing the name. Once in the workspace, the block's behavior is set by double-clicking the block, which opens

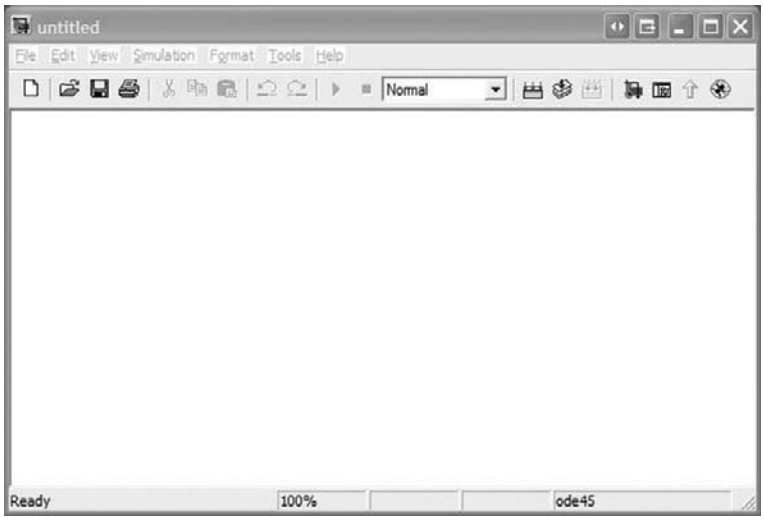


FIGURE A.9 SIMULINK workspace used to create the model.

another window where parameters are specified. The block size can be changed by selecting the block with the mouse pointer, and then, while the left mouse button is held down on the handle, drag the handle to its new size. To remove a block, simply select and delete it. For convenience, one can use the copy and paste commands to reproduce a block.

Blocks are connected together with line segments. To connect two blocks, move the mouse pointer to the angle bracket (>) on the output of one block, hold the left mouse button down, and then drag the pointer to the input < of another block. If the line segment is drawn incorrectly, use the mouse to select, and then delete it. The line segment can be manipulated by selecting it, and while the left mouse button is held down, drag the line segment to its desired shape. One can also break into a line segment by holding the mouse pointer over the line segment with the "CTRL" key pressed down, clicking the left mouse button, and dragging the pointer to the desired location.

Shown in Figure A.11 (Top) is the workspace that consists of a sine wave, integrator, and scope, connected together as the output of the sine wave to the input of the integrator to the output of the integrator to the input of the scope. Double-clicking on the sine wave reveals its properties as shown in Figure A.11 (Middle). Double-clicking on the scope block results in a graph shown in Figure A.11 (Bottom). The scope is an excellent debugging tool to verify all simulation outputs are correct, but other graphing capabilities in MATLAB and EXCEL provide far superior plots for professional reports.

Once a SIMULINK model is defined, a simulation of the model is run whereby all outputs and inputs are computed from the simulation start time to the end time for each time period. Before running the simulation, the simulation parameters need to be specified by clicking "Simulation" on the menu and then selecting the pull down "Simulation Parameters". Shown in Figure A.12 is the "Simulation Parameter" window with default parameters. The default values are usually acceptable for many simple applications. Obviously, the "Stop time" should be set to be at least 5 time constants (dominant time constant).

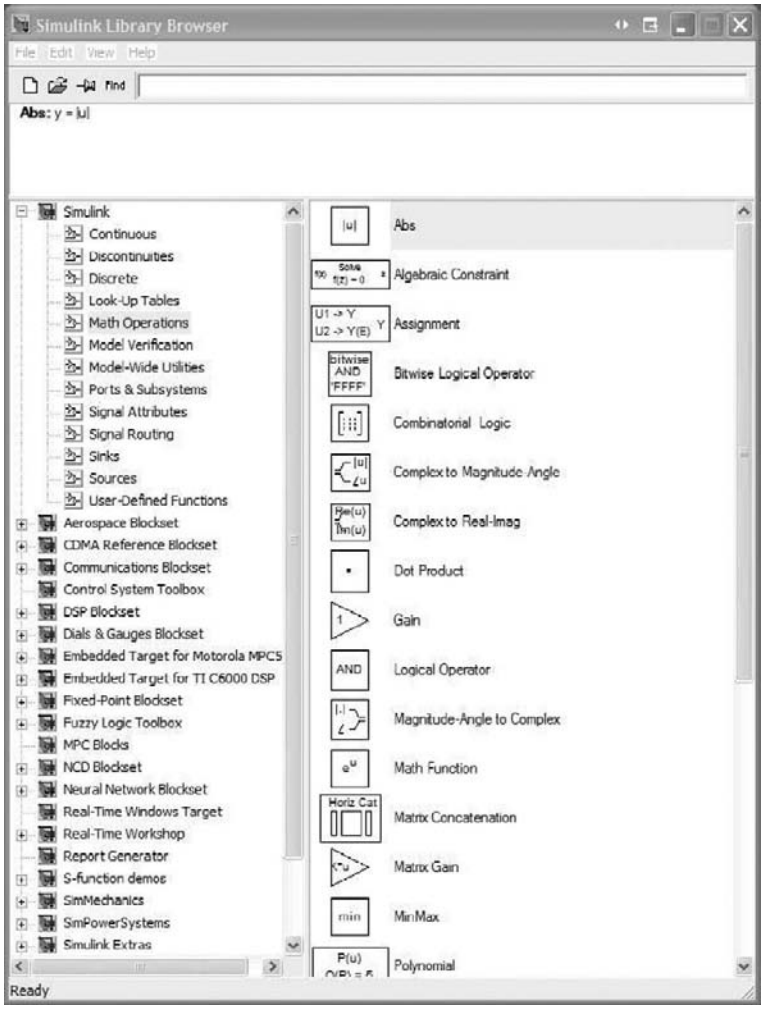
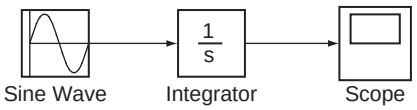


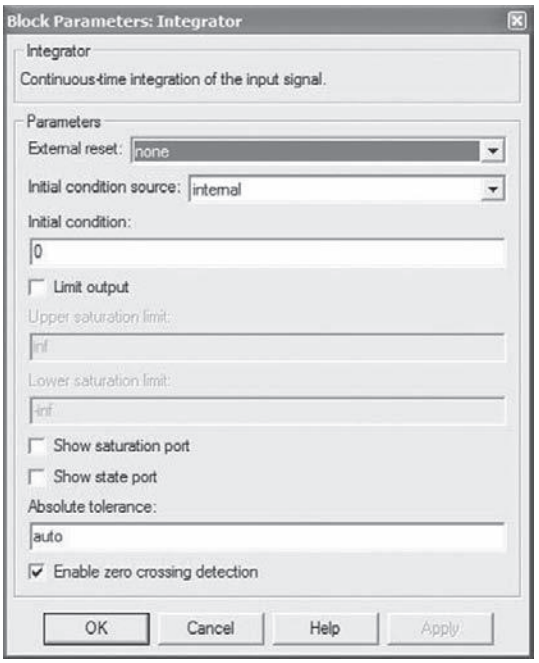
FIGURE A.10 The "Math Operations" library.

The start time can be set to a value appropriate for the simulation. The default simulation algorithm (ode45, the Dormand-Prince formula) is usually adequate for most applications. This algorithm is the best first choice as it involves calculating the output value at the next state based only on the previous output value.

Simulations can be run with either a fixed-step solver or a variable-step solver. The fixed-step solver calculates the model at equally spaced time steps, beginning at the start time. The step size must be small enough so that the approximations are valid, but not too small so that the program takes too long to run. A good rule of thumb is to start at

$$\text{step size} = \frac{\text{Total simulation time}}{500 \text{ to } 5000}$$





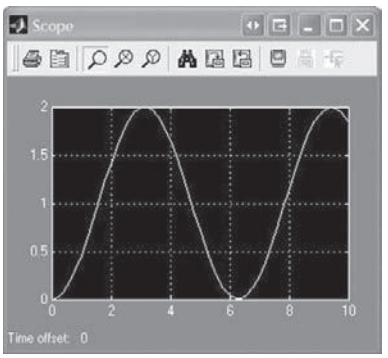


FIGURE A.11 (Top) SIMULINK blocks. (Middle) Block parameters for the integrator block. (Bottom) The output of the scope block.

and then to reduce the step size by 10. If the simulation output does not change, retain the step size, otherwise continue to reduce the step size by 10 until the output stabilizes. The variable-step size adjusts the step size, decreasing the size when the outputs are changing rapidly and increasing the step size when the outputs are changing slowly. The

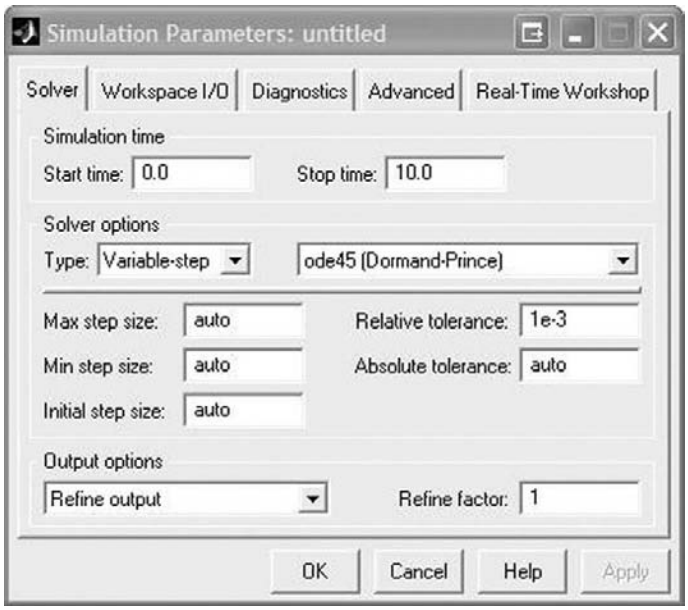


FIGURE A.12 Simulation Parameters window.

variable-step size simulation often saves simulation calculation time, resulting in a faster simulation. To run the simulation, select the command "Start" from the "Simulation" menu or the run command icon.

After creating a SIMULINK model, use "Save" or "Save As" from the "File" menu to save the model. Many simulation models have a number of parameters, some of which are changed from simulation to simulation. For convenience, it makes sense to define the parameters in MATLAB (i.e., K=5, M=3, etc.), with the parameter values saved in an m-file. The m-file can then be opened in MATLAB and these parameters can then be used in the SIMULINK model. To print the SIMULINK model, select the "Print" command from the "File" menu, or select the "Copy Model to Clipboard" command from the "Edit" menu and paste it into another application like Microsoft Word.

EXAMPLE PROBLEM A.8

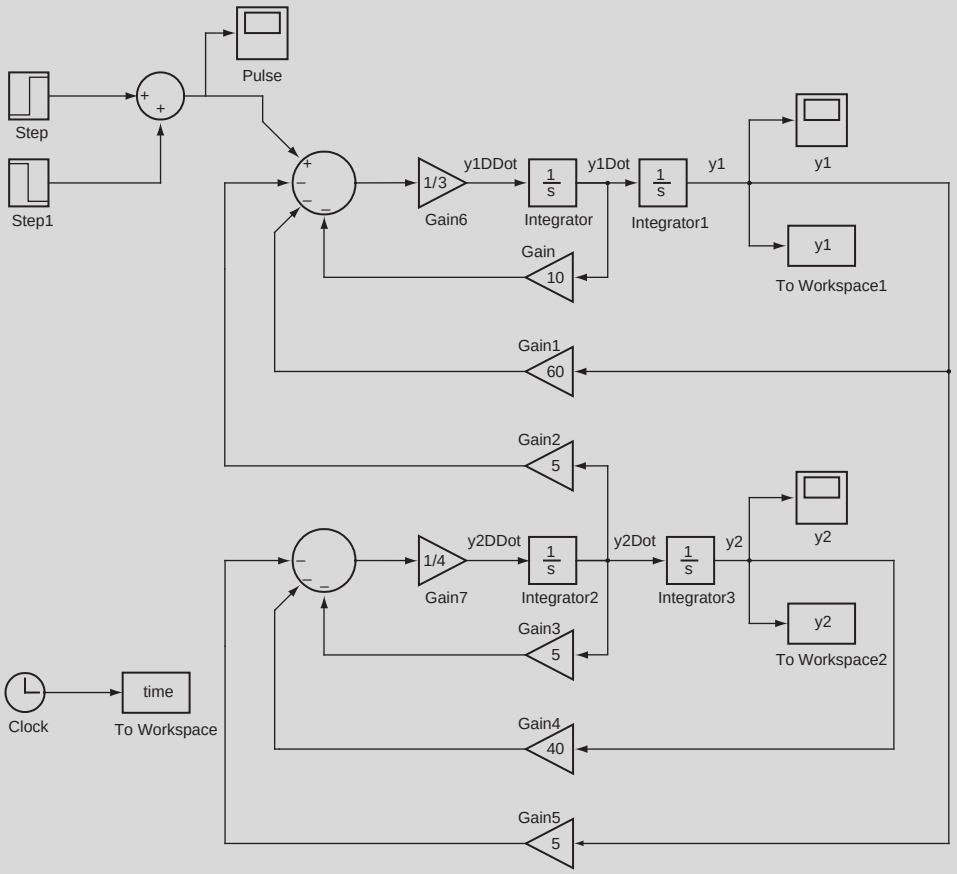
Simulate the response of the system described in Example A.7 with zero initial conditions if the input $f(t)$ is a pulse with magnitude 120 and duration 10 applied at $t = 0$.

Solution

Whenever we create a SIMULINK simulation, it is best to begin by drawing the block diagram for the system. In Example A.7, we already created the block diagram where the highest derivative was first solved in each equation and all of the other terms were placed on the right-hand side

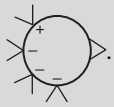
of the equation. The block diagram was then constructed using a summer for the equal sign, the integrator to remove the derivative terms, and gains for the feedback elements. The SIMULINK model is essentially a reproduction of the block diagram.

To create the SIMULINK model, we open a SIMULINK session, open a SIMULINK workspace, and then open the appropriate library to begin dragging and dropping the icons to create the model. The final SIMULINK model is shown in the following figure. Details on how to create the model are provided in the remainder of the example.



To begin drawing the model, first drag the summer, which is located in the "Math Operations" library. The summer icon looks like  . This block can be changed to a rectangle from the default circle by double-clicking the block, which opens the "Block Parameters Sum" window. Also, the number of inputs can be increased by adding additional "pluses" or "minuses" in the

Continued

list of signs. For the first equation, we have four elements connected to the summer (a positive for the input and three negative feedback elements), so the summer is modified by removing one positive sign and adding three negative signs, giving  .

The input is a pulse with magnitude 120 and duration of 10. To create a pulse, we use two step inputs from the Library "Sources". The step input block looks like 

to enter the magnitude and duration, double-click on the block to open the "Block Parameters Step" and enter for the first block 0 for "Step Time", 10 for the "Initial value", and 0 for the "Final value". The second step function uses the values 120 for the "Step Time", 0 for the "Initial value", and -120 for the "Final value". Another summer with line segments from the output of the step to the input of the summer is used to connect the two inputs. The output of the input summer is connected to the "+" input of the first summer.

After the summer, a gain of 1/3 is needed. The gain block looks like 

, which is found in the "Math Operations" library. After dragging it from the library, the magnitude is set by double-clicking and entering a value of 1/3. The number inside the gain block is the value of the gain if it can fit, otherwise "-k-" is used. A line segment is used to connect the summer to the gain.

Next, two integrators are dragged across to the workspace from the "Continuous" library. The integrator has the symbol $\frac{1}{s}$ in it that comes from the Laplace transform representation of integration. While not necessary in this example, integrators sometimes have nonzero initial conditions. To enter the initial condition for an integrator, double-click on the integrator block, which opens the "Block Parameters Integrator" window, where the initial condition can be entered. By default, the initial conditions for the integrator are zero. The output of the summer is connected with a line segment to the input of the first integrator, the output of the first integrator is connected to the input of the second integrator, and the output of the second integrator is connected to the input of the third integrator. It is generally good practice to label the output of each integrator. A label can be entered by clicking anywhere in the open workspace. A convenient label for $\ddot{y}_1$ is "y1DDot" (borrowed from the D-operator), $\dot{y}_1$ is "y1Dot", and so on.

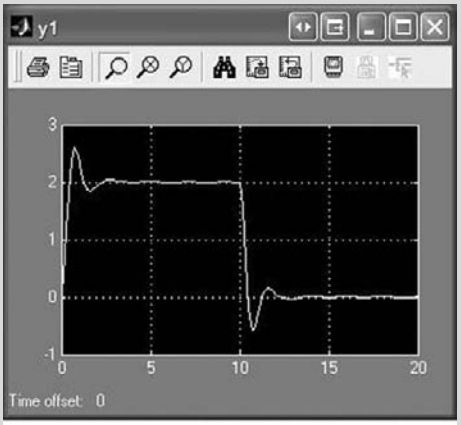
Three gain blocks are then dragged across to the workspace from the "Math Operations" library. To rotate the gain block so that it is in the correct orientation, select it and press the "CTRL and R" keys together twice for a 180° counterclockwise rotation (each click rotates 90°). The value for each gain block is entered by double-clicking and entering 10, 60, and 5. Line segments are used to connect the output of the integrators to the gain block by moving the mouse pointer over the integrator output line segment, pressing "CTRL" and left-clicking the mouse (to break into the line), and then moving the mouse pointer to the input of the gain block. The output of the gain blocks are then connected to the summer. Note that the term $\dot{y}_2$ is not connected to the gain block until the second equation is drawn.

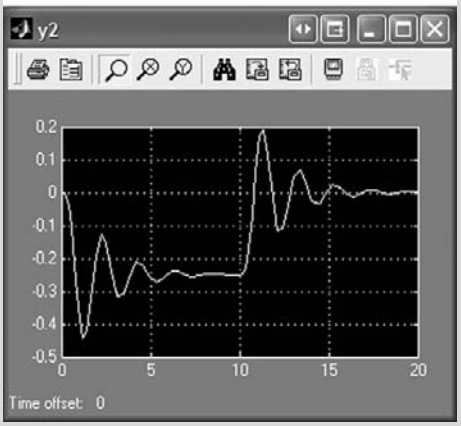
The same steps are used to create the SIMULINK model for the second equation. Notice that the input for the Gain2 block is $\dot{y}_2$, and the input for Gain5 is y_1.

To view the output, two scopes are used from the "Sinks" library. For clarity the labels on each scope are renamed to "y1" and "y2" by double-clicking on the labels below the scope and making

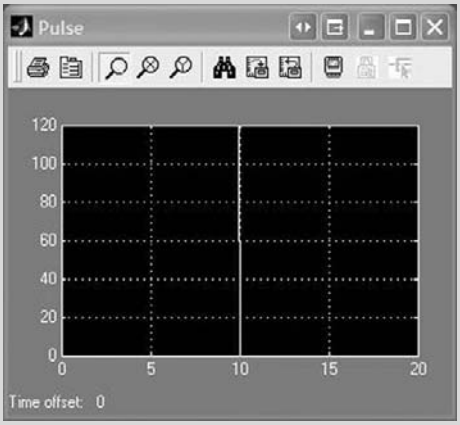
the change. Double-clicking on the scope icon opens a new window that contains a graph for the output variable. A scope is also drawn for the pulse input as good practice to ensure that the input is correctly specified. The binocular icon on the menu of the scope graph auto scales the graph.

The "Clock" block from the "Sources" library is used to output the simulation time at each simulation step. To move the simulation data from SIMULINK to the MATLAB workspace for further analysis and plotting, the "To Workspace" block is used from the "Sinks" library. Three "To Workspace" blocks are used in the example to output data for the variables y_1, y_2, and *Time*. The "To Workspace" writes the data to an array specified by the block's "Variable name". By default, the "To Workspace" uses the "Variable name" "simout" for the first, "simout2" for the second, and so on. These names are easily renamed by double-clinking on the block and entering meaningful names, such as y1, y2, and Time in this example. While the block is open, the "Save format" should be changed to "Array" and the "Sample Time" to a number that provides enough data points. In this example, the simulation time is 20, so the sample time is set at .2 to provide 100 data points for each variable in the MATLAB workspace. Once in the workspace, one can manipulate or plot the variables, i.e., using the MATLAB command "plot(Time, y1)".





Continued



The final step is to open the "Simulation Parameters" to change the simulation "Stop time" to 20 and click the run button. The preceding figure displays the output of the three scopes.

In some simulations, a model's performance is evaluated by changing parameter values. For the model in this example, the block parameters would need to be changed individually. For a model with hundreds of parameters, this would entail a considerable effort. It is far easier to enter the parameter values in MATLAB and use the parameter names in the blocks rather than values. For instance, enter the following in MATLAB

```
>> PulseGain = 120;
>> PulseDuration = 10;
>> Gain = 10;
>> Gain1 = 60;
>> Gain2 = 5;
>> Gain3 = 5;
>> Gain4 = 40;
>> Gain5 = 5;
>> Gain6 = 1/3;
>> Gain7 = 1/4;
```

Then in each gain block, use the parameter name rather than the value. Any changes in the MATLAB values are automatically updated in SIMULINK. Defining initial conditions for integrator blocks can also be done in MATLAB. To avoid entering the parameter values each time MATLAB is opened, store the values in an m-file.

It is often convenient to use subsystems to collect blocks together when drawing a SIMULINK model to make a model more readable. In the previous example, for instance, the pulse input is easily drawn as a subsystem by selecting the blocks by collecting them in a bounding box with the mouse pointer (i.e., use the mouse to draw a box around the elements), right-clicking the mouse and then selecting "Create subsystem". The inputs and the outputs to the subsystem are those of the elements in the selected blocks. Shown in Figure A.13 is part of the model from the previous example

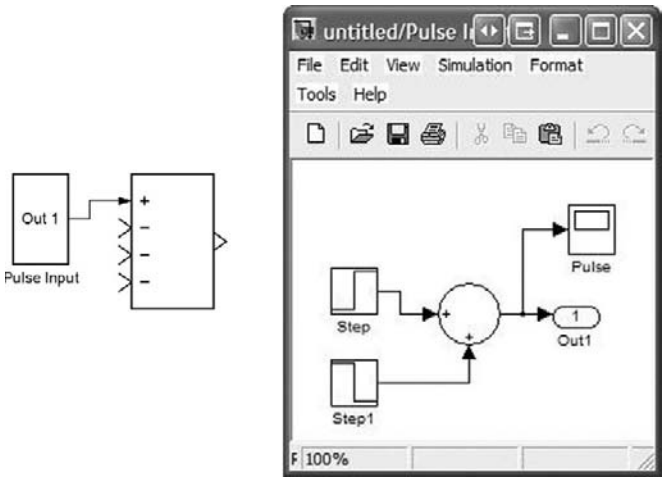


FIGURE A.13 (Left) A subsystem for the pulse input from the previous example. (Right) Double-clicking the subsystem block opens a new window that displays the blocks within the subsystem.

using a subsystem for the pulse. Double-clicking the subsystem block opens a new window with the blocks in the subsystem with the output from the subsystem indicated by the Out1 block.

Two blocks that are useful in simulations are the "Math Function" and "Fcn". The "Math Function" from the "Math Operations" library is convenient for entering mathematical functions such as the exponential, logarithm, and power functions from a drop down list. For example, when the input variable name is "u" and the function "exp", the output is "exp (u)". The label within the block is the math function. The "Fcn" from the "User-Defined Functions" library is convenient for entering general expressions. Once again, the input variable name is "u". Any arithmetic operator ($+\ -*/$) and mathematical function (cos, sin, tan, exp, power, ...) can be used in the block, as well as any parameter from MATLAB. For instance, in writing the potassium conductance for a membrane, the expression $\bar{g}_k n^4$ is needed, where $\bar{g}_k$ is a constant and n is a voltage and time dependent rate constant. In SIMULINK, this expression is written using the "Fnc" block with the parameter "Gkbar*u^4" where "Gkbar" is defined in MATLAB and "u" the input to the block representing n.

In modeling systems, some parameters may change based on threshold conditions. For example, in the saccadic eye movement system, the input to the agonist muscle, F_{ag}, is a low-pass filtered plus-step waveform given by the following equation.

$$\dot{E}_{ag} = \frac{N_{ag} - F_{ag}}{\tau_{ag}}$$

where N_{ag} is the neural control input (pulse-step waveform), and

$$\tau_{ag} = \tau_{ac}\left(u\left(t\right) - u\left(t - t_1\right)\right) + \tau_{de} u\left(t - t_1\right)$$

where τ_{ag} is the time-varying time constant made up of τ_{ac} and τ_{de}, the activation and deactivation time constants. To simulate this subsystem we use the "Switch" block in the "Signal

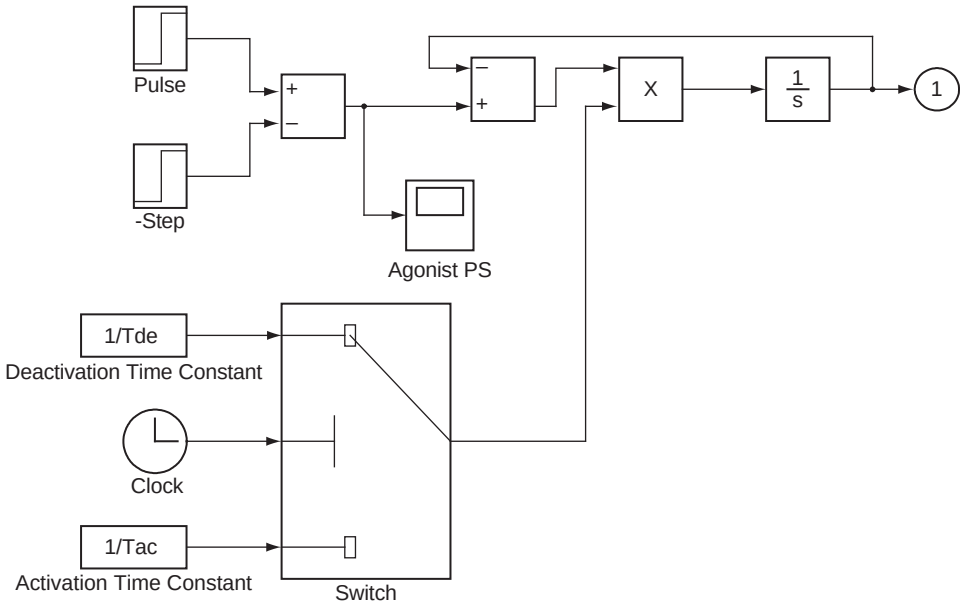


FIGURE A.14 A low-pass filter with a time-varying time constant.

Routing" library as shown in Figure A.14. The "clock" block is used to monitor the simulation time. At the start of the simulation, the time constant τ_{ac} is used. When the simulation time is greater than t_1, threshold is reached and the switch changes to the time constant τ_{de}. The remainder of the blocks in Figure A.13 solves for F_{ag}.

For information on the other operations in SIMULINK, the reader is encouraged to consult the help menu.

Index

Note: Page numbers followed by *b* indicate boxes, *f* indicate figures, and *t* indicate tables.

B

C

G

H

I

M

P

T