

Contents

Preface	<i>xv</i>
About the Companion Website	<i>xvii</i>

1	The Tree of Life (I)	<i>1</i>
1.1	Introduction	<i>1</i>
1.2	Emergence of Life	<i>1</i>
1.2.1	Timeline Disagreements	<i>3</i>
1.3	Classifications and Mechanisms	<i>4</i>
1.4	Chromatin Structure	<i>5</i>
1.5	Molecular Mechanisms	<i>9</i>
1.5.1	Precursor Messenger RNA	<i>9</i>
1.5.2	Precursor Messenger RNA to Messenger RNA	<i>10</i>
1.5.3	Classes of Introns	<i>10</i>
1.5.4	Messenger RNA	<i>10</i>
1.5.5	mRNA to Proteins	<i>11</i>
1.5.6	Transfer RNA	<i>12</i>
1.5.7	Small RNA	<i>12</i>
1.5.8	The Transcriptome	<i>13</i>
1.5.9	Gene Networks and Information Processing	<i>13</i>
1.5.10	Eukaryotic vs. Prokaryotic Regulation	<i>14</i>
1.5.11	What Is Life?	<i>14</i>
1.6	Known Species	<i>14</i>
1.7	Approaches for Compartmentalization	<i>15</i>
1.7.1	Two Main Approaches for Organism Formation	<i>16</i>
1.7.2	Size and Metabolism	<i>16</i>
1.8	Sizes in Eukaryotes	<i>16</i>
1.8.1	Sizes in Unicellular Eukaryotes	<i>17</i>
1.8.2	Sizes in Multicellular Eukaryotes	<i>17</i>
1.9	Sizes in Prokaryotes	<i>17</i>

1.10	Virus Sizes	18
1.10.1	Viruses vs. the Spark of Metabolism	20
1.11	The Diffusion Coefficient	20
1.12	The Origins of Eukaryotic Cells	21
1.12.1	Endosymbiosis Theory	21
1.12.2	DNA and Organelles	22
1.12.3	Membrane-bound Organelles with DNA	23
1.12.4	Membrane-bound Organelles Without DNA	23
1.12.5	Control and Division of Organelles	24
1.12.6	The Horizontal Gene Transfer	24
1.12.7	On the Mechanisms of Horizontal Gene Transfer	25
1.13	Origins of Eukaryotic Multicellularity	26
1.13.1	Colonies Inside an Early Unicellular Common Ancestor	26
1.13.2	Colonies of Early Unicellular Common Ancestors	26
1.13.3	Colonies of Inseparable Early Unicellular Common Ancestors	27
1.13.4	Chimerism and Mosaicism	28
1.14	Conclusions	29
2	Tree of Life: Genomes (II)	31
2.1	Introduction	31
2.2	Rules of Engagement	31
2.3	Genome Sizes in the Tree of Life	32
2.3.1	Alternative Methods	33
2.3.2	The Weaving of Scales	33
2.3.3	Computations on the Average Genome Size	36
2.3.4	Observations on Data	38
2.4	Organellar Genomes	40
2.4.1	Chloroplasts	40
2.4.2	Apicoplasts	40
2.4.3	Chromatophores	42
2.4.4	Cyanelles	42
2.4.5	Kinetoplasts	42
2.4.6	Mitochondria	43
2.5	Plasmids	43
2.6	Virus Genomes	44
2.7	Viroids and Their Implications	46
2.8	Genes vs. Proteins in the Tree of Life	47
2.9	Conclusions	49
3	Sequence Alignment (I)	51
3.1	Introduction	51
3.2	Style and Visualization	51

3.3	Initialization of the Score Matrix	54
3.4	Calculation of Scores	57
3.4.1	Initialization of the Score Matrix for Global Alignment	57
3.4.2	Initialization of the Score Matrix for Local Alignment	62
3.4.3	Optimization of the Initialization Steps	65
3.4.4	Curiosities	66
3.5	Traceback	71
3.6	Global Alignment	75
3.7	Local Alignment	79
3.8	Alignment Layout	84
3.9	Local Sequence Alignment – The Final Version	87
3.10	Complementarity	91
3.11	Conclusions	97
4	Forced Alignment (II)	99
4.1	Introduction	99
4.2	Global and Local Sequence Alignment	100
4.2.1	Short Notes	100
4.2.2	Understanding the Technology	101
4.2.3	Main Objectives	102
4.3	Experiments and Discussions	102
4.3.1	Alignment Layout	104
4.3.2	Forced Alignment Regime	105
4.3.3	Alignment Scores and Significance	105
4.3.4	Optimal Alignments	108
4.3.5	The Main Significance Scores	110
4.3.6	The Information Content	110
4.3.7	The Match Percentage	111
4.3.8	Significance vs. Chance	112
4.3.9	The Importance of Randomness	112
4.3.10	Sequence Quality and the Score Matrix	113
4.3.11	The Significance Threshold	113
4.3.12	Optimal Alignments by Numbers	114
4.3.13	Chaos Theory on Sequence Alignment	115
4.3.14	Image-Encoding Possibilities	115
4.4	Advanced Features and Methods	115
4.4.1	Sequence Detector	116
4.4.2	Parameters	116
4.4.3	Heatmap	116
4.4.4	Text Visualization	122
4.4.5	Graphics for Manuscript Figures and Didactic Presentations	123

4.4.6	Dynamics	123
4.4.7	Independence	123
4.4.8	Limits	123
4.4.9	Local Storage	124
4.5	Conclusions	127
5	Self-Sequence Alignment (I)	129
5.1	Introduction	129
5.2	True Randomness	130
5.3	Information and Compression Algorithms	130
5.4	White Noise and Biological Sequences	131
5.5	The Mathematical Model	131
5.5.1	A Concrete Example	132
5.5.2	Model Dissection	133
5.5.3	Conditions for Maxima and Minima	136
5.6	Noise vs. Redundancy	137
5.7	Global and Local Information Content	137
5.8	Signal Sensitivity	138
5.9	Implementation	140
5.9.1	Global Self-Sequence Alignment	140
5.9.2	Local Self-Sequence Alignment	144
5.10	A Complete Scanner for Information Content	147
5.11	Conclusions	149
6	Frequencies and Percentages (II)	151
6.1	Introduction	151
6.2	Base Composition	152
6.3	Percentage of Nucleotide Combinations	152
6.4	Implementation	153
6.5	A Frequency Scanner	156
6.6	Examples of Known Significance	158
6.7	Observation vs. Expectation	160
6.8	A Frequency Scanner with a Threshold	161
6.9	Conclusions	163
7	Objective Digital Stains (III)	165
7.1	Introduction	165
7.2	Information and Frequency	166
7.3	The Objective Digital Stain	170
7.3.1	A 3D Representation Over a 2D Plane	173
7.3.2	ODSs Relative to the Background	177

7.4	Interpretation of ODSs	181
7.5	The Significance of the Areas in the ODS	183
7.6	Discussions	184
7.6.1	A Similarity Between Dissimilar Sequences	186
7.7	Conclusions	186
8	Detection of Motifs (I)	187
8.1	Introduction	187
8.2	DNA Motifs	187
8.2.1	DNA-binding Proteins vs. Motifs and Degeneracy	188
8.2.2	Concrete Examples of DNA Motifs	188
8.3	Major Functions of DNA Motifs	191
8.3.1	RNA Splicing and DNA Motifs	191
8.4	Conclusions	195
9	Representation of Motifs (II)	197
9.1	Introduction	197
9.2	The Training Data	197
9.3	A Visualization Function	198
9.4	The Alignment Matrix	201
9.5	Alphabet Detection	203
9.6	The Position-Specific Scoring Matrix (PSSM) Initialization	206
9.7	The Position Frequency Matrix (PFM)	207
9.8	The Position Probability Matrix (PPM)	208
9.8.1	A Kind of PPM Pseudo-Scanner	209
9.9	The Position Weight Matrix (PWM)	212
9.10	The Background Model	215
9.11	The Consensus Sequence	218
9.11.1	The Consensus – Not Necessarily Functional	219
9.12	Mutational Intolerance	221
9.13	From Motifs to PWMs	222
9.14	Pseudo-Counts and Negative Infinity	226
9.15	Conclusions	229
10	The Motif Scanner (III)	231
10.1	Introduction	231
10.2	Looking for Signals	232
10.3	A Functional Scanner	235
10.4	The Meaning of Scores	239
10.4.1	A Score Value Above Zero	239
10.4.2	A Score Value Below Zero	241

- 10.4.3 A Score Value of Zero 241
- 10.5 Conclusions 242

- 11 Understanding the Parameters (IV) 243**
 - 11.1 Introduction 243
 - 11.2 Experimentation 243
 - 11.2.1 A Scanner Implementation Based on Pseudo-Counts 244
 - 11.2.2 A Scanner Implementation Based on Propagation of Zero Counts 246
 - 11.3 Signal Discrimination 249
 - 11.4 False-Positive Results 250
 - 11.5 Sensitivity Adjustments 251
 - 11.6 Beyond Bioinformatics 252
 - 11.7 A Scanner That Uses a Known PWM 253
 - 11.8 Signal Thresholds 256
 - 11.8.1 Implementation and Filter Testing 258
 - 11.9 Conclusions 262

- 12 Dynamic Backgrounds (V) 263**
 - 12.1 Introduction 263
 - 12.2 Toward a Scanner with Two PFMs 263
 - 12.2.1 The Implementation of Dynamic PWMs 264
 - 12.2.2 Issues and Corrections for Dynamic PWMs 271
 - 12.2.3 Solutions for Aberrant Positive Likelihood Values 274
 - 12.3 A Scanner with Two PFMs 279
 - 12.4 Information and Background Frequencies on Score Values 283
 - 12.5 Dynamic Background vs. Null Model 284
 - 12.6 Conclusions 286

- 13 Markov Chains: The Machine (I) 287**
 - 13.1 Introduction 287
 - 13.2 Transition Matrices 287
 - 13.3 Discrete Probability Detector 291
 - 13.3.1 Alphabet Detection 292
 - 13.3.2 Matrix Initialization 293
 - 13.3.3 Frequency Detection 295
 - 13.3.4 Calculation of Transition Probabilities 297
 - 13.3.5 Particularities in Calculating the Transition Probabilities 306
 - 13.4 Markov Chains Generators 307
 - 13.4.1 The Experiment 308
 - 13.4.2 The Implementation 312
 - 13.4.3 Simulation of Transition Probabilities 315

- 13.4.4 The Markov machine 315
- 13.4.5 Result Verification 317
- 13.5 Conclusions 318

- 14 Markov Chains: Log Likelihood (II) 319**
 - 14.1 Introduction 319
 - 14.2 The Log-Likelihood Matrix 319
 - 14.2.1 A Log-Likelihood Matrix Based on the Null Model 320
 - 14.2.2 A Log-Likelihood Matrix Based on Two Models 322
 - 14.3 Interpretation and Use of the Log-Likelihood Matrix 326
 - 14.4 Construction of a Markov Scanner 328
 - 14.5 A Scanner That Uses a Known LLM 337
 - 14.6 The Meaning of Scores 340
 - 14.7 Beyond Bioinformatics 344
 - 14.8 Conclusions 345

- 15 Spectral Forecast (I) 347**
 - 15.1 Introduction 347
 - 15.2 The Spectral Forecast Model 347
 - 15.3 The Spectral Forecast Equation 349
 - 15.4 The Spectral Forecast Inner Workings 350
 - 15.4.1 Each Part on a Single Matrix 351
 - 15.4.2 Both Parts on a Single Matrix 352
 - 15.4.3 Both Parts on Separate Matrices 353
 - 15.4.4 Concrete Example 1 354
 - 15.4.5 Concrete Example 2 357
 - 15.4.6 Concrete Example 3 359
 - 15.5 Implementations 360
 - 15.5.1 Spectral Forecast for Signals 362
 - 15.5.2 What Does the Value of d Mean? 364
 - 15.5.3 Spectral Forecast for Matrices 368
 - 15.6 The Spectral Forecast Model for Predictions 372
 - 15.6.1 The Spectral Forecast Model for Signals 372
 - 15.6.2 Experiments on the Similarity Index Values 381
 - 15.6.3 The Spectral Forecast Model for Matrices 384
 - 15.7 Conclusions 389

- 16 Entropy vs. Content (I) 391**
 - 16.1 Introduction 391
 - 16.2 Information Entropy 391
 - 16.3 Implementation 395

16.4	Information Content vs. Information Entropy	400
16.4.1	Implementation	403
16.4.2	Additional Considerations	409
16.5	Conclusions	409

17 Philosophical Transactions 411

17.1	Introduction	411
17.2	The Frame of Reference	411
17.2.1	The Fundamental Layer of Complexity	412
17.2.2	On the Complexity of Life	414
17.3	Random vs. Pseudo-random	415
17.4	Random Numbers and Noise	418
17.5	Determinism and Chaos	419
17.5.1	Chaos Without Noise	420
17.5.2	Chaos with Noise	427
17.5.3	Limits of Prediction	430
17.5.4	On the Wings of Chaos	431
17.6	Free Will and Determinism	431
17.6.1	The Greatest Disappointment	432
17.6.2	The Most Powerful Processor in Existence	433
17.6.3	Certainty vs. Interpretation	435
17.6.4	A Wisdom that Applies	436
17.7	Conclusions	439

Appendix A Appendix A 441

A.1	Association of Numerical Values with Letters	441
A.2	Sorting Values on Columns	443
A.3	The Implementation of a Sequence Logo	446
A.4	Sequence Logos Based on Maximum Values	451
A.5	Using Logarithms to Build Sequence Logos	455
A.6	From a Motif Set to a Sequence Logo	459

References 467

Index 489