

Tandy Warnow
Editor

Bioinformatics and Phylogenetics

Seminal Contributions of Bernard Moret

 Springer

Editor

Tandy Warnow
University of Illinois at Urbana-Champaign
Urbana, IL, USA

ISSN 1568-2684

ISSN 2662-2432 (electronic)

Computational Biology

ISBN 978-3-030-10836-6

ISBN 978-3-030-10837-3 (eBook)

<https://doi.org/10.1007/978-3-030-10837-3>

Library of Congress Control Number: 2018966860

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*Dedicated to Bernard M. E. Moret on his
retirement.*

Preface

This Festschrift is in honor of Bernard M. E. Moret, whose retirement from Ecole Polytechnique Fédérale de Lausanne (EPFL) in December 2016 culminated a nearly 40-year career. Bernard's research spanned several areas in computer science, including algorithm engineering, high-performance computing, and algorithmic computational biology. Many of Bernard's contributions were concerned with inferring and using phylogenies (i.e., evolutionary trees and phylogenetic networks), especially for large and challenging datasets. His work in genome rearrangement phylogeny is the best known, where he has been one of a small number of leading researchers in establishing theory and developing novel methods and open-source software based on rigorous mathematical theory. In addition, he has also contributed to the understanding and developing divide-and-conquer strategies, supertree construction, absolute fast converging phylogeny estimation, phylogenetic networks, and the use of phylogenies to answer biological questions. He has also trained (both directly and indirectly) many of the leading people in computational biology, including several of the contributors to this volume. For example, Alexandros Stamatakis and Daniel Doerr were postdoctoral researchers of Bernard's; Jijun Tang was his Ph.D. student; and Mark Holder, Luay Nakhleh, and Sébastien Roch were funded by the CIPRES project (see <http://www.phylo.org>), which Moret directed from 2003–2006.

Two conferences were organized in honor of Bernard's retirement: the first, organized by his current and former students and postdocs, was held at EPFL on November 7–8, 2016 (see <https://lcbp.epfl.ch/climb/>), and the second, which I organized, was held at Berkeley on June 2, 2017 (see <http://tandy.cs.illinois.edu/Moret-Festschrift.html>). These two conferences led to this Festschrift.

The chapters in this volume represent some of the areas in which Bernard's work has had an influence, including methods for phylogenetic tree and network estimation, genome rearrangements, cancer phylogeny, species trees, divide-and-conquer strategies, and the use of integer linear programming in computational biology. Each chapter provides an introduction to a cutting-edge problem in computational biology that is computationally and mathematically interesting. Hence, the volume is designed to be useful as a text for a graduate course in

computational biology and bioinformatics, aiming at computer scientists, applied mathematicians, and statisticians. Although much of the work that is described is advanced, each chapter provides references to the literature to enable the reader to obtain additional background, if needed.

The chapters are ordered based on the statistical complexity of the problem they address. The first three chapters address challenges in developing accurate and efficient software for the NP-hard maximum likelihood phylogeny estimation problem. Chapters 1 (by Alexandros Stamatakis) and 2 (by Stéphane Guindon and Olivier Gascuel) focus on optimizing maximum likelihood codes (including but not limited to numerical optimization), while Chap. 3 (by David Bader and Kamesh Madduri) focuses on high-performance computing aspects; together, the three chapters provide complementary insights into how to design codes for this important basic problem in computational phylogenetics. Chapter 4 (by Sébastien Roch) is also about phylogeny estimation from aligned sequences, and addresses a basic statistical question: how much data does a phylogeny estimation method require (or need) to recover the true tree with high probability, as a function of the model tree parameters (e.g., number of leaves and lengths of the branches in the tree)? Roch's chapter comes with a Jupyter notebook and provides scripts for analyses and simulations. Chapter 5 (by Nadia El-Mabrouk and Emmanuel Noutahi) addresses algorithms to infer gene trees when the input includes aligned gene sequences and also a species tree. The inference of species trees is covered in Chaps. 6 and 7. Chapter 6 (which I contributed) is about divide-and-conquer strategies to scale phylogeny estimation methods to large datasets, and specifically addresses limitations in current supertree methods. Chapter 7 (by Benjamin Redelings and Mark Holder) is about taxonomic supertrees and the challenge of constructing them when some taxa in the input have unknown placements within a taxonomic hierarchy. Chapter 8 (by Santos Muñoz et al.) addresses the inference of ultrametric distances from additive distance matrices, and so is related to the problem of assigning dates to internal nodes in phylogenetic trees. Chapters 9 (by Jijun Tang) addresses the inference of ancestral genomes under genome rearrangement events. Chapters 10 (by Ron Zeira and Ron Shamir) and 11 (by Russell Schwartz) address complementary approaches to inferring evolutionary histories in cancer; Zeira and Shamir focus on how chromosomal rearrangements can be used as indicators of evolutionary history and Schwartz provides an overview of different approaches to tumor phylogenetics. Chapters 12 (by Louxin Zhang) and 13 (by R. A. Leo Elworth et al.) examine problems in phylogenetic networks, with Zhang focusing on discrete mathematics questions and Elworth et al. addressing statistical estimation issues. Chapter 14 (by Daniel Doerr and Jens Stoye) uses evolution to provide a framework within which to understand comparative and functional genomics. Chapter 15 (by Dan Gusfield) provides an introduction to Integer Linear Programming and its use in computational biology, with specific focus on how ILP can be used to solve the NP-hard Traveling Salesman Problem; his chapter also comes with software and datasets.

While the chapters are designed to be self-contained (and each contains a substantial bibliography to enable the reader to get additional background), some background in computer science (algorithms and running time analysis) and statistics (e.g., the use of probabilistic models and statistical estimation under these models) is assumed. Background in computational phylogenetics is helpful but not required, but many readers may find it helpful to read some of the textbooks in phylogenetics. For a statistical perspective on this research area, see [1] and [3]. A computer science perspective on algorithm design for phylogeny estimation (especially for large datasets) is provided in [2].

Urbana, USA

Tandy Warnow

References

1. Felsenstein, J.: *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts (2004)
2. Warnow, T.: *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, Cambridge, UK (2018)
3. Yang, Z.: *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford (2014)

Contents

1	A Review of Approaches for Optimizing Phylogenetic Likelihood Calculations	1
	Alexandros Stamatakis	
1.1	Introduction	1
1.2	Calculating the Likelihood on Phylogenies	2
1.3	Sequential PLF Optimization via Algorithmic Means	4
	1.3.1 Saving Computations	5
	1.3.2 Saving Memory	7
1.4	Sequential Optimization via Technical Means	8
	1.4.1 Standard Techniques: Tip–Tip and Tip–Inner Optimizations	9
	1.4.2 Vectorization	9
1.5	Partial and Full Terraces in Tree Space	10
1.6	Parallel PLF Computations	12
	1.6.1 Preprocessing and Parallel I/O	13
	1.6.2 Parallelization Approaches	13
	1.6.3 Data Distribution Algorithms	14
1.7	Open Problems and Future Challenges	16
	References	17
2	Numerical Optimization Techniques in Maximum Likelihood Tree Inference	21
	Stéphane Guindon and Olivier Gascuel	
2.1	Introduction	21
2.2	Modeling Sequence Evolution	23
2.3	Matrix-Based Calculation of the Likelihood Function	26
2.4	Likelihood Calculation and Pruning Algorithm Using Vector Operations Only	27
2.5	Inferring Edge Lengths	29

- 2.5.1 Speeding Up the Likelihood Calculation 29
- 2.5.2 Optimizing One Length 30
- 2.5.3 Optimizing All Lengths 32
- 2.6 Inferring Parameters of Mixture Models 34
- 2.7 Conclusion 36
- References 37
- 3 High-Performance Phylogenetic Inference 39**
- David A. Bader and Kamesh Madduri
- 3.1 Introduction 39
- 3.2 Faster Likelihood Calculations 40
- 3.3 Performance Optimizations and Multi-node Parallelism 41
- 3.4 Conclusions 42
- References 42
- 4 Hands-on Introduction to Sequence-Length Requirements in
Phylogenetics 47**
- Sébastien Roch
- 4.1 Introduction 47
- 4.2 Definitions 48
- 4.3 A Simple Setting 49
- 4.4 Phylogenetic Signal 52
- 4.4.1 Short Branches 54
- 4.4.2 Depth 55
- 4.5 Not All Reconstruction Methods are Created Equal 56
- 4.6 What About Maximum Likelihood Estimation? 59
- 4.6.1 Likelihood Ratio Test 60
- 4.6.2 Optimizing the Branch Lengths 64
- 4.7 Lower Bound on the Best Achievable Requirement 68
- 4.8 Scaling Up to Large Trees 73
- 4.8.1 Signal Decay 74
- 4.8.2 Depth v. Branching 77
- 4.9 Bibliographic Remarks 84
- References 84
- 5 Gene Family Evolution—An Algorithmic Framework 87**
- Nadia El-Mabrouk and Emmanuel Noutahi
- 5.1 Introduction 88
- 5.2 Trees 90
- 5.3 Reconciliation of a Binary Gene Tree with a Binary Species
Tree 91
- 5.3.1 DL Reconciliation 93
- 5.3.2 DTL Reconciliation 94
- 5.3.3 Binary Gene Tree Reconciliation in Presence of
ILS 96

- 5.4 Reconciliation with a Non-binary Species Tree 99
- 5.5 Reconciliation of a Non-binary Gene Tree with a Binary Species Tree 100
 - 5.5.1 PolytomySolver 101
 - 5.5.2 Extensions to DTL Reconciliation 104
- 5.6 Inferring a Gene Tree from a Set of Trees 104
 - 5.6.1 Amalgamation: Gene Tree Inference from a Set of Clades 105
 - 5.6.2 Supertree: Inferring a Tree from a Set of Subtrees 106
- 5.7 A Unifying View for the DL Model 109
- 5.8 Discussion 113
- References 115
- 6 Divide-and-Conquer Tree Estimation: Opportunities and Challenges 121**

Tandy Warnow

 - 6.1 Introduction 121
 - 6.2 Background 126
 - 6.2.1 Terminology 126
 - 6.2.2 Representations of Trees 127
 - 6.2.3 Bipartition-Based Supertree Methods 128
 - 6.2.4 Quartet-Based Supertree Methods 131
 - 6.2.5 Distance-Based Supertree Methods 132
 - 6.3 Accuracy and Scalability of Existing Supertree Methods 133
 - 6.4 Improving Scalability of Supertree Methods 134
 - 6.4.1 SuperFine: Boosting Supertree Methods 135
 - 6.4.2 Explicitly Constraining the Search Space 137
 - 6.5 Relationship to Phylogenomic Species Tree Estimation 139
 - 6.6 Further Reading 141
 - 6.7 Concluding Remarks 142
 - References 143
- 7 Taxonomic Supertree Construction with *Incertae sedis* Taxa 151**

Benjamin D. Redelings and Mark T. Holder

 - 7.1 Introduction 151
 - 7.1.1 Background 154
 - 7.1.2 A Naive Semantics: Consequences of Ignoring the Annotation 156
 - 7.2 Semantics of *Incertae sedis* Taxa 157
 - 7.2.1 Goals for an *Incertae sedis* Semantics 157
 - 7.2.2 Terminology for Rooted Splits 158
 - 7.2.3 Split-Based Semantics for *Incertae sedis* Taxa 159
 - 7.2.4 Unrestricted Range and Ignoring Additional Information 160

7.2.5	Naming	161
7.2.6	Taxonomic Revision	164
7.2.7	Ambiguity of Split Representation for <i>Incertae sedis</i> Taxa	166
7.3	Handling <i>Incertae sedis</i> Taxa in a Software Pipeline	168
7.3.1	Subproblem Decomposition	169
7.3.2	Subproblem Solution	170
7.4	Discussion	171
	References	172
8	Evolutionary Rate Change and the Transformation from Additive to Ultrametric: Modal Similarity of Orthologs in Fish and Flower Phylogenomics	175
	Daniella Santos Muñoz, Eric Lam and David Sankoff	
8.1	Introduction	175
8.2	Approaches to Transformation	177
8.2.1	Farris Transform	177
8.2.2	Nonparametric Rate Smoothing (NPRS)	178
8.2.3	Penalized Likelihood	179
8.3	Pipeline	180
8.4	Applications	181
8.4.1	Fish	181
8.4.2	Solanaceae	185
8.4.3	Malvaceae	185
8.5	Discussion	188
	References	188
9	Ancestral Genome Reconstruction	193
	Jijun Tang	
9.1	Introduction	193
9.2	Definitions	194
9.3	Pairwise Distance and Sorting	196
9.4	Solving the Median Problem	197
9.4.1	Computing with Multiple Genomes	199
9.5	Conclusions	200
	References	201
10	Genome Rearrangement Problems with Single and Multiple Gene Copies: A Review	205
	Ron Zeira and Ron Shamir	
10.1	Introduction	206
10.1.1	Genomes and Rearrangements	206
10.1.2	Genome Rearrangements in Species Evolution	208
10.1.3	Genome Rearrangements in Cancer	210

- 10.2 Single-Gene Models, Operation Types, and Distance Measures 212
 - 10.2.1 Genome Representation 212
 - 10.2.2 Breakpoint Distance 218
 - 10.2.3 Reversal and Translocation Distances 219
 - 10.2.4 DCJ Distance 220
 - 10.2.5 SCoJ Distance 221
- 10.3 Multi-copy Models in Species Evolution 221
 - 10.3.1 Polyploidy 222
 - 10.3.2 Single-Copy Models with Indels 223
 - 10.3.3 Multi-copy Models Without Duplications/Deletions 224
 - 10.3.4 Models with Duplications or Deletions 225
- 10.4 Multi-copy Models in Cancer 226
 - 10.4.1 Models with Duplications/Deletions 228
 - 10.4.2 Copy Number Profile Distances 228
 - 10.4.3 Other Cancer Models 231
- References 234
- 11 Computational Models for Cancer Phylogenetics 243**
 - Russell Schwartz
 - 11.1 Introduction 243
 - 11.1.1 Background 245
 - 11.1.2 Scope and Organization 247
 - 11.2 Estimating Evolutionary Distance 248
 - 11.2.1 Single Nucleotide Variants (SNVs) 249
 - 11.2.2 Copy Number Aberrations (CNAs) 250
 - 11.2.3 Other Data Types 255
 - 11.3 Median Nodes 257
 - 11.3.1 SNV Median Nodes 257
 - 11.3.2 CNA Median Nodes 258
 - 11.3.3 Median Nodes for Other Data Types 260
 - 11.4 Steiner Tree Problems 260
 - 11.4.1 SNV Steiner Trees 261
 - 11.4.2 CNA Phylogenetics 262
 - 11.4.3 Hybrid and Alternative Distance Measures 263
 - 11.5 Phylogenetics and Deconvolution 263
 - 11.5.1 Deconvolutional Phylogenies on SNVs 263
 - 11.5.2 Deconvolutional Phylogenies on CNAs 266

11.5.3	Hybrid Deconvolutional Methods	267
11.5.4	Other Marker Types	268
11.6	Emerging Directions	269
11.7	Conclusions	270
	References	271
12	Clusters, Trees, and Phylogenetic Network Classes	277
	Louxin Zhang	
12.1	Mathematical Models of Evolution	277
12.1.1	Phylogenetic Trees	277
12.1.2	Rooted Phylogenetic Networks	278
12.1.3	Applications of Rooted Phylogenetic Networks	280
12.2	Decomposition of Rooted Phylogenetic Networks	281
12.3	Clusters in Rooted Phylogenetic Networks	282
12.3.1	Clusters in Phylogenetic Trees	282
12.3.2	Cluster Networks and Regular Networks	283
12.3.3	Softwired Clusters in Rooted Phylogenetic Networks	285
12.3.4	The Cluster Containment Problem	286
12.3.5	Robinson–Foulds Distances	290
12.4	Phylogenetic Trees and Rooted Phylogenetic Networks	290
12.4.1	Trees in Rooted Phylogenetic Networks	290
12.4.2	Tree-Based Phylogenetic Networks	291
12.4.3	The Tree Containment Problem	295
12.5	Reticulation-Visible Phylogenetic Networks	295
12.5.1	The Node Visibility Property	295
12.5.2	Reticulation-Visible Networks	297
12.5.3	Nearly Stable Networks	299
12.5.4	A Characterization of Galled Networks	301
12.5.5	Sizes of Reticulation-Visible Networks	301
12.5.6	Linear-Time Algorithms for the Cluster Containment Problem	304
12.5.7	Fast Algorithms for the Tree Containment Problem	308
12.6	Relationships Among Network Classes	311
12.7	Bibliographic Notes	312
	References	313
13	Advances in Computational Methods for Phylogenetic Networks in the Presence of Hybridization	317
	R. A. Leo Elworth, Huw A. Ogilvie, Jiafan Zhu and Luay Nakhleh	
13.1	Introduction	318
13.2	Background for Nonbiologists	322
13.2.1	Terminology	322

13.2.2	Phylogenetic Trees and Their Likelihood	324
13.3	From Humble Beginnings: Smallest Displaying Networks	327
13.3.1	The Topology of a Phylogenetic Network	327
13.3.2	Inferring Smallest Displaying Networks	327
13.3.3	Phylogenetic Networks as Summaries of Trees	329
13.3.4	A Step Toward More Complexity: Minimizing Deep Coalescences	331
13.4	Phylogenetic Networks: A Generative Model of Molecular Sequence Data	332
13.4.1	Parameterizing the Network’s Topology	333
13.4.2	The Multispecies Network Coalescent and Gene Tree Distributions	334
13.5	Maximum Likelihood Inference of Phylogenetic Networks . . .	335
13.5.1	Inference	336
13.6	Bayesian Inference of Phylogenetic Networks	339
13.6.1	Probability Distributions Over Species Networks . . .	340
13.6.2	Sampling the Posterior Distribution	341
13.6.3	Inference Under MSC Versus MSNC When Hybridization Is Present	343
13.7	Phylogenetic Invariants Methods	345
13.8	Phylogenetic Networks in the Population Genetics Community	348
13.8.1	TreeMix	348
13.9	Data, Methods, and Software	349
13.9.1	Limitations	351
13.10	Conclusions and Future Directions	352
	References	353
14	A Perspective on Comparative and Functional Genomics	361
	Daniel Doerr and Jens Stoye	
14.1	Introduction	361
14.2	Background	363
14.3	Comparative Detection of Functional Regions	364
14.4	Statistical Analysis	368
14.5	Analysis of Seven Amniote Genomes	369
14.6	Conclusion and Outlook	370
	References	371
15	Integer Linear Programming in Computational Biology: Overview of ILP, and New Results for Traveling Salesman Problems in Biology	373
	Dan Gusfield	
15.1	Introduction	373
15.2	Brief Overview of ILP	375

- 15.2.1 LP- and ILP-Solvers 375
- 15.3 Biological Graphs and Networks 376
 - 15.3.1 High-Density Subgraphs: A Nontrivial Biological Feature 377
- 15.4 The Maximum Clique and Maximum Independent Set Problems and Their Solutions Using ILP 379
 - 15.4.1 An Abstract ILP Formulation for the Maximum Independent Set Problem 379
- 15.5 New Results on the Traveling Salesman Problem (TSP) in Biology 380
 - 15.5.1 Introduction to TSP 380
- 15.6 The Traveling Salesman Problem in Genomics 381
 - 15.6.1 Examples of the TS Problems in Computational Biology 381
- 15.7 Solving TS Problems with Integer Linear Programming 382
 - 15.7.1 DFJ: The Classical ILP Formulation 382
- 15.8 A Compact ILP Solution to the TS Tour Problem on G' 384
 - 15.8.1 The GG TSP Formulation 385
 - 15.8.2 The MTZ Formulation 386
- 15.9 Empirical Results 388
 - 15.9.1 Results for the GG Formulation 388
 - 15.9.2 Comparing Empirical Results for the Other Compact Formulations 390
- 15.10 Take Home Lessons 393
- 15.11 On Strength 394
 - 15.11.1 Is Strength a Good Predictor of Efficiency in Practice? 396
- References 402
- Index 405**