COMPUTATIONAL PHYLOGENETICS

An Introduction to Designing Methods for Phylogeny Estimation

A comprehensive account of both basic and advanced material in phylogeny estimation, focusing on computational and statistical issues. No background in biology or computer science is assumed, and there is minimal use of mathematical formulas, meaning that students from many disciplines, including biology, computer science, statistics, and applied mathematics, will find the text accessible.

The mathematical and statistical foundations of phylogeny estimation are presented rigorously, following which more advanced material is covered. This includes substantial chapters on multi-locus phylogeny estimation, supertree methods, Markov models of sequence evolution, multiple sequence alignment techniques, and designing methods for large-scale phylogeny estimation. The author provides key analytical techniques to prove theoretical properties about methods, as well as addressing performance in practice for methods for estimating trees. Research problems requiring novel computational methods are also presented, so that graduate students and researchers from varying disciplines will be able to enter the broad and exciting field of computational phylogenetics.

TANDY WARNOW is a Founder Professor of Engineering at the University of Illinois at Urbana-Champaign. Her awards include the National Science Foundation Young Investigator Award (1994), the David and Lucile Packard Foundation Award in Science and Engineering (1996), a Radcliffe Institute for Advanced Study Fellowship (2003), and a John Simon Guggenheim Memorial Foundation Fellowship (2011). She was elected a Fellow of the Association for Computing Machinery (ACM) in 2006, and of the International Society for Computational Biology (ISCB) in 2017.

COMPUTATIONAL PHYLOGENETICS

An Introduction to Designing Methods for Phylogeny Estimation

TANDY WARNOW
University of Illinois, Urbana-Champaign

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi – 110002, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the Universitys mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107184718
DOI: 10.1017/9781316882313

First published 2018

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-18471-8 Hardback

Dedicated to Eugene Leighton (Gene) Lawler (1933–1994), my doctoral advisor,
whose enthusiasm and generosity inspired me throughout graduate school,
and who introduced me to the field of computational phylogenetics.

Contents

Preface

Overview

The evolutionary history of a set of genes, species, or individuals provides a context in which biological questions can be addressed. For this reason, phylogeny estimation is a fundamental step in many biological studies, with many applications throughout biology, such as protein structure and function prediction, analyses of microbiomes, inference of human migrations, etc. In fact, there is a famous saying by Dobzhansky that "Nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1973).

Because phylogenies represent what has happened in the past, they cannot be directly observed but rather must be estimated. Consequently, increasingly sophisticated statistical models of sequence evolution have been developed, and are now used to estimate phylogenetic trees. Indeed, over the last few decades, hundreds of software packages and novel algorithms have been developed for phylogeny estimation, and this influx of computational approaches into phylogenetic estimation has transformed systematics. The availability of sophisticated computational methods, fast computers and high-performance computing (HPC) platforms, and large sequence datasets enabled through DNA sequencing technologies, has led to the expectation that highly accurate large-scale phylogeny estimation, potentially answering open questions about how life evolved on earth, should be achievable.

Yet large-scale phylogeny estimation turns out to be much more difficult than expected, for multiple reasons. First, all the best methods are computationally intensive, and standard techniques do not scale well to large datasets; for example, maximum likelihood phylogeny estimation is NP-hard, so exact solutions cannot be found efficiently (unless P = NP), and Bayesian MCMC methods can take a long time to reach stationarity. While massive parallelism can ameliorate these challenges to some extent, it doesn't really address the basic challenge inherent in searching an exponential search space. However, another issue is that the statistical models of sequence evolution that properly address genomic data are substantially more complex than the ones that model individual loci, and methods to estimate genome-scale phylogenies are (relatively speaking) in their infancy compared to methods for single gene phylogenies. Finally, there is a substantial gap between performance as suggested by mathematical theory (which is used to establish guarantees about

methods under statistical models of evolution) and how well the methods actually perform on data – even on data generated under the same statistical models! Indeed, this gap is one of the most interesting things about doing research in computational phylogenetics, because it means that the most impactful research in the area must draw on mathematical theory (especially probability theory and graph theory) as well as on observations from data.

The main goal of this text is to enable researchers (typically graduate students in computer science, applied mathematics, or statistics) to be able to contribute new methods for phylogeny estimation, and in particular to develop methods that are capable of providing improved accuracy for large heterogeneous datasets that are characteristic of the types of inputs that are increasingly of interest in practice. The secondary goal is to enable biologists to understand the methods and their statistical guarantees under these models of evolution, so that they can select appropriate methods for their datasets, and select appropriate datasets given the available methods.

Some of the material in the textbook is fairly mathematical, and presumes undergraduate coursework in discrete mathematics and algorithm design and analysis. However, no background in biology is assumed, and the assumed statistics background is relatively lightweight. While some students without the expected background in computer science may find it difficult to understand some of the proofs, my goal has been to enable all students to understand the theoretical guarantees for phylogeny estimation methods and the statistical models on which they are based, so that they can adequately critique the scientific literature, and also choose methods and datasets that are best able to address the scientific questions they wish to answer.

Outline of the Textbook

Part I provides the "Discrete Mathematics for Phylogenetics" foundations for the textbook; the concepts and mathematics introduced in this part are the building blocks for algorithm design in phylogenetics, especially for developing methods that can scale to large datasets; understanding these concepts makes it possible to understand theoretical guarantees of methods under statistical models of evolution. Chapter 1 introduces the major themes involved in computational phylogenetics, addressing both theory (e.g., statistical consistency under a statistical model of evolution) and performance on both simulated and biological data. This chapter uses the Cavender–Farris–Neyman model of binary sequence evolution since understanding issues in analyzing data generated by this very simple model is helpful to understanding statistical estimation under the commonly used models of molecular sequence evolution. Chapter 2 introduces trees as graph-theoretic objects, and presents different representations of trees that will be useful for method development. Chapters 3, 4, and 5 present different types of methods for phylogenetic tree estimation (based on combining subtrees, using character data, or using distances, respectively). Chapter 6 presents methods for analyzing sets of trees, each on the same set of taxa, and for computing consensus trees and agreement subtrees; it also discusses how these methods are used to estimate support for different phylogenetic hypotheses. Chapter 7 examines the

topic of supertree estimation, where the input is a set of trees on overlapping sets of taxa and the objective is a tree on the full set of taxa. Supertree methods are of interest in their own right and also because they are key algorithmic ingredients in divide-and-conquer methods, a topic we return to in Chapter 11.

Part II of the textbook is concerned with molecular phylogenetics. Chapter 8 presents commonly used statistical models of molecular sequence evolution and statistical methods for phylogeny estimation under these models. However, standard sequence evolution models do not include events such as insertions, deletions, and duplications, which can change the sequence length. These are very common processes, so biological sequences are usually of different lengths and must first be put into a *multiple sequence alignment* before they can be analyzed using phylogeny estimation methods; the subject of how to compute a multiple sequence alignment is covered in Chapter 9. Constructing a species tree or even a phylogenetic network from different gene trees in the presence of gene tree heterogeneity due to incomplete lineage sorting, gene duplication and loss, horizontal gene transfer, or hybridization is a fascinating research area that we present in Chapter 10. We end with Chapter 11, which addresses method development for estimating trees on large datasets. Large-scale phylogeny estimation is increasingly important since nearly all good approaches to phylogeny and multiple sequence alignment estimation are computationally intensive (either heuristics for NP-hard optimization problems or Bayesian methods), and many large datasets are being assembled that cannot be accurately analyzed using existing methods.

Each chapter ends with a set of review questions and homework problems. The review questions are easy to answer and do not require any significant problem solving or calculation. The homework problems are largely pen and paper problems that reinforce the mathematical content of the text.

The textbook comes with four appendices. Appendix A provides an introduction to biological evolution and data; the textbook can be read without it, but the reader who wishes to analyze biological data will benefit from this material. Appendix B provides an introduction to algorithm design and analysis; this material is not necessary for students with undergraduate computer science backgrounds, but may be a helpful introduction for students without this background. Appendix C provides some guidelines about how to write papers that introduce new methods or evaluate existing methods. Appendix D provides computational projects ranging from short term (i.e., a few days) to research projects that could lead to publications. In fact, several of the final projects for my Computational Phylogenetics courses have grown into journal publications (e.g., Bayzid et al. (2014); Zimmermann et al. (2014); Davidson et al. (2015); Chou et al. (2015); Nute and Warnow (2016)).

I wish to thank my editor, David Tranah, for his detailed and insightful comments on the many earlier versions of the text. I also wish to thank my students, colleagues, and family members who gave helpful criticism, including Edward Braun, Sarah Christensen, Steve Evans, Dan Gusfield, Joseph Herman, Ally Kaminsky, Laura Kubatko, Ari Löytynoja, Siavash Mirarab, Erin Molloy, Luay Nakhleh, Mike Nute, David Posada, Bhanu Renukuntla, Ehsan Saleh, Erfan Sayyari, Kimmen Sjölander, Travis Wheeler, and Mark Wilkinson. The several anonymous reviewers also gave very useful comments.

The images of the Monterey Cypress tree on the front and back covers are in honor of the CIPRES project (www.phylo.org), an NSF-funded project for phylogenetic research that I co-led with Bernard Moret from 2003–2010. Many of the algorithmic advances discussed in the text came out of research supported by CIPRES.

Glossary

afc: Absolute fast-converging methods

ASSU: The algorithm by Aho, Sagiv, Szymanski, and Ullman for determining if a set of rooted triplet trees is compatible, and constructing the compatibility tree if it exists

centroid edge: An edge in a tree T whose deletion defines a decomposition of the leafset into two parts that is as close to balanced as possible

c-gene: A region within a set of genomes that is recombination-free

GTR: General Time Reversible model

GTR-GAMMA: The GTR model with gamma-distributed rates across sites

HMM: Hidden Markov model

homologous: Two sequences are homologous if they have descended from a common ancestor

homoplasy: Evolution with back-mutation or parallel evolution

indels: Insertions and deletions

JC69: Jukes–Cantor model

K2P: Kimura 2-parameter model

MRCA: Most recent common ancestor

MSC: Multi-species coalescent model

NP: The class of decision (i.e., yes/no) problems for which the yes-instances can be verified in polynomial time

NP-hard: A problem that is at least as difficult as the hardest problems in the class NP

NP-complete: A decision problem that is NP-hard and also in the class NP

polytomy: A node in an unrooted tree of degree greater than three, or a node in a rooted tree with more than two children

Notation

- λ: The empty string
- $ab|cd$: Quartet tree on leafset a,b,c,d with one internal edge separating a,b from c,d
- $(a,(b,c))$: Rooted tree on three leaves a,b,c in which b and c are siblings
- $Clades(T)$: The set of clades of a rooted tree T, where a clade is the set of leaves below some internal node in T
- $C(T)$: The set of bipartitions on the leafset induced by edge deletions in a tree T
- $C_I(T)$: The set of non-trivial bipartitions on the leafset induced by deletions of internal edges in a tree T
- $\mathscr{L}(T)$: The set of leaves of a tree T
- $L_\infty(M,M')$: For matrices $\mathbf{M}$ and $\mathbf{M}'$ with the same dimensions, this is $max_{ij}\,|M_{ij}-M'_{ij}|$
- $M[i,j]$: For matrix $\mathbf{M}$, this is the entry in row i and column j. This is also denoted by M_{ij}.
- $MP(T,M)$: The maximum parsimony score of a tree T given the character matrix $\mathbf{M}$
- $Q(T)$: The set of homeomorphic unrooted quartet trees induced by T on its leafset
- $Q_r(T)$: The set of unrooted fully resolved (i.e., binary) quartet trees in $Q(T)$
- $|S|$: The number of elements in the set S
- $S \setminus S'$: The set $\{x : x \in S \text{ and } x \notin S'\}$ (i.e., the elements of S that are not in S')
- $T|X$: The subtree of T induced on leafset X, with nodes of degree two suppressed
- T_u: The unrooted tree obtained by suppressing the root for T