Estimating Species Trees

Practical and Theoretical Aspects

EDITED BY

L. Lacey Knowles and Laura S. Kubatko

WILEY-BLACKWELL

ESTIMATING SPECIES TREES

Practical and Theoretical Aspects

Edited by

L. LACEY KNOWLES
Department of Ecology and Evolutionary Biology and Museum of Zoology
University of Michigan

LAURA S. KUBATKO
Department of Statistics and Department of Evolution, Ecology, and Organismal Biology
The Ohio State University

WILEY-BLACKWELL

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

978-0-470-52685-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ESTIMATING SPECIES TREES

PREFACE

In January 2009, the first workshop on species tree estimation was held at the University of Michigan. Because the merger of phylogenetic and population approaches is central to methods for estimating species trees, the workshop was initiated to address the gap between the population genetic principles on which these methods are based and the backgrounds of those interested in applying these procedures (i.e., students interested in phylogenetics are often not trained in population genetics). Building bridges and filling this knowledge gap was a key goal of the workshop. Not only were the practicalities of using software taught, but key concepts were also discussed in order to enable researchers to make informed decisions as they delve into this exciting new area of phylogenetic study.

The influx of multilocus data in phylogenetics is a primary driving force behind this new development in molecular systematics—the direct estimation of species trees, as opposed to relying on gene trees for phylogenetic inference. However, it is the inherent appeal of these new approaches for species tree estimation that explains the tremendous interests in their application and their analytical development. Specifically, with the unprecedented access to molecular data and improvements in computational techniques, there is no justification for ignoring an inescapable biological reality—gene trees differ from one another for a variety of reasons. Moreover, an explicit accounting of this variance not only can usher in more accurate estimates of species relationships but also can reveal the biological processes that have influenced the diversification history and shaped organismal genomes.

In the book, we highlight, by example, how species tree estimation differs from traditional phylogenetic estimation. This includes both conceptual and practical issues related to improving species tree estimates. The first half of the book devotes six chapters to methodological developments, whereas the last half of the book, also with six chapters, focuses on empirical applications. The contributors were among the original participants from the workshop. Through the collection of chapters (each of which represents the authors' own perspective on aspects of species tree estimation arising from their individual research programs), a diversity of perspectives and backgrounds are presented. This diversity means that the book speaks to people with varying levels of familiarity with the topic of species tree estimation. However, it does not (nor is it intended to) provide a comprehensive overview of the subject. The combination of theoretical and empirical work is meant to provide readers with a level of knowledge of both the advances and limitations of this nascent area of phylogenetics in order to assist researchers in applying the methods, while also inspiring future advances among those researchers with an interest in methodological development. Such cross talk (between empiricists and theoreticians) is vital to the growth of the new area of molecular phylogenetics as it refocuses attention on the biological history of diversification (i.e., the timing and pattern of species divergence), including the processes generating the observed patterns of genetic variation (e.g., sorting of ancestral polymorphism and gene flow, in addition to mutation models of nucleotide evolution). In this way, the book provides access to a molecular phylogenetic perspective that, unlike the vast majority of phylogenetic methods that focus on the estimation of gene trees, places the focus on the primary target of interest—the species tree.

L. Lacey Knowles
Laura S. Kubatko

CONTRIBUTORS

Cécile Ané, Department of Statistics and Department of Botany, University of Wisconsin-Madison, Medical Science Center, 1300 University Ave., Madison, WI 53706-1532; Email: ane@stat.wisc.edu

Natalia M. Belfiore, Visiting Scholar, Department of Statistics, University of California-Berkeley; Email: nmb@berkeley.edu

Robb T. Brumfield, Museum of Natural Science and Department of Biological Sciences, 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803; Email: brumfld@lsu.edu

Matthew D. Carling, Laboratory of Ornithology, Department of Ecology and Evolutionary Biology, 159 Sapsucker Woods Road, Cornell University, Ithaca, NY 14850; Email: mdc248@cornell.edu

Santiago Castillo-Ramírez, Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Apartado Postal 565-A, CP 62210, Cuernavaca, Morelos, México; Email: iago@ccg.unam.mx

Karen A. Cranston, Biodiversity Synthesis Center, Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago, IL 60605; Email: kcranston@fieldmuseum.org

James H. Degnan, Department of Mathematics and Statistics, Private Bag 4800, University of Canterbury, Christchurch 8140 New Zealand; Email: J.Degnan@math.canterbury.ac.nz

Scott V. Edwards, Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138; Email: sedwards@fas.harvard.edu

H. Lisle Gibbs, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, 370 Aronoff Laboratory, 318 W. 12th Avenue, Columbus, OH 43210; Email: gibbs.128@osu.edu

L. Lacey Knowles, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079; Email: knowlesl@umich.edu

Laura S. Kubatko, Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210; Email: lkubatko@stat.osu.edu

Catherine R. Linnen, Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138; Email: clinnen@oeb.harvard.edu

Liang Liu, Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138; Email: lliu@oeb.harvard.edu

Chen Meng, Monsanto Company, 800 North Lindbergh Boulevard, O3A, St. Louis, MO 63167; Email: chen.meng@monsanto.com

Luay Nakhleh, Department of Computer Science, Rice University, 6100 Main Street, MS 132, Houston, TX 77005; Email: nakhleh@cs.rice.edu

Dennis Pearl, Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210; Email: dkp@stat.osu.edu

Cuong Than, Department of Human Genetics, Bioinformatics Program, 2017 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218; Email: cvthan@cs.rice.edu

CONTENTS

CHAPTER 4 *PROBABILITIES OF GENE TREE TOPOLOGIES WITH INTRASPECIFIC SAMPLING GIVEN A SPECIES TREE* 53

James H. Degnan

CHAPTER 5 *INFERENCE OF PARSIMONIOUS SPECIES TREE FROM MULTILOCUS DATA BY MINIMIZING DEEP COALESCENCES* 79

Cuong Than and Luay Nakhleh

CHAPTER 6 *ACCOMMODATING HYBRIDIZATION IN A MULTILOCUS PHYLOGENETIC FRAMEWORK* 99

Laura S. Kubatko and Chen Meng

CHAPTER 1

ESTIMATING SPECIES TREES: AN INTRODUCTION TO CONCEPTS AND MODELS

L. Lacey Knowles
Laura S. Kubatko

1.1 INTRODUCTION

The estimation of relationships among species in an evolutionary context broadly falls within the purview of the discipline of systematics. However, as the central framework in evolutionary (and some ecological) study, the enormous impact of this single endeavor—phylogenetic estimation—is unquestionable. How, and whether, species relationships are accurately inferred are, consequently, issues of broad and far-reaching concern.

The goal of this book is to provide an overview of several recently developed methods for phylogenetic estimation that focus explicitly on the challenges and strengths inherent in the analysis of multilocus data while giving practical guidelines on implementing these approaches. Decreased sequencing costs and increased access to primer sets enhance the relative ease of data collection, providing unprecedented amounts of multilocus sequence for molecular phylogenetic analysis across all of biodiversity (e.g., Goldman and Yang 2008; Hughes et al. 2006; Wiens et al. 2008). Detailed suggestions and discussion throughout the chapters focus on both conceptual and methodological issues, addressing such topics as how results should be interpreted and how to recognize the signs of a problem with an analysis. The combination of theoretical and empirical studies contained herein serves to identify both the strengths and the limitations of these new methods under not only idealized situations with simulated data but also with empirical sequence data. The guidelines also serve to draw attention to the impact that sampling design, marker choice, and taxon sampling will have on the performance of the new methods.

1.1.1 Different Tree Types and Their Relationship to Phylogeny

As a characterization of the history of species divergence (including both the pattern and relative timing of lineage splitting), a phylogeny is a tree where both the topology and branch lengths portray information about the evolutionary history of species (Fig. 1.1). While molecular data predominate the pursuit of estimating the evolutionary history of species, the trees estimated from DNA sequences are clearly distinct from, and are not

Estimating Species Trees: Practical and Theoretical Aspects, Edited by L. Lacey Knowles and Laura S. Kubatko

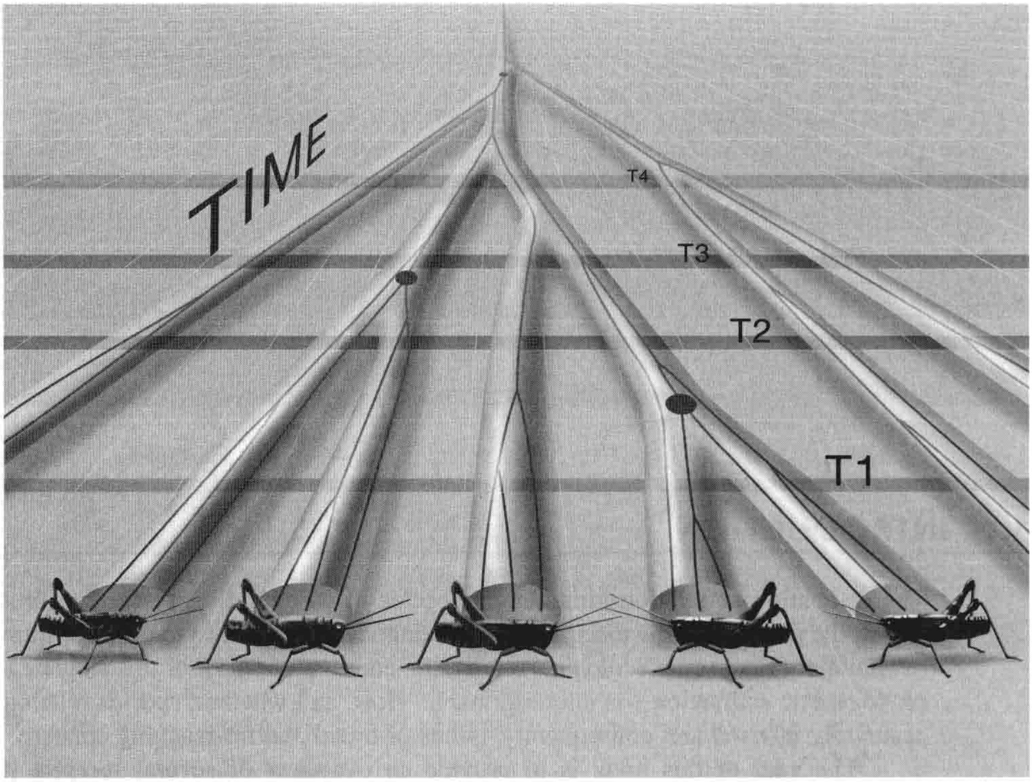

Figure 1.1 Species trees contain information on both the pattern (topology) and timing (branch lengths) of species diversification. This phylogenetic history can be inferred from the gene trees that are embedded within the species lineages, which may or may not be concordant with the species tree (e.g., the deep coalescence of gene lineages marked with the red dots). By incorporating a model of gene lineage coalescence (in addition to the models of nucleotide substitution), the phylogenetic history of species (i.e., the species tree) can be estimated, despite widespread incomplete lineage sorting (i.e., sequences from multiple individuals per species—three individuals for this locus in this case—do not form monophyletic clades). (Illustration by John Megahan.)

synonymous with, the underlying species history—the species tree (Maddison 1997; Slowinski and Page 1999). In contrast to the differing genealogical histories (i.e., gene trees) that might characterize a locus (or a nonrecombining DNA fragment), there is only one species history, whether that history is strictly bifurcating (i.e., a species tree) or involves reticulations, which may or may not obscure species relationships.

The patterns of similarity and differences in the DNA sequences of organisms related by descent from common ancestors implicitly contain information about species relationships. That is, there is an intimate link between gene trees and the species tree in which they are embedded. This link means that gene trees are informative about species phylogenies, yet it is clear that a gene tree should not be equated with a species phylogeny since the evolutionary processes that determine the structure of gene trees differ from those governing species trees. The structure of a species tree is determined by the process of speciation, extinction, and in some cases, hybridization, whereas the gene tree structure reflects not only the proliferation and loss of species lineages but also the population genetic process of mutation and gene lineage coalescence within species lineages, and in some cases, the locus-specific effects of migration between species lineages.

Enormous attention has been dedicated to understanding the theoretical and computational challenges associated with estimating gene trees from molecular data, as well as the practical complications that arise with empirical investigations. For example, in addition to the development of very sophisticated methods for estimating a gene tree from DNA sequences (e.g., accommodating complex models of nucleotide evolution and evaluating the full probability of the data for a set of tree topologies and branch lengths; reviewed in Felsenstein 2004), the impact of various data properties on tree accuracy is also well studied (e.g., the number of base pairs analyzed and taxon sampling; Flynn et al. 2005; Graybeal 1998; Rannala et al. 1998; Rosenberg and Kumar 2001; Wiens 2003; Zwiki and Hillis 2002). In contrast, we are only beginning to understand the theoretical and computational challenges, as well as the practical complications of empirical data, when the target is to obtain an estimate of the species tree. For example, multiple processes may determine the relationship between species and their contained loci (e.g., gene lineage coalescence alone or in combination with gene flow). Moreover, the collection of possible bifurcating trees (i.e., the *tree space*) becomes enormous even for a moderate number of species. For example, even if only bifurcating processes are considered, and ignoring differences in branch lengths, there are approximately 2×10^6 trees for 10 taxa. The difficulties posed by such issues, as well as strategies for contending with these challenges, are discussed in the following sections that trace the steps from species tree estimation back to the collection of DNA sequence data.

While much of the research on obtaining direct estimates of species trees has been driven by computational developments, these methodological changes do not represent the inception of new core phylogenetic concepts. The recent advances (paradoxically) provide a practical means of returning to the systematic tradition of estimating species relationship. Thus, in spite of the fact that estimating species trees involves a fundamental shift in how molecular data are used and interpreted, the target is still the phylogeny. Estimation of a species tree, in addition to putting the focus on the object of systematic interest, also provides a framework for studying the processes generating a set of contained gene trees because of the explicit distinction between the species tree and gene trees. For example, the discord among gene trees may be biologically meaningful (as opposed to being due to tree-building errors, for example; Jeffroy et al. 2006). The different gene trees may provide insights about the diversification process (e.g., the population size of the taxa relative to the divergence time separating speciation events, or the extent of gene flow among taxa), or whether species trees are meaningful if there is significant horizontal gene transfer, a question that requires empirical evaluation (e.g., Galtier and Daubin 2008).

1.2 THE RELATIONSHIP BETWEEN GENE TREES AND SPECIES TREES

Gene trees and species trees are different from one another for a variety of reasons. The most important of these is the possibility that evolutionary processes such as horizontal gene transfer, hybridization, gene duplication, or incomplete lineage sorting lead to differences in the underlying histories of each gene for a given species phylogeny. Understanding these evolutionary processes and their effect on the relationship between gene trees and species trees is thus a problem of central importance to the development of methods for estimating species phylogenies: the goal is estimation of species trees; the data available to do this come in the form of DNA sequences arising from the histories of individual genes. We must therefore strive to understand and effectively model the process by which sequence data arise on the individual gene trees, conditional on the overall species-level relationships.

The methods described and illustrated in this book incorporate one or more of the evolutionary processes mentioned above, and many of these models are common to several of the subsequent chapters on species tree estimation. For this reason, we will devote the next few sections to giving a relatively broad overview of the common models used to relate gene trees to species trees, with ample references to which the reader is directed to obtain a more detailed explanation. Section 1.2.1 defines the processes of horizontal gene transfer, gene duplication, hybridization, and incomplete lineage sorting, and briefly describes their effects on relationships between gene trees and species trees. Section 1.2.2 gives a more detailed description of the coalescent process because it is fundamental to several of the methods included in this book (e.g., Chapters 2, 4, 5, and 6). Section 1.3 then builds on this by describing methods for modeling nucleotide sequence evolution along gene trees.

1.2.1 Evolutionary Mechanisms for Gene Tree Discord

Maddison (1997) provides a very comprehensive description of the processes mentioned below, with explicit discussion of the effects of these processes on individual gene histories. Here we provide the following brief descriptions:

- *Horizontal gene transfer* is a term used to describe a process by which genetic material is transferred from one species to another at a given point in time (thus corresponding to genetic exchange that occurs “horizontally” across a phylogeny), rather than from parent to offspring (which occurs “vertically” on a phylogeny). This could happen, for instance, when a vector such as a virus carries DNA from one species to another and this genetic material is subsequently integrated into the genome of the infected organism. Horizontal gene transfer events are known to occur commonly in the bacteria (Medigue et al. 1991; Syvanen 1994; Valdez and Pinero 1992). Horizontally transferred genes will, at least initially, be more closely related to the ancestors of the organism from which they were derived than to those in which they currently reside, thus leading to gene trees that differ from the species tree.
- *Gene duplication* refers to the event that a copy of a particular gene is inserted into the genome, followed by the subsequent (and separate) evolution of the two copies. If a single copy of the gene is sampled from each organism, the sampling of a duplicated gene might result in the observation of a gene tree that differs from the species tree. Gene duplication events are prevalent in plants, fish, and insects.
- *Hybridization* between species occurs when two distinct species interbreed, with the resulting formation of hybrid organisms that share some genetic material from each of the parental organisms. When hybridization occurs without formation of a new taxonomic lineage that is distinct from the parental lineages from which it was formed, the process is often referred to as *introgression* or *introgressive hybridization*. Hybridization is ubiquitous in nature, with current estimates that approximately 25% of plants and 10% of animals hybridize (Mallet 2007).
- *Incomplete lineage sorting* occurs when multiple gene lineages persist through speciation events. Following a speciation event, some forms of the gene may be lost, while others are maintained and continue to evolve. This process is illustrated in Figure 1.2a, which shows a species tree for three taxa (outlined in bold, black lines) with several embedded gene trees (thinner, colored lines). For example, in the green gene tree, gene lineage C fails to find a most recent common ancestor with gene lineage B during time interval t, and instead finds a most recent common ancestor with gene lineage A above the root of the species tree. This leads to a gene tree that

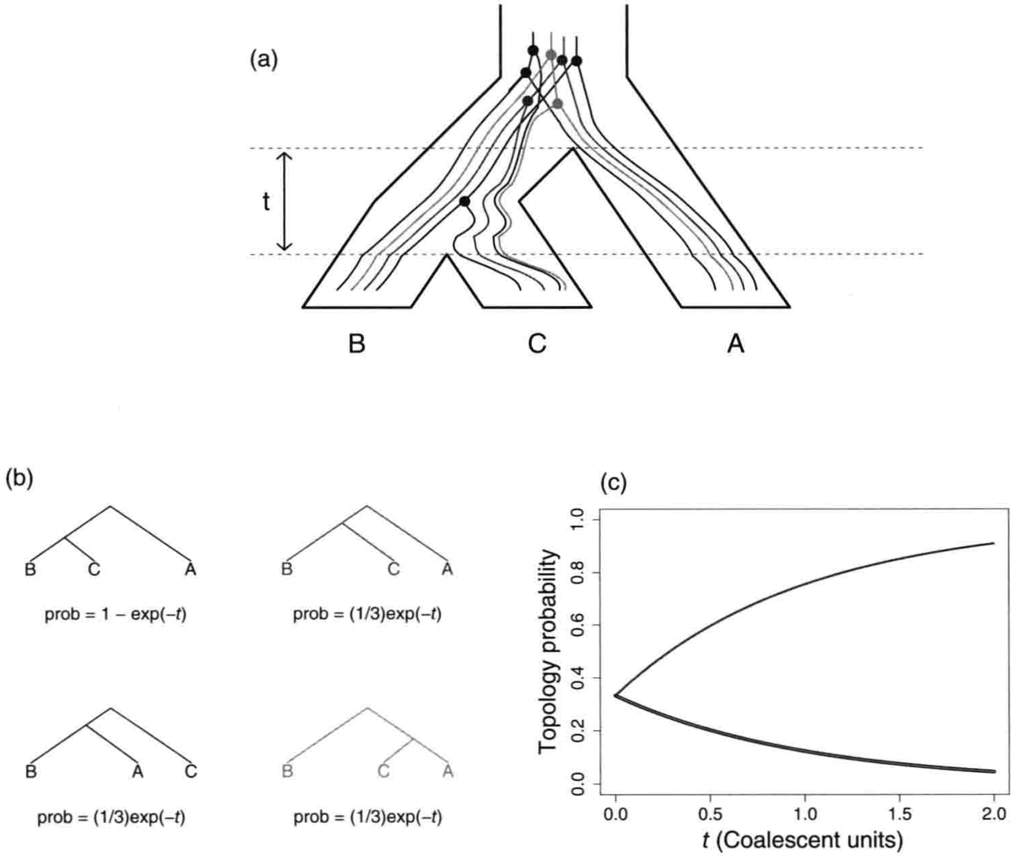

Figure 1.2 Topology probabilities under the coalescent model for three-taxon trees. (a) The species tree is shown outlined in black. The time interval between the two speciation events is *t*, and should be interpreted in coalescent units (number of 2*N* generations). The four embedded trees are the four possible gene histories when deep coalescent events are allowed. (b) The four possible gene histories from (a) are shown separately, with their probabilities under the coalescent model given beneath. Note that the two gene histories in the first row are the same when only the topology is considered, so that the probability of this gene tree topology under the coalescent model is the sum of these two probabilities. Thus, there are only three distinct gene tree topologies in the three-taxon case. (c) Probabilities of each of three gene tree topologies under the coalescent model as a function of the interval of time between speciation events, *t*. Note that the "blue" and "green" gene trees always have the same probabilities. Note also that as *t* increases, the probability of the "red" gene tree (which is the gene tree with the same topology as the species tree) approaches 1.

differs from the species tree (Fig. 1.2b). It is clear that the possibility of such events can result in gene trees that differ in substantial and important ways from the species tree. This process is commonly modeled by the coalescent.

1.2.2 The Coalescent Process and Gene Tree Distributions

Several of the chapters included in this volume develop methodologies for species tree estimation that utilize the coalescent process as a model for the relationship between gene trees and a species tree. For this reason, we include here a more detailed introduction to the basic ideas underlying this process. Excellent books on this topic include the recent works of Hein et al. (2004) and Wakeley (2009).

The *coalescent*, or the *coalescent process*, refers to a mathematical model for the random joining of sampled gene lineages as they are followed back in time. In most