

SpringerBriefs in Systems Biology

David R. Bickel



# Phylogenetic Trees and Molecular Evolution

A Hands-on Introduction  
with Uncertainty  
Quantification Corrected

 Springer

# **SpringerBriefs in Systems Biology**

SpringerBriefs in Systems Biology is an exciting new series of concise publications of cutting-edge research and practical applications in Systems Biology. Systems Biology is the study of the complex interactions between the components of biological systems (genes, proteins, mechanisms, etc), and how these interactions give rise to the function and behavior of that system. The structure and dynamics of cellular and organismal function are examined as a whole, rather than as isolated parts. The interaction of these parts gives rise to new properties and functions which are called “emergent properties”.

David R. Bickel

# Phylogenetic Trees and Molecular Evolution

A Hands-on Introduction with Uncertainty  
Quantification Corrected



Springer

David R. Bickel   
Informatics and Analytics  
University of North Carolina at Greensboro  
Greensboro, NC, USA

ISSN 2193-4746 ISSN 2193-4754 (electronic)  
SpringerBriefs in Systems Biology  
ISBN 978-3-031-11957-6 ISBN 978-3-031-11958-3 (eBook)  
<https://doi.org/10.1007/978-3-031-11958-3>

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed 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

*To Mary Anna, beloved and highly favored*

μακαρια η πιστευσασα οτι εσται τελειωσις τοις  
λελαλημενοις αυτη παρα κυριου

# Preface

## Why This Book?

### *Brief Description*

The book introduces molecular phylogenetics and explains how to interpret the results of evolutionary models, which by their nature only quantify a fraction of the uncertainty involved. The book is intended to introduce upper-level undergraduate and graduate students of biology to phylogenetic tree reconstruction and its dependence on models of molecular evolution. The scope encompasses basic concepts of estimating divergence times and ancestral sequences and of quantifying the uncertainty of the estimates.

### **Key features:**

- Selective use of mathematics
  - An informal, graphical introduction to the main concepts of molecular phylogenetics
  - A competitive dice game to teach evolution modeling without equations
  - Gradual exposure to simple formulas for understanding phylogenetic methods
  
- Uncertainty quantification
  - Thorough explanations of sources of uncertainty in phylogenetics
  - Simple methods for quantifying uncertainty not captured by phylogenetics software
  
- Concepts underlying present and future software more than details of today's programs

## ***Why the Subtitle “A Hands-On Introduction with Uncertainty Quantification Corrected?”***

This book serves two purposes. First, it introduces biology students to phylogenetic trees and molecular evolution without requiring any mathematics beyond elementary algebra. It does so by presenting the main concepts in a variety of ways: first in graphics, then in some history of the field, next in a dice game, and finally in simple equations. Completing the exercises will prepare students for textbooks with more realistic mathematical modeling and details of current software.

Second, the book explores the interface between molecular evolution and uncertainty quantification, of potential benefit to both fields. For not only does phylogenetics clearly illustrate the need for uncertainty quantification in science, but quantifying the uncertainty involved in estimating phylogenetic trees can lead to reliable biological interpretations. This book equips students with statistical tools for correcting uncertainty quantification for the study of evolutionary relationships between DNA and protein sequences.

### **Chapter-by-Chapter Synopsis**

1. Chapter 1 starts slow, without assuming more than a basic knowledge of biology or mathematics. Readers are exposed to the main concepts of the book through graphics and an easy-to-read, conversational exposition. The absence of formulas facilitates comprehension for readers intimidated by mathematics. The chapter ends with simple exercises to reinforce the basic ideas.
2. This chapter continues to expose readers to the main concepts of the book. It does so by sketching a history of the main developments of molecular phylogenetics and molecular evolution since the 1960s. All mathematical formulas related to this chapter are postponed until Appendix B and Appendix C. The exercises at the end of the chapter encourage reflection, especially on how the history can inform assessments of the uncertainty in estimated trees.
3. The third chapter covers a probability model of molecular evolution in a way designed to engage readers with little interest in mathematics. That is accomplished by stating the model in terms of a competitive game rather than mathematical formulas: one team rolls dice to simulate molecular evolution, and the other team then uses the resulting simulated sequence data to estimate the phylogenetic tree. The description of the game rules gradually eases the readers into some simple mathematical formulas needed to understand how evolutionary distances are corrected for multiple substitutions. Exercises provided at the end of the chapter give readers hands-on experience with the sequence simulation and tree reconstruction methods.
4. The fourth chapter goes into more detail about estimating divergence times from molecular sequence data. It explains the inadequacy of confidence intervals for



quantifying the uncertainty in divergence times. A simple method of improving the uncertainty quantification is provided. Exercises at the end of the chapter give readers experience with quantifying the uncertainty in divergence times.

5. The fifth chapter explains three ways to infer ancestral sequences from sequence data. After explaining a heuristic approach to build the intuition of the readers, the chapter explains maximum likelihood estimation and Bayesian estimation with some simplifications to keep the formulas minimal. To adjust the results for uncertainty not quantified in the models, the previous chapter's method of uncertainty quantification is adapted. Exercises provided at the end of the chapter give readers practical experience with estimating ancestral sequences and intuition about the uncertainty involved.
6. This chapter lifts the uncertainty quantification of the previous chapters up to the level of molecular evolution hypotheses. A relatively recent rival to the neutral theory of molecular evolution is explained in terms of how its predictions differ. The exercises at the end of the chapter give readers experience with quantifying the extent to which sequence data support one evolutionary hypothesis more than another.
7. The last chapter is a guide to further reading on molecular phylogenetics, uncertainty quantification, and other topics encountered in the book.

The three appendices fill in some mathematical gaps.

## Acknowledgments

Molecular evolution models were the subject of my PhD research, which would not have been possible without the mentorship of Bruce J. West. This book started in the form of lecture notes at the Medical College of Georgia. I drafted Appendices **B** and **C** at the University of Ottawa. Teaching computational biology courses at the University of North Carolina at Greensboro led to many updates and additions, and I am grateful to all the students who provided feedback.

The comments of the anonymous reviewers on the book proposal led to the addition of Chap. 7 and to an expansion of Sect. 2.7. I thank Shi Huang for carefully reading a draft of Chap. 6 and for suggesting corrections.

Larissa Albright, my first Springer contact for this book, ensured efficient peer review and publishing agreement processes. I appreciate how quickly Sanjana Sundaram and Merry Stuber answered all my questions as I finalized the manuscript for submission to Springer. I thank Tiffany Lu for helpful correspondence and both Sanjana Sundaram and Vinodhini Srinivasan for coordinating the publication process. I am grateful to A. Meenahkumary at Straive for efficiently managing production. The comments of the two anonymous reviewers led to improved clarity and scope.

My interest in dice games has its origins in unforgettable campaigns with Chris, Brian, John, and David. I warmly thank Dorothy Johnson for her sustained

interest in the progress of the book project. Mary Anna deserves mention for her professionalism in securing the epigraphs' permissions. I enjoyed discussing the topic of the book with her, Evelyn, and Christian and seeing the reaction of Faith and Lydia to the plots of upside-down trees!

Greensboro, NC, USA  
May 2022

David R. Bickel

# Contents

<b>1</b>	<b>Introduction to Phylogenetic Trees</b>	1
1.1	What Are Phylogenetic Trees All About?	1
1.2	Assumptions Behind Models of Molecular Phylogenetics	12
1.2.1	Common Ancestry	12
1.2.2	Molecular Clock Hypothesis	13
1.2.3	Statistical Assumptions	13
1.3	Exercises	14
<b>2</b>	<b>Adaptation of the Molecular Clock: A Divergence Time Story</b>	15
2.1	1960s: Starting the Clock	15
2.2	1970s: Neutralism Versus Selectionism	17
2.3	1980s: Fluctuating Rates	17
2.4	1990s: Confronting the Fossil Record	18
2.5	2000s: Molecular Punctuated Equilibrium	19
2.6	2010s: Integration of Data	20
2.7	2020s: Time Will Tell	21
2.8	Excursus: Models with Molecular Evolution over All Time Scales	21
2.8.1	Why Models with Evolution over All Time Scales?	21
2.8.2	Fractal point process models	21
2.8.3	Fractal-Rate Poisson Models	22
2.8.4	Multiplicative Model of Molecular Evolution	23
2.9	Exercises	23
2.10	Bibliographic Notes	23
<b>3</b>	<b>Estimating Phylogenetic Trees</b>	25
3.1	Substituter: The Poisson Game (3 Sequences of 12 Nucleotides Each)	25
3.1.1	Rules for Simulation Mode	26
3.1.2	Rules for Estimation Mode	29
3.2	Relations of Different Types of Trees	34
3.3	Software for Tree Estimation	35