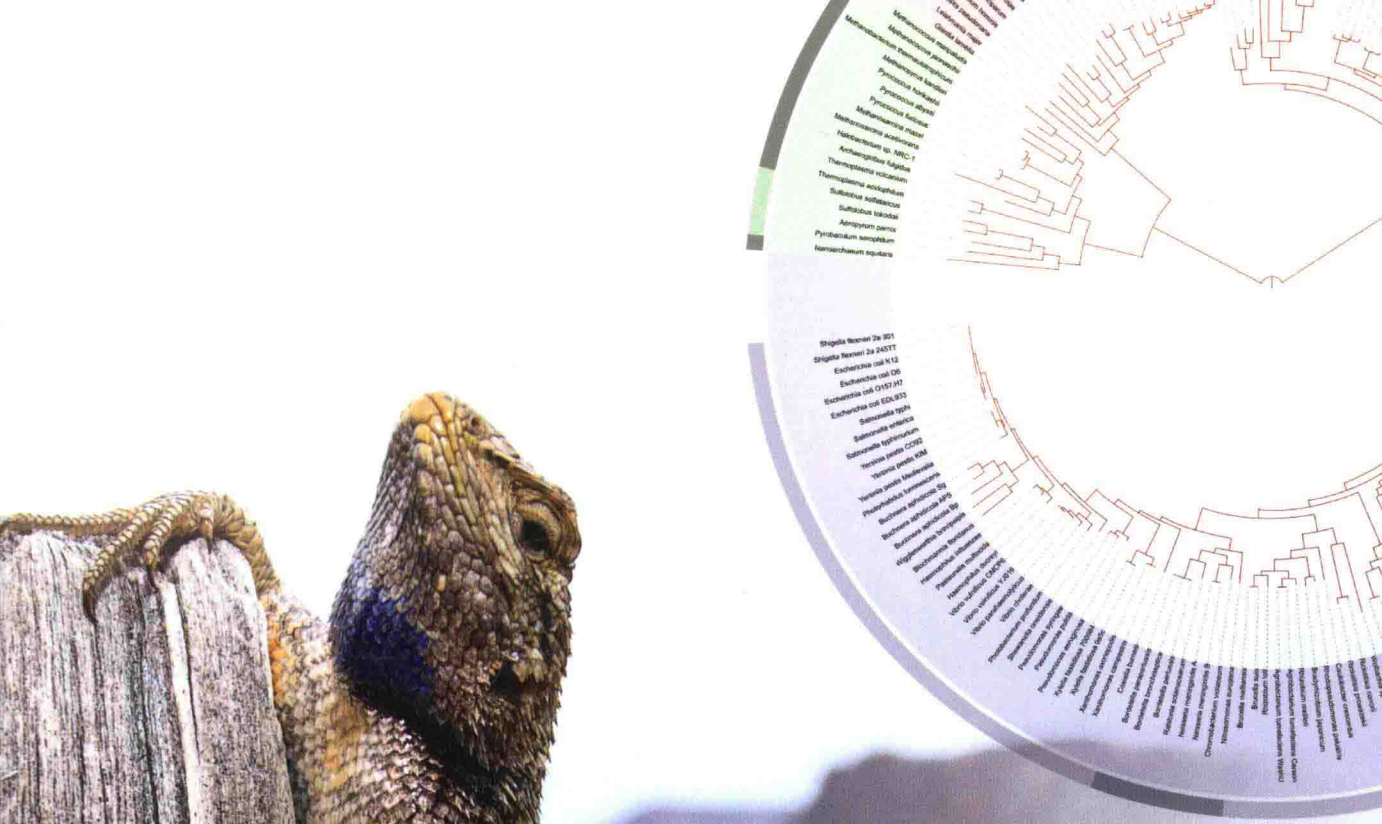
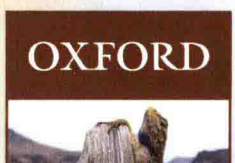


Ziheng Yang

Molecular Evolution

A STATISTICAL APPROACH



Molecular Evolution

A Statistical Approach

ZIHENG YANG

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2014

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2013956540

ISBN 978–0–19–960260–5 (hbk.)
ISBN 978–0–19–960261–2 (pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Molecular Evolution

Foreword

Over the last two decades, Ziheng Yang has been a leading architect of the emergent field of computational molecular evolution. His first book, *Computational Molecular Evolution*, was published in 2006 and became an instant classic. The book broke new ground both in terms of its subject matter and expository style. It presented an up-to-date, detailed, and comprehensive account of computational and statistical aspects of molecular evolutionary analysis, while retaining an informal style and pragmatic perspective that made it highly accessible. The book targeted a readership that included both biologists and applied mathematicians, yet it did not oversimplify in catering to biologists by avoiding advanced calculus or linear algebra, or pandering to mathematicians with the usual theorem-proof format. Somehow, this middle-of-the-road approach seems to have worked. Furthermore, despite the book's graduate textbook flavour the chapters were peppered with Yang's original interpretations and suggestions making it part textbook and part research monograph. Even individuals who were already experienced in computational evolutionary analysis will have gained new insights.

Yang's knowledge and practical experience are evident on every page of his new book, *Molecular Evolution: A Statistical Approach*. What is particularly remarkable is his ability to translate for non-specialists the key developments of this rapidly changing field so effectively. The content represents a significant expansion of his previous book; in particular, the treatment of Bayesian inference is much more extensive. Bayesian inference has become a cornerstone of phylogenetic inference over the last decade, as many programs such as MRBAYES and BEAST are now available which implement Markov chain Monte Carlo (MCMC) simulation methods for this purpose. The book devotes new chapters to the fundamentals of Bayesian inference and MCMC methodologies. Biologists using MCMC programs for molecular evolutionary analyses will benefit from the ground-up approach of these chapters, which introduce the basic principles using motivating examples based on evolutionary processes of obvious practical importance that will be familiar to molecular evolutionists. In this way, remarkably clear explanations are provided for such notoriously difficult concepts as reversible-jump MCMC, Dirichlet processes, Bayes factor calculations for model comparison, and so on. Several excellent books exist on phylogenetic inference, written from either an applied statistical perspective (Felsenstein 2004) or a more rigorous mathematical one (Semple and Steel 2003). However, I am unaware of any book that contains the extensive details found in Yang's book concerning the MCMC implementations (proposal moves, prior distributions, etc.) underlying currently available programs for Bayesian phylogenetic inference.

In this era of cheap next-generation sequencing, multi-locus genomic data are the new norm and therefore the distinction between inference of locus-specific gene trees and multi-locus species trees has become key. *Molecular Evolution: A Statistical Approach* thus contains a new chapter that covers the multi-species coalescent, species tree inference, and species delimitation methods. Yang has been a key contributor to the development of this theory during the last decade and provides one of the clearest explanations of the

multi-species coalescent that I have read. For persons whose research interests include computational molecular evolution and molecular phylogenetics this new book from Ziheng Yang is essential reading.

Bruce Rannala

Davis, California
April 2014

Preface

The main objective of this book is to present and explain the statistical methods and computational algorithms developed in molecular evolution, phylogenetics, and phylogeography for the comparative analysis of genetic sequence data. Reconstruction of molecular phylogeny and inference of the molecular evolutionary process are considered problems of statistical inference, and likelihood and Bayesian methods are treated in depth as standard methods of data analysis. Heuristic and approximate methods are discussed from such a viewpoint as well and are often used to introduce the central concepts, because of their simplicity and intuitive appeal. However, the book does not dwell on proofs or mathematical niceties; it emphasizes care but not rigour.

Molecular Evolution: A Statistical Approach represents an expanded and updated treatment of my earlier research monograph *Computational Molecular Evolution*, published by Oxford University Press in 2006. The major change has been the far more comprehensive and extensive coverage of Bayesian methods, while the target audience has been expanded to include upper level undergraduate as well as graduate students. It can also be read by researchers working in such diverse fields as evolutionary biology, molecular systematics, population genetics, statistical phylogeography, bioinformatics and computational biology, computer science, and computational statistics. It is hoped that biologists who have used software programs to analyse their own data will find the book particularly useful in helping them understand the principles of the methods. For applied mathematicians, molecular studies of evolution are 'a source of novel statistical problems' (Neyman 1971), and this book will provide an accessible summary of the exciting and often unconventional inference problems in the field, some of which are yet unsolved.

Although this new book is written at a similar level of mathematical sophistication as my 2006 work, I have taken care to assist the biologist readers who may find the mathematical arguments challenging. First, every important mathematical result is followed by a verbal rendering, and it is reportedly possible to read the book while skipping the equations, at least at first reading. Second, I have included numerous examples of real data analysis and numerical calculations to illustrate the theory, in addition to the working problems at the end of each chapter. Many biologists find numerical calculations less intimidating than abstract formulae. Example datasets and small C and R programs that implement computational algorithms discussed in the book are posted on the web site for the book: http://abacus.gene.ucl.ac.uk/MESA/. Third, I have prepared a primer on probability and statistics, with an overview of mathematical results used in this book, for biologists who would like to grapple with the mathematical details in the book. This has been used as the pre-course reading material for an advanced workshop on Computational Molecular Evolution (CoME) that runs annually in Hinxton, Cambridge, and Heraklion, Crete, co-organized by Aidan Budd, Nick Goldman, Alexandros Stamatakis, and me. It is available at: http://abacus.gene.ucl.ac.uk/PPS/PrimerProbabilityStatistics.pdf.

The 2006 book was used as a textbook for graduate courses on bioinformatics and computational genomics in Peking University (2010) and in ETH Zurich (2011). I thank the students in those courses for their useful feedback. For instructors, I have found an early

coverage of the simulation chapter to be useful, as afterwards simulation projects can be assigned as homework when other chapters are taught.

I am grateful to a number of colleagues who read earlier drafts of chapters of this book and provided constructive comments and criticisms: Konstantinos Angelis, Mario dos Reis, Ed Susko, Chi Zhang, and Tianqi Zhu. The following colleagues read and commented on Chapter 9: Daniel Dalquen, Adam Leaché, Liang Liu, and Jim Mallet. Needless to say, all errors that remain are mine. (Please report errors and typos you discover to me at z.yang@ucl.ac.uk. Errata will be posted on the book's web site.) Thanks are also due to Helen Eaton, Lucy Nash, and Ian Sherman at Oxford University Press for their support and patience throughout the project.

Ziheng Yang

London
April 2014

Contents

Reader note: The asterisk next to a heading indicates a more difficult or technical section/problem.

CHAPTER 1

Models of nucleotide substitution

1.1 Introduction

Calculation of the distance between two sequences is perhaps the simplest phylogenetic analysis, yet it is important for two reasons. First, calculation of pairwise distances is the first step in distance matrix methods of phylogeny reconstruction, which use cluster algorithms to convert a distance matrix into a phylogenetic tree. Second, Markov process models of nucleotide substitution used in distance calculation form the basis of likelihood and Bayesian methods of phylogeny reconstruction. Indeed, joint analysis of multiple sequences can be viewed as a natural extension of pairwise distance calculation. Thus, besides discussing distance estimation, this chapter introduces the theory of Markov chains used in modelling nucleotide substitutions in a DNA sequence. It also introduces the method of maximum likelihood (ML). Bayesian estimation of pairwise distances and Bayesian phylogenetics are introduced in Chapters 6–8.

The distance between two sequences is defined as the expected number of nucleotide substitutions per site. If the evolutionary rate is constant over time, the distance will increase linearly with the time of divergence. A simplistic distance measure is the proportion of different sites, sometimes called the p distance. If 10 sites are different between two sequences, each 100 nucleotides long, then $p = 10\% = 0.1$. This raw proportion works fine for very closely related sequences but is otherwise a clear underestimate of the number of substitutions that have occurred. A variable site may result from more than one substitution, and even a constant site, with the same nucleotide observed in the two sequences, may harbour back or parallel substitutions (Figure 1.1). Multiple substitutions at the same site or *multiple hits* cause some changes to be hidden. As a result, p is not a linear function of evolutionary time. Thus the raw proportion p is usable only for highly similar sequences, with $p < 5\%$, say.

To estimate the number of substitutions, we need a probabilistic model to describe changes between nucleotides over evolutionary time. Continuous-time Markov chains are commonly used for this purpose. The nucleotide sites in the sequence are assumed to be evolving independently of each other. Substitutions at any particular site are described by a Markov chain, with the four nucleotides to be the *states* of the chain. The main feature of a Markov chain is that it has no memory: 'given the present, the future does not depend on the past'. In other words, the probability with which the chain jumps into other nucleotide states depends on the current state, but not on how the current state is reached. This is known as the *Markovian property*. Besides this basic assumption, we often place further constraints on substitution rates between nucleotides, leading to

Molecular Evolution: A Statistical Approach. Ziheng Yang. © Ziheng Yang 2014.
Published 2014 by Oxford University Press.

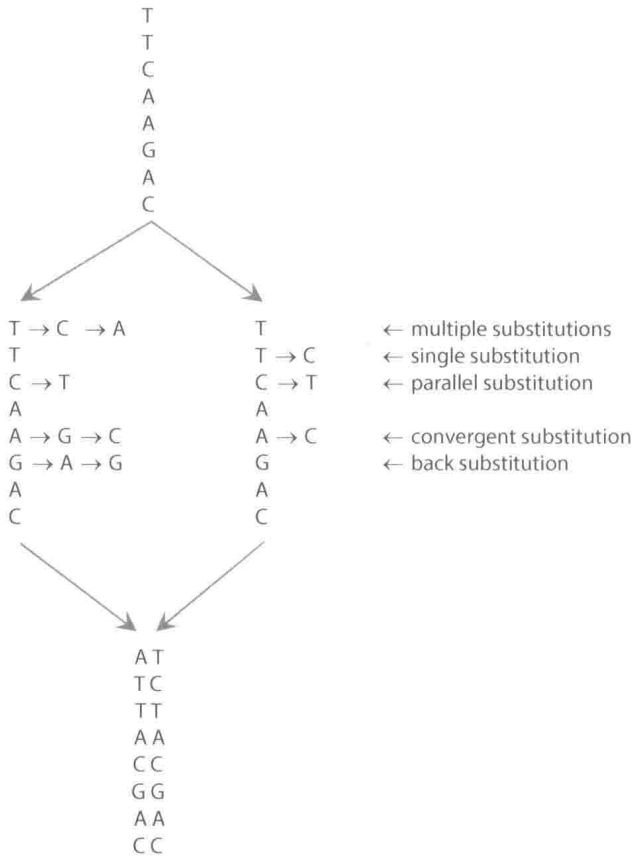


Fig. 1.1 Illustration of multiple substitutions at the same site or multiple hits. An ancestral sequence has diverged into two sequences and has since accumulated nucleotide substitutions independently along the two lineages. Only two *differences* are observed between the two present-day sequences, so that the proportion of different sites is $\hat{p} = 2/8 = 0.25$, while in fact as many as 10 *substitutions* (seven on the left lineage and three on the right lineage) occurred so that the true distance is $10/8 = 1.25$ substitutions per site. Constructed following Graur and Li (2000).

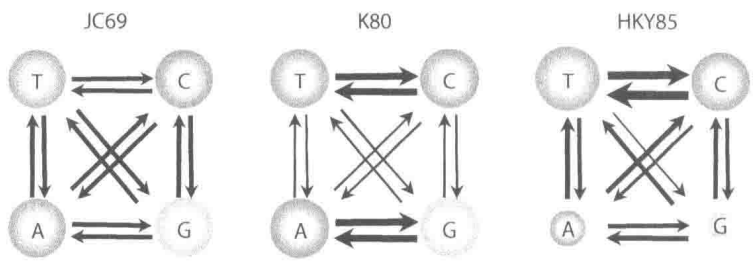


Fig. 1.2 Relative substitution rates between nucleotides under three Markov chain models of nucleotide substitution: JC69, K80, and HKY85. The thickness of the lines represents the substitution rates, while the sizes of the circles represent the steady-state distribution.