

Studies of evolution at the molecular level have experienced phenomenal growth in the last few decades, due to rapid accumulation of genetic sequence data, improved computer hardware and software, and the development of sophisticated analytical methods. The flood of genomic data has generated an acute need for powerful statistical methods and efficient computational algorithms to enable their effective analysis and interpretation.

*Molecular Evolution: A Statistical Approach* presents and explains modern statistical methods and computational algorithms for the comparative analysis of genetic sequence data in the fields of molecular evolution, molecular phylogenetics, statistical phylogeography, and comparative genomics. Written by an expert in the field, the book emphasizes conceptual understanding rather than mathematical proofs. The text is enlivened with 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. The coverage of maximum likelihood and Bayesian methods are in particular up-to-date, comprehensive, and authoritative.

This advanced textbook is aimed at graduate level students and professional researchers (both empiricists and theoreticians) in the fields of bioinformatics and computational biology, statistical genomics, evolutionary biology, molecular systematics, and population genetics. It will also be of relevance and use to a wider audience of applied statisticians, mathematicians, and computer scientists working in computational biology.

Cover images: Lizard photograph courtesy of Charles W. Linkem Ph.D.  
(Detail from) Tree of Life Image © itol.embl.de

**OXFORD**  
UNIVERSITY PRESS  
[www.oup.com](http://www.oup.com)

ISBN 978-0-19-960261-2



9 780199 602612

OXFORD

**Molecular Evolution**  
A STATISTICAL APPROACH

Yang

Ziheng Yang

# Molecular Evolution

A STATISTICAL APPROACH

OXFORD



# **Molecular Evolution**

## *A Statistical Approach*

ZIHENG YANG

**OXFORD**  
UNIVERSITY PRESS

*Molecular Evolution: A Statistical Approach.* Ziheng Yang. © Ziheng Yang 2014.  
Published 2014 by 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

© Ziheng Yang 2014

The moral rights of the author have been asserted

First Edition published in 2014

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in  
a retrieval system, or transmitted, in any form or by any means, without the  
prior permission in writing of Oxford University Press, or as expressly permitted  
by law, by licence or under terms agreed with the appropriate reprographics  
rights organization. Enquiries concerning reproduction outside the scope of the  
above should be sent to the Rights Department, Oxford University Press, at the  
address above

You must not circulate this work in any other form  
and you must impose this same condition on any acquirer

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

Links to third party websites are provided by Oxford in good faith and  
for information only. Oxford disclaims any responsibility for the materials  
contained in any third party website referenced in this work.

# 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

vi FOREWORD

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*  
*September 2013*



# 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

## viii PREFACE

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](mailto: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*

*September 2013*

# Contents

---

<b>1 Models of nucleotide substitution</b>	<b>1</b>
1.1 Introduction	1
1.2 Markov models of nucleotide substitution and distance estimation	4
1.2.1 The JC69 model	4
1.2.2 The K80 model	7
1.2.3 HKY85, F84, TN93, etc.	9
1.2.4 The transition/transversion rate ratio	13
1.3 Variable substitution rates across sites	15
1.4 Maximum likelihood estimation of distance	17
1.4.1 The JC69 model	18
1.4.2 The K80 model	22
1.4.3 Likelihood ratio test of substitution models	22
*1.4.4 Profile and integrated likelihood methods	24
1.5 Markov chains and distance estimation under general models	26
1.5.1 Markov chains	26
*1.5.2 Distance under the unrestricted (UNREST) model	27
*1.5.3 Distance under the general time-reversible model	29
1.6 Discussions	32
1.6.1 Distance estimation under different substitution models	32
1.6.2 Limitations of pairwise comparison	32
1.7 Problems	33
<b>2 Models of amino acid and codon substitution</b>	<b>35</b>
2.1 Introduction	35
2.2 Models of amino acid replacement	35
2.2.1 Empirical models	35
2.2.2 Mechanistic models	39
2.2.3 Among-site heterogeneity	39
2.3 Estimation of distance between two protein sequences	40
2.3.1 The Poisson model	40
2.3.2 Empirical models	41
2.3.3 Gamma distances	41
2.4 Models of codon substitution	42
2.4.1 The basic model	42
2.4.2 Variations and extensions	44
2.5 Estimation of $d_S$ and $d_N$	47
2.5.1 Counting methods	47
2.5.2 Maximum likelihood method	55



## x CONTENTS

2.5.3	Comparison of methods	57
2.5.4	More distances and interpretation of the $d_N/d_S$ ratio	58
2.5.5	Estimation of $d_S$ and $d_N$ in comparative genomics	61
*2.5.6	Distances based on the physical-site definition	63
*2.5.7	Utility of the distance measures	65
*2.6	Numerical calculation of the transition probability matrix	65
2.7	Problems	68
<b>3</b>	<b>Phylogeny reconstruction: overview</b>	<b>70</b>
3.1	Tree concepts	70
3.1.1	Terminology	70
3.1.2	Species trees and gene trees	79
3.1.3	Classification of tree reconstruction methods	81
3.2	Exhaustive and heuristic tree search	82
3.2.1	Exhaustive tree search	82
3.2.2	Heuristic tree search	82
3.2.3	Branch swapping	84
3.2.4	Local peaks in the tree space	86
3.2.5	Stochastic tree search	88
3.3	Distance matrix methods	88
3.3.1	Least-squares method	89
3.3.2	Minimum evolution method	91
3.3.3	Neighbour-joining method	91
3.4	Maximum parsimony	95
3.4.1	Brief history	95
3.4.2	Counting the minimum number of changes on a tree	95
3.4.3	Weighted parsimony and dynamic programming	96
3.4.4	Probabilities of ancestral states	99
3.4.5	Long-branch attraction	99
3.4.6	Assumptions of parsimony	100
3.5	Problems	101
<b>4</b>	<b>Maximum likelihood methods</b>	<b>102</b>
4.1	Introduction	102
4.2	Likelihood calculation on tree	102
4.2.1	Data, model, tree, and likelihood	102
4.2.2	The pruning algorithm	103
4.2.3	Time reversibility, the root of the tree, and the molecular clock	107
4.2.4	A numerical example: phylogeny of apes	108
4.2.5	Amino acid, codon, and RNA models	110
*4.2.6	Missing data, sequence errors, and alignment gaps	110
4.3	Likelihood calculation under more complex models	114
4.3.1	Mixture models for variable rates among sites	114
4.3.2	Mixture models for pattern heterogeneity among sites	122
4.3.3	Partition models for combined analysis of multiple datasets	123
4.3.4	Nonhomogeneous and nonstationary models	125

4.4	Reconstruction of ancestral states	125
4.4.1	Overview	125
4.4.2	Empirical and hierarchical Bayesian reconstruction	127
*4.4.3	Discrete morphological characters	130
4.4.4	Systematic biases in ancestral reconstruction	131
*4.5	Numerical algorithms for maximum likelihood estimation	133
*4.5.1	Univariate optimization	134
*4.5.2	Multivariate optimization	136
4.6	ML optimization in phylogenetics	138
4.6.1	Optimization on a fixed tree	138
4.6.2	Multiple local peaks on the likelihood surface for a fixed tree	139
4.6.3	Search in the tree space	140
4.6.4	Approximate likelihood method	143
4.7	Model selection and robustness	144
4.7.1	Likelihood ratio test applied to rbcL dataset	144
4.7.2	Test of goodness of fit and parametric bootstrap	146
*4.7.3	Diagnostic tests to detect model violations	147
4.7.4	Akaike information criterion (AIC and AIC <sub>c</sub> )	148
4.7.5	Bayesian information criterion	149
4.7.6	Model adequacy and robustness	150
4.8	Problems	151
<b>5</b>	<b>Comparison of phylogenetic methods and tests on trees</b>	<b>153</b>
5.1	Statistical performance of tree reconstruction methods	153
5.1.1	Criteria	154
5.1.2	Performance	156
5.2	Likelihood	157
5.2.1	Contrast with conventional parameter estimation	157
5.2.2	Consistency	158
5.2.3	Efficiency	159
5.2.4	Robustness	163
5.3	Parsimony	165
5.3.1	Equivalence with misbehaved likelihood models	165
5.3.2	Equivalence with well-behaved likelihood models	168
5.3.3	Assumptions and justifications	169
5.4	Testing hypotheses concerning trees	171
5.4.1	Bootstrap	172
5.4.2	Interior-branch test	177
5.4.3	K-H test and related tests	178
5.4.4	Example: phylogeny of apes	179
5.4.5	Indexes used in parsimony analysis	180
5.5	Problems	181
<b>6</b>	<b>Bayesian theory</b>	<b>182</b>
6.1	Overview	182
6.2	The Bayesian paradigm	183

## xii CONTENTS

6.2.1	The Bayes theorem	183
6.2.2	The Bayes theorem in Bayesian statistics	184
*6.2.3	Classical versus Bayesian statistics	189
6.3	Prior	197
6.3.1	Methods of prior specification	197
6.3.2	Conjugate priors	198
6.3.3	Flat or uniform priors	199
*6.3.4	The Jeffreys priors	200
*6.3.5	The reference priors	202
6.4	Methods of integration	203
*6.4.1	Laplace approximation	203
6.4.2	Mid-point and trapezoid methods	204
6.4.3	Gaussian quadrature	205
6.4.4	Marginal likelihood calculation for JC69 distance estimation	206
6.4.5	Monte Carlo integration	210
6.4.6	Importance sampling	210
6.5	Problems	212
<b>7</b>	<b>Bayesian computation (MCMC)</b>	<b>214</b>
7.1	Markov chain Monte Carlo	214
7.1.1	Metropolis algorithm	214
7.1.2	Asymmetrical moves and proposal ratio	218
7.1.3	The transition kernel	219
7.1.4	Single-component Metropolis–Hastings algorithm	220
7.1.5	Gibbs sampler	221
7.2	Simple moves and their proposal ratios	221
7.2.1	Sliding window using the uniform proposal	222
7.2.2	Sliding window using the normal proposal	223
7.2.3	Bactrian proposal	223
7.2.4	Sliding window using the multivariate normal proposal	224
7.2.5	Proportional scaling	225
7.2.6	Proportional scaling with bounds	226
7.3	Convergence, mixing, and summary of MCMC	226
7.3.1	Convergence and tail behaviour	226
7.3.2	Mixing efficiency, jump probability, and step length	230
7.3.3	Validating and diagnosing MCMC algorithms	241
7.3.4	Potential scale reduction statistic	242
7.3.5	Summary of MCMC output	243
7.4	Advanced Monte Carlo methods	244
7.4.1	Parallel tempering (MC <sup>3</sup> )	245
7.4.2	Trans-model and trans-dimensional MCMC	247
7.4.3	Bayes factor and marginal likelihood	256
7.5	Problems	260
<b>8</b>	<b>Bayesian phylogenetics</b>	<b>263</b>
8.1	Overview	263
8.1.1	Historical background	263

8.1.2	A sketch MCMC algorithm	264
8.1.3	The statistical nature of phylogeny estimation	264
8.2	Models and priors in Bayesian phylogenetics	266
8.2.1	Priors on branch lengths	266
8.2.2	Priors on parameters in substitution models	269
8.2.3	Priors on tree topology	276
8.3	MCMC proposals in Bayesian phylogenetics	279
8.3.1	Within-tree moves	279
8.3.2	Cross-tree moves	281
8.3.3	NNI for unrooted trees	284
8.3.4	SPR for unrooted trees	287
8.3.5	TBR for unrooted trees	289
8.3.6	Subtree swapping	291
8.3.7	NNI for rooted trees	292
8.3.8	SPR on rooted trees	293
8.3.9	Node slider	294
8.4	Summarizing MCMC output	295
8.5	High posterior probabilities for trees	296
8.5.1	High posterior probabilities for trees or splits	296
8.5.2	Star tree paradox	298
*8.5.3	Fair coin paradox, fair balance paradox, and Bayesian model selection	300
8.5.4	Conservative Bayesian phylogenetics	305
8.6	Problems	306
<b>9</b>	<b>Coalescent theory and species trees</b>	<b>308</b>
9.1	Overview	308
9.2	The coalescent model for a single species	309
9.2.1	The backward time machine	309
9.2.2	Fisher–Wright model and the neutral coalescent	309
9.2.3	A sample of $n$ genes	312
9.2.4	Simulating the coalescent	315
9.2.5	Estimation of $\theta$ from a sample of DNA sequences	316
9.3	Population demographic process	320
9.3.1	Homogeneous and nonhomogeneous Poisson processes	321
9.3.2	Deterministic population size change	322
9.3.3	Nonparametric population demographic models	323
9.4	Multispecies coalescent, species trees and gene trees	325
9.4.1	Multispecies coalescent	325
9.4.2	Species tree–gene tree conflict	331
9.4.3	Estimation of species trees	335
9.4.4	Migration	343
9.5	Species delimitation	349
9.5.1	Species concept and species delimitation	349
9.5.2	Simple methods for analysing genetic data	351
9.5.3	Bayesian species delimitation	352

## xiv CONTENTS

9.5.4	The impact of guide tree, prior, and migration	355
9.5.5	Pros and cons of Bayesian species delimitation	358
9.6	Problems	359
<b>10</b>	<b>Molecular clock and estimation of species divergence times</b>	<b>361</b>
10.1	Overview	361
10.2	Tests of the molecular clock	363
10.2.1	Relative-rate tests	363
10.2.2	Likelihood ratio test	364
10.2.3	Limitations of molecular clock tests	365
10.2.4	Index of dispersion	366
10.3	Likelihood estimation of divergence times	366
10.3.1	Global clock model	366
10.3.2	Local clock model	367
10.3.3	Heuristic rate-smoothing methods	368
10.3.4	Uncertainties in calibrations	370
10.3.5	Dating viral divergences	372
10.3.6	Dating primate divergences	373
10.4	Bayesian estimation of divergence times	375
10.4.1	General framework	375
10.4.2	Approximate calculation of likelihood	376
10.4.3	Prior on evolutionary rates	377
10.4.4	Prior on divergence times and fossil calibrations	378
10.4.5	Uncertainties in time estimates	382
10.4.6	Dating viral divergences	384
10.4.7	Application to primate and mammalian divergences	385
10.5	Perspectives	388
10.6	Problems	389
<b>11</b>	<b>Neutral and adaptive protein evolution</b>	<b>390</b>
11.1	Introduction	390
11.2	The neutral theory and tests of neutrality	391
11.2.1	The neutral and nearly neutral theories	391
11.2.2	Tajima's <i>D</i> statistic	393
11.2.3	Fu and Li's <i>D</i> , and Fay and Wu's <i>H</i> statistics	394
11.2.4	McDonald–Kreitman test and estimation of selective strength	395
11.2.5	Hudson–Kreitman–Aquadre test	397
11.3	Lineages undergoing adaptive evolution	398
11.3.1	Heuristic methods	398
11.3.2	Likelihood method	399
11.4	Amino acid sites undergoing adaptive evolution	400
11.4.1	Three strategies	400
11.4.2	Likelihood ratio test of positive selection under random-site models	402
11.4.3	Identification of sites under positive selection	405
11.4.4	Positive selection at the human MHC	406

11.5	Adaptive evolution affecting particular sites and lineages	408
11.5.1	Branch-site test of positive selection	408
11.5.2	Other similar models	409
11.5.3	Adaptive evolution in angiosperm phytochromes	410
11.6	Assumptions, limitations, and comparisons	411
11.6.1	Assumptions and limitations of current methods	412
11.6.2	Comparison of methods for detecting positive selection	413
11.7	Adaptively evolving genes	414
11.8	Problems	416
<b>12</b>	<b>Simulating molecular evolution</b>	<b>418</b>
12.1	Introduction	418
12.2	Random number generator	418
12.3	Generation of discrete random variables	420
12.3.1	Inversion method for sampling from a general discrete distribution	420
12.3.2	The alias method for sampling from a discrete distribution	421
12.3.3	Discrete uniform distribution	422
12.3.4	Binomial distribution	423
12.3.5	The multinomial distribution	423
12.3.6	The Poisson distribution	423
12.3.7	The composition method for mixture distributions	424
12.4	Generation of continuous random variables	424
12.4.1	The inversion method	425
12.4.2	The transformation method	425
12.4.3	The rejection method	425
12.4.4	Generation of a standard normal variate using the polar method	428
12.4.5	Gamma, beta, and Dirichlet variables	430
12.5	Simulation of Markov processes	430
12.5.1	Simulation of the Poisson process	430
12.5.2	Simulation of the nonhomogeneous Poisson process	431
12.5.3	Simulation of discrete-time Markov chains	433
12.5.4	Simulation of continuous-time Markov chains	435
12.6	Simulating molecular evolution	436
12.6.1	Simulation of sequences on a fixed tree	436
12.6.2	Simulation of random trees	439
12.7	Validation of the simulation program	439
12.8	Problems	440
	<b>Appendices</b>	<b>442</b>
	Appendix A. Functions of random variables	442
	Appendix B. The delta technique	446
	Appendix C. Phylogenetic software	448
	<i>References</i>	450
	<i>Index</i>	488