Article details

Article title: Molecular Phylogenetics
Article ID: 9780199941728-0098
Article author(s): Ziheng Yang
Publishing Group: Reference-US

☐Revision

Substantive Update: Y/N
Heavily Revised article (with new author(s)): Y/N
Title revised? Y/N
Previous title:

Table of contents:

Module details

Module: Evolutionary Biology **Module code:** **Module ISBN:** 9780199941728
EVB

Style and XML details

Citation style: Scientific
Special characters/fonts/elements:
Notes to Copyeditor:

Format neutral

Note: All content will be format neutral compliant, unless otherwise indicated here

Exceptions:

- Oxford Bibliographies articles do not currently have keywords. Introductions serve as abstracts.

Additional Notes to Copyeditor
NONE.

Molecular Phylogenetics

INTRODUCTION

Molecular phylogenetics is the science of using molecular data (DNA and protein sequences) to infer the phylogenetic relationships among species. While traditionally morphological characters were used to infer species phylogenies, nowadays molecules are the dominating type of data for nearly all species groups. Both the development of statistical methods for phylogenetic reconstruction and divergence time estimation and the use of such methods to infer species phylogenies are major undertakings of the field.

BOOKS AND REVIEWS

A number of textbooks have been written. Some are written for biologists with minimal requirement for statistics while others focus on the statistical methods and require a basic knowledge of calculus, linear algebra, probability theory, and mathematical statistics. These volumes can be read by applied statisticians, bioinformaticians, computational biologists, and empirical biologists who are comfortable with quantitative arguments.

Biologically Oriented Textbooks

Page and Holmes 1998 is one of the earliest textbooks. It takes a phylogenetic approach to molecular evolution and explains the basic ideas of phylogenetics very well. Nei and Kumar 2000 describes simple methods of phylogeny reconstruction such as parsimony and distance methods. Bromham 2016 takes an equation-free approach to molecular phylogenetics and evolution, so the book may be useful for generating interest among biology or genomics students in molecular phylogenetics. Graur 2016 is an introductory text focused on the biological questions of the field. This volume covers the same ground as Li 1997 but includes more up-to-date materials. Modern phylogenetic analyses are dominated by large data sets and sophisticated statistical methods such as maximum likelihood and Bayesian inference. Coverage of likelihood and Bayesian methods in those books is either absent or minimal.

Bromham, L. 2016. *An introduction to molecular evolution and phylogenetics*. Oxford: Oxford Univ. Press. [ISBN: 9780198736363]

This is an elementary equation-free textbook written for biologists.

Graur, D. 2016. *Molecular and genome evolution*. Sunderland, MA: Sinauer. [ISBN: 9781605354699]

This introductory textbook does not assume much prerequisite knowledge of molecular biology, evolution, or mathematics. Chapter 5 covers phylogeny-reconstruction methods.

Li, W. -H. 1997. *Molecular evolution*. Sunderland, MA: Sinauer. [ISBN: 9780878934638]

This book presents a synthesis of evolutionary studies at the molecular level up to 1997, including a chapter on reconstructing molecular phylogenies.

Nei, M., and S. Kumar. 2000. *Molecular evolution and phylogenetics*. Oxford: Oxford Univ. Press. [ISBN: 9780195135855]

This book includes a detailed coverage of distance and parsimony methods of phylogeny reconstruction at an elementary level. It is particularly useful for users of the MEGA software.

Page, R. D. M., and E. C. Holmes. 1998. *Molecular evolution: A phylogenetic approach*. Oxford: Blackwell Science. [ISBN: 9780865428898]

This introductory text illustrates the basic concepts of molecular phylogenetics very clearly, with lots of diagrams and examples.

Statistically Oriented Textbooks

Hillis, et al. 1996 is an early book on molecular phylogenetics. For about ten years, the chapters on substitution models and on phylogeny reconstruction were the major source for learning statistical phylogenetics. Nielsen 2006 and Gascuel 2005 are two edited books focusing on mathematical models and statistical methods in molecular phylogenetics. Similarly Balding, et al. 2007 includes a number of chapters on topics of molecular phylogenetics. Felsenstein 2004 is extremely comprehensive and discusses nearly everything related to phylogenies. Yang 2014 covers statistical models and methods (in particular, likelihood and Bayesian methods) in greater depth. Semple and Steel 2003 is a mathematical treatment of phylogenetics and discusses many mathematical and combinatorial problems inspired by the parsimony method of phylogeny reconstruction.

Balding, D., M. Bishop, and C. Cannings. 2007. *Handbook of statistical genetics*. 3d ed. Hoboken, NJ: John Wiley. [ISBN: 9780470058305]

Written by leaders in the diverse field of statistical genetics, this two-volume handbook includes a number of chapters that cover topics in molecular phylogenetics, such as nucleotide and amino acid substitution models, adaptive protein evolution, phylogeny reconstruction, and molecular clock dating.

Felsenstein, J. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer. [ISBN: 9780878931774]

This encyclopedic book discusses nearly everything related to phylogenies. The book includes historical background and anecdotes concerning the early development of the field, which make it particularly interesting reading. The book discusses statistical models and methods of phylogeny reconstruction in great detail, including distance methods, maximum likelihood, and bootstrap. Its coverage of Bayesian methods is now somewhat out of date, since most of the Bayesian methods and programs were developed after the publication of the book.

Gascuel, O., ed. 2005. *Mathematics of evolution and phylogeny*. Oxford: Oxford Univ. Press. [ISBN: 9780199231348]

This edited book provides an introduction to models of nucleotide and amino acid substitution, including a chapter on Bayesian phylogenetics.

Hillis, D. M., C. Moritz, and B. K. Mable, eds. 1996. *Molecular systematics*. Sunderland, MA: Sinauer. [ISBN: 9780878932825]

This is an important early book for modern phylogenetics and systematics. Chapter 11, on phylogenetic inference by Swofford, et al., provides a summary of statistical methods, including maximum likelihood. Bayesian methods were developed after the publication of this book.

the reader from basic introductory material to state-of the-art statistical methods.

Nielsen, R., ed. 2006. *Statistical methods in molecular evolution*. New York: Springer. [ISBN: 9781441919724]

This edited book reviews probabilistic models and statistical methods for phylogenetic comparative analysis of genetic sequence data. It covers advanced computational methods, such as maximum likelihood optimization and Bayesian Markov Chain Monte Carlo. The book includes an introductory section, suitable for readers new to the field. The chapters are written by the leaders in the field and they will take

Semple, C., and M. Steel. 2003. *Phylogenetics*. New York: Oxford Univ. Press. [ISBN: 9780198509424]

This book is written for mathematicians and discusses mathematical and combinatorial problems related to phylogenetic trees, inspired by phylogenetic parsimony. The book does not cover statistical methods of data analysis, such as maximum likelihood and Bayesian methods.

Yang, Z. 2014. *Molecular evolution: A statistical approach*. Oxford: Oxford Univ. Press. [ISBN: 9780199602605]

This is an expanded and updated edition of Yang's *Computational Molecular Evolution*, (Oxford: Oxford University Press, 2006). It summarizes the Markov models of nucleotide, amino acid and codon substitution, statistical methods and computational algorithms of phylogeny reconstruction. It includes such topics as molecular clock dating, detection of molecular adaptation, and computer simulation. This is the only book currently available that includes a comprehensive coverage of Bayesian phylogenetics.

Software-Oriented Books

A number of computer software packages have been developed in the field of molecular phylogenetics and evolution. A few books have been written to guide beginners to the important packages. Hall 2011 is a how-to manual for beginners with no experience in phylogenetic tree reconstruction. Lemey, et al. 2009 is written by software developers and achieves the same goal at a more advanced level.

Hall, B. G. 2011. *Phylogenetic trees made easy: A how-to manual*. 4th ed. Sunderland, MA: Sinauer. [ISBN: 9780878936069]

This simple book helps the reader to get started in creating phylogenetic trees from protein or DNA sequence data. It is aimed at molecular and cell biologists, who may not be familiar with phylogenetics or evolutionary theory. Software packages covered include CLUSTAL, MEGA, MrBayes, and codeml.

Lemey, P., M. Salemi, and A. -M. Vandamme, eds. 2009. *The phylogenetic handbook*. Cambridge, UK: Cambridge Univ. Press. [ISBN: 9780521877107]

This book is a hands-on guide to both the theory and the practice of molecular phylogenetic analysis. A rich collection of software packages is covered, with many of them described by their original authors. Most chapters include a theory section followed by a software practice section. The book may be useful for teaching advanced-level undergraduate and graduate students. Written by different authors, the chapters are somewhat uneven.

Books on Special Topics

Several edited books deal with major research areas in molecular phylogenetics and evolution. Rosenberg 2009 discusses multiple sequence alignment, which is often the first step in a phylogenetic analysis. Liberles 2010 treats ancestral sequence reconstruction, the inference of sequences in extinct ancestors on the phylogeny given the observed sequences for the modern

species. Cannarozzi and Schneider 2012 summarizes modern developments in codon-substitution models (introduced in 1994). Chen, et al. 2014 deals with recent advancements in the Bayesian approach to molecular phylogenetics, since their introduction around 1996. Knowles and Kubatko 2010 discusses species tree estimation in the presence of gene-tree conflicts, which is becoming increasingly important in analysis of modern phylogenomic data.

Cannarozzi, G., and A. Schneider. 2012. *Codon evolution: Mechanisms and models*. New York: Oxford Univ. Press. [ISBN: 9780199601165]

This edited book summarizes recent developments in codon-substitution models and their use to detect adaptive evolution at the molecular level.

Chen, M. -H., L. Kuo, and P. Lewis. 2014. *Bayesian phylogenetics: Methods, algorithms, and applications*. Boca Raton, FL: CRC. [ISBN: 9781466500792]

This edited book summarizes recent developments in the Bayesian approach to phylogenetic analysis. Several chapters deal with modern computational algorithms (such as path sampling) for calculating Bayes factors for model comparison.

Knowles, L. L., and L. S. Kubatko. 2010. *Estimating species trees: Practical and theoretical aspects*. Oxford: Wiley-Blackwell. [ISBN: 9780470526859]

This edited book discusses species tree estimation using multi-locus data when the gene trees from the different loci differ from each and from the species tree. It describes species tree estimation under the multispecies coalescent model, which accommodates species tree-gene tree conflicts, as a new phylogenetic paradigm for the 21st century. This field has been advancing fast, and many methods have been developed since the publication of the book.

Liberles, D. A., ed. 2010. *Ancestral sequence reconstruction*. New York: Oxford Univ. Press. [ISBN: 9780199299188]

This edited book discusses the methods and applications of ancestral sequence reconstruction. It includes examples of practical applications as well as the theoretical and experimental detail.

Rosenberg, M. S., ed. 2009. *Sequence alignment: Methods, models, concepts, and strategies*. Berkeley: Univ. of California Press. [ISBN: 9780520256972]

Sequence alignment is a complicated and perhaps under-appreciated aspect of molecular phylogenetics. This book provides a gentle introduction to different aspects of multiple sequence alignment, such as global and local alignment algorithms, assessment of alignment quality, and alignment software. The book does not discuss statistical alignment, which uses models of insertions and deletions as well as substitutions to evaluate different alignments. Statistical alignment has great theoretical appeal but has not produced useful software, so that, in practice, alignment is conducted using heuristic methods.

Reviews

A number of reviews have been published on various aspects of phylogenetic analysis. We mention only a few recent ones. Whelan, et al. 2001 reviews Markov models of sequence evolution. Goldman, et al. 2000 summarizes statistical tests of phylogenies in the likelihood framework. Yang and Rannala 2012 is a beginner's introduction to molecular phylogenetics. Several reviews have been published on the molecular clock and its use to date species divergences, including Kumar 2005, Bromham and Penny 2003, and dos Reis 2016. Huelsenbeck, et al. 2001 is a widely read review on Bayesian phylogenetics.

Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* 4:216–224.

This review discusses the molecular clock, factors that cause the rate variation (violation of the clock), and the use of the clock on divergence time estimation.

dos Reis, M., P. C. J. Donoghue, and Z. Yang. 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* 17:71–80.

This review of molecular clock dating focuses on the Bayesian method, which is the dominating method for integrating information from fossils and molecules. It paints a gloomy picture of clock dating, marred by uncertainties in fossil calibrations and the confounding effects of time and rate in molecular sequence comparisons.

Goldman, N., J. Anderson, and A. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* 49.4: 652–670.

This review discusses likelihood-based tests for comparing phylogenetic trees, including the Kishino-Hasegawa test, Shimodaira-Hasegawa test, and parametric bootstrap.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294.5550: 2310–2314.

This widely read review on the Bayesian approach to phylogenetics served well to increase the visibility of the method in the field.

Kumar, S. 2005. Molecular clocks: Four decades of evolution. *Nature Reviews Genetics* 6:654–662.

This review of the molecular clock provides a chronology of molecular clock studies.

Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends in Genetics* 17.5: 262–272.

This is a highly readable review on models of nucleotide substitution and methods of phylogenetic reconstruction.

Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics* 13:303–314.

This is a basic introduction to concepts in molecular phylogenetics.

JOURNALS

****Systematic Biology**** and ****Molecular Biology and Evolution **** are the two leading journals devote to molecular phylogenetics and evolution. ****Molecular Phylogenetics and Evolution**** and the ****Journal of Molecular Evolution**** also publish mostly studies in the field, as are the review journals ****Trends in Ecology & Evolution**** and ****Annual Review of Ecology, Evolution, and Systematics****. Comparative analysis of molecular data has permeated nearly all branches of biological sciences. As a result, many generalist journals regularly publish papers in molecular phylogenetics, such as *Proceedings of the National Academy of Sciences of the United States of America*, *Nature*, *Science*, and *Cell*. In addition, specialist journals such as ****Genetics****, ****Genome Research****, ****Evolution ****, ****Molecular Ecology****, ****Philosophical Transactions of the Royal Society of London: Series B****, ****Proceedings B: Biological Sciences****, ****Current Biology****, ****Biology Letters****, and ****Virology**** regularly publish studies in the field of molecular phylogenetics. Methodological papers are frequently published in ****Bioinformatics****, ****BMC Bioinformatics****, and ****Journal of Computational Biology****, or even statistics journals.

**Annual Review of Ecology, Evolution, and Systematics*[http://www.annualreviews.org/journal/ecolsys]*. 2003–. [class:periodical]

This annual review journal publishes articles in the fields of ecology, evolutionary biology, and systematics, with topics ranging from phylogeny, speciation, and molecular evolution, among others. In 1970 published as the *Annual Review of Ecology and Systematics*.

**Bioinformatics*[http://academic.oup.com/bioinformatics]*. 1998–. [class:periodical]
This journal regularly publishes statistical methods and computational algorithms for phylogenetic analysis of molecular data.

**BMC Evolutionary Biology*[http://bmcevolbiol.biomedcentral.comwww.biomedcentral.com/bmcevolbiol]*. 2001–. [class:periodical]
This is an open-access journal that publishes on molecular evolution and phylogenetics.

**Genome Biology and Evolution*[http://gbe.oxfordjournals.org/]*. 2009–. [class:periodical]
This is the online journal of the Society for Molecular Biology and Evolution. While having a genomic focus, this journal regularly publishes papers on phylogenetic analysis of genomic data.

**Journal of Molecular Evolution*[http://link.springer.com/journal/239]*. 1971–. [class:periodical]
This is the earliest journal devoted to the field of molecular evolution and continues to publish papers on both methodological developments and biological discoveries in molecular evolution and phylogenetics.

**Molecular Biology and Evolution*[http://mbe.oxfordjournals.org/]*. 1983–. [class:periodical]
This is a journal of the Society for Molecular Biology and Evolution, which publishes research at the interface of molecular biology and evolutionary biology, including molecular phylogenetics.

**Molecular Phylogenetics and Evolution*[http://www.journals.elsevier.com/molecular-phylogenetics-and-evolution]*. 1992–. [class:periodical]
This journal publishes papers on methodologies and applications of molecular phylogenetics.

**Nature Reviews Genetics*[http://www.nature.com/nrg]*. 2000–. [class:periodical]
This is a high-profile review journal that publishes on genetics, including molecular phylogenetics and evolution.

**Systematic Biology*[http://sysbio.oxfordjournals.org/]*.1952–. [class:periodical]
This is the bimonthly journal of the Society of Systematic Biologists. Its main focus is phylogenetics and systematics.

**Trends in Ecology & Evolution*[http://www.cell.com/trends/ecology-evolution/home]*. 1980–. [class:periodical]
This is a leading review journal devoted to ecology and evolution.

**Virus Evolution*[http://academic.oup.com/ve]*. 2015–. [class:periodical]
This is a new open-access journal started in 2015 focusing on the evolution of viruses, or viruses as a model system for studying the evolutionary process.

HISTORY OF MOLECULAR PHYLOGENETICS

The appearance of protein sequences in the late 1950s and early 1960s heralded the beginning of the field of molecular phylogenetics. Indeed, the authors of Zuckerkandl and Pauling 1962 and Zuckerkandl and Pauling 1965 recognized that macromolecules (protein and DNA sequences) are the best kind of data for inferring the evolutionary history among species. The major methods of phylogenetic reconstruction, including parsimony (minimum evolution), and distance and likelihood methods, were already developed by the early 1970s. Even the ideas for the Bayesian

approach to phylogenetics are found in Edwards 1970 (cited under *Maximum Likelihood*). Felsenstein 2004 provides much material for the early history of molecular phylogenetics, and Edwards 2009 gives a personal account of the origin of statistical methods, including minimum evolution, least squares, and maximum likelihood.

Edwards, A. W. F. 2009. Statistical methods for evolutionary trees. *Genetics* 183.1: 5–12.
This article recounts the exciting origin of the statistical approach to molecular phylogenetics. In the 1960s, Cavali-Sforza and Edwards made an effort to adapt R. A. Fisher's maximum likelihood method to the problem of phylogeny reconstruction. They were not very successful. In the meantime, they invented the parsimony (minimum evolution) and distance methods. The origin of the Bayesian method may also be traced this effort.

Felsenstein, J. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer. [ISBN: 9780878931774]
This book includes many historical accounts and anecdotes from the early history of the field.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in biochemistry*. Edited by M. Kasha and B. Pullman, 189–225. New York: Academic Press.
This chapter includes the first molecular-clock dating analysis, even though the term "molecular evolutionary clock" was first used in the 1965 paper.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology* 8.2: 357–366
This classic paper is highly readable and relevant even today. The excitement at the newly emerging data and field is palpable. Among many of its insightful discussions, we found the following concerning the utility of molecular data to phylogenetics: "We may ask the questions where in the now living systems the greatest amount of history has survived and how it can be extracted. . . . Best fit are the different types of macromolecules (sequences) which carry the genetic information." (p. 357). The term *molecular evolutionary clock* was coined in this paper.

Minimum-Evolution Principle or Parsimony

The minimum-evolution principle or parsimony was proposed as a criterion for comparing different evolutionary histories using continuous characters in Edwards and Cavalli-Sforza 1963. This was suggested as an approximation to the maximum likelihood method, which the authors were attempting to develop. The authors of Camin and Sokal 1965 applies the principle to discrete morphological characters (that is, the parsimony principle of minimizing the number of changes), and they provide a philosophical justification. The philosophical-statistical debate was to develop into a major controversy that lasted well into the late 1990s. For molecular data, a number of pioneers of the field clearly found it natural to minimize the number of changes on the tree. For example, the authors of Pauling and Zuckerkandl 1963 did it to infer ancestral proteins for "paleogenetic" studies of their chemical properties. The authors of Eck and Dayhoff 1966 did it to construct empirical matrices of amino acid substitution rates. Walter Fitch was the first to present a systematic algorithm to enumerate the most parsimonious reconstructions of ancestral states (Fitch 1971). The statistician John Hartigan provided a mathematical proof and his algorithm accommodates multifurcating trees as well (Hartigan 1973). The early histories of phylogenetic methods are reviewed in Edwards 1996 and Felsenstein 2004 (cited under *History of Molecular Phylogenetics*).

Camin, J. H., and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19.3: 311–326.

This paper minimized the number of changes to infer the species phylogeny using discrete morphological characters. This is the minimum-evolution (parsimony) principle discussed in Edwards and Cavalli-Sforza 1963.

Eck, R. V., and M. O. Dayhoff. 1966. Inference from protein sequence comparisons. In *Atlas of protein sequence and structure*. Edited by M. O. Dayhoff. Silver Spring, MD: National Biomedical Research Foundation.

Dayhoff and Eck used a parsimony-style argument to count amino acid changes along branches of the phylogenetic tree, leading to the well-known PAM (for "point accepted mutations") matrices, which is now used in every likelihood or Bayesian phylogenetic program.

Edwards, A. W. F. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. *Systematic Biology* 45.1: 79–91.

As the title suggests, this article traces the origin of the parsimony (minimum-evolution) method of phylogenetic tree reconstruction.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1963. The reconstruction of evolution (Abstract). *Annals of Human Genetics* 27:105–106.

This abstract announces the minimum-evolution (parsimony) principle: "The most plausible estimate of the evolutionary tree is that which invokes the minimum net amount of evolution."

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* 20.4: 406–416.

This classic paper is remarkable in providing an algorithm for inferring the most parsimonious reconstructions of ancestral states given the tree and the observed character states at the tips of the tree. This is a version of the dynamical programming algorithm.

Hartigan, J. A. 1973. Minimum evolution fits to a given tree. *Biometrics* 29.1: 53–65.

This paper provides a dynamical programming algorithm for reconstructing ancestral states using parsimony, with mathematical proofs. The algorithm works on multifurcating trees as well as binary trees.

Pauling, L., and E. Zuckerkandl. 1963. Chemical paleogenetics: Molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica* 17:S9–S16.

This excellent paper foreshadows modern interest in ancestral sequence reconstruction. This is remarkable given that very few protein sequences and no DNA sequence had been determined at the time. The idea of parsimony counting of changes on the tree, to be formalized as the Fitch-Hartigan algorithm, is explicitly used.

Distance Methods

Besides proposing the minimum-evolution (parsimony) principle, Edwards and Cavalli-Sforza also initiated the statistical approach to phylogenetics. The authors of Cavalli-Sforza and Edwards 1967 develop the additive-tree method, which uses the least squares criterion to fit the branch lengths on the tree to the matrix of estimated pairwise distances. The least squares method was made popular in Fitch and Margoliash 1967. Edwards 2009 and Felsenstein 2004 (cited under *History of Molecular Phylogenetics*) provide reviews of the early history of distance methods. The most commonly used distance-matrix method of phylogeny reconstruction is the neighbor-joining method, developed in Saitou and Nei 1987.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution* 21.3: 550–570.

This classic paper is noteworthy in developing both the parsimony and the distance (least-squares) methods of phylogeny reconstruction. It discusses the maximum likelihood method as well, although its correct formation had to wait until Felsenstein 1973.

Edwards, A. W. F. 2009. Statistical methods for evolutionary trees. *Genetics* 183.1: 5–12.
This article relates the author's personal accounts of the origin of least squares and likelihood methods of phylogeny reconstruction.

Fitch, W. M., and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochemical Genetics* 1.1: 65–71.
The authors implemented the least-squares method for inferring a phylogenetic tree using protein sequences.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4.4: 406–425.
This paper describes the neighbor-joining method of phylogeny reconstruction, which may be the most commonly used phylogeny-reconstruction method based on citations.

Maximum Likelihood

In the early 1960s, Anthony Edwards and Luca Cavalli-Sforza, both working with R. A. Fisher, made an effort to apply Fisher's maximum likelihood method to estimate genealogical trees of human populations using gene frequency data. They used the Yule process (pure-birth process) to describe the probabilities of rooted trees and Brownian motion to describe the drift of gene frequencies over time. Nevertheless, they experienced difficulties in applying ML, including "singularities" in the likelihood function. The nature of the estimation problem was clarified in Edwards 1970, which points out that the tree has a distribution specified by the Yule process, and should be estimated from the conditional (posterior) distribution of the tree given the data. Edwards 1970 is thus closer to Bayesian than to maximum likelihood. Inspired by Alan Wilson, the statistician Jerzy Neyman (both working in Berkeley) considered the statistical problem of estimating phylogenetic trees (Neyman 1971). The first correct formulation of the problem in the likelihood framework is found in Felsenstein 1973, for discrete morphological characters, and in Felsenstein 1981, for DNA sequences. PhyML and RAxML are two fast programs for phylogenetic reconstruction by maximum likelihood, described in Guindon and Gascuel 2003 and Stamatakis, et al. 2005.

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process (with discussion). *Journal of the Royal Statistical Society: Series B* 32:155–174.
This remarkable paper, published in a statistics journal, clarifies the nature of the estimation problem of phylogeny reconstruction. If one takes the view that the branching process specifies a prior on the phylogenies, this paper may be considered the first attempt to apply the Bayesian approach to phylogenetics.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. In *Phenetic and phylogenetic classifications: A symposium, Univ. of Liverpool, 8 and 9 April 1964*. Edited by V. H. Heywood and J. MacNeill, 67–76. Systematics Association Publication 6. London: Systematics Association. [class:conference-paper]
This paper records Edwards and Cavalli-Sforza's initial attempt to apply maximum likelihood to phylogeny estimation. The data considered are gene frequencies for major blood groups in different human populations. The authors met with considerable challenges, part of which

stemmed from the mistake of treating random variables (tree and times that have distributions under the branching model) as parameters. This was recognized in Edwards 1970.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* 22.3: 240–249.
This paper introduces the pruning algorithm (the dynamic programming algorithm) for likelihood calculation on a phylogeny when the data consist of discrete characters from the modern species. This introduces the maximum likelihood method of phylogenetics.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17.6: 368–376.
This paper introduces the maximum likelihood method for phylogeny reconstruction using DNA sequence data. It discusses the pruning algorithm, time reversibility, and the pully principle. It also introduces the likelihood ratio test of the molecular clock.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52.5: 696–704.
This article describes the fast likelihood program PhyML.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics: Proceedings of a symposium held at Purdue Univ., 1970*. Edited by S. S. Gupta and J. Yackel, 1–27. New York: Academic Press. [class:conference-paper]
Jerzy Neyman is a statistician known for some basic statistical concepts such as confidence intervals, null and alternative hypotheses, type-I and type-II errors, and the Neyman-Pearson lemma. This paper is sometimes described as the first application of maximum likelihood to phylogenetics. While this may be debatable, Neyman is certainly correct to suggest that molecular phylogenetics is a source of novel statistical problems. Nowadays the hottest domain of statistics appears to be Bayesian, and it may be hard to find any new exciting Bayesian computational algorithms that have not been trialed in phylogenetics.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21.4: 456–463.
This article describes the fast likelihood program RAxML, which is particularly powerful on modern multicore computers.

Bayesian methods

The Bayesian approach to phylogenetics applied to sequence data was introduced by three groups working independently at about the same time in the mid 1990s: Rannala and Yang 1996; Rannala and Yang 1996; Mau and Newton 1997; and Li, et al. 2000. These early studies assumed the molecular clock and inferred rooted trees. Bayesian phylogenetics really took off after Huelsenbeck and Ronquist published their Bayesian MCMC program MrBayes (Huelsenbeck and Ronquist 2001). Later Drummond, et al. 2006 described the BEAST software, which works on rooted trees under clock and relaxed-clock models. The authors of Lartillot, et al. 2007 published the PhyloBayes program, which implements nonstationary models to deal with substitution heterogeneity among sites and among branches, which may be important for deep phylogenies. Yang 2014 (chapter 8) provides a comprehensive coverage of Bayesian MCMC algorithms in phylogenetics. Chen, et al. 2014 summarizes recent developments in Bayesian phylogenetics.

Chen, M. -H., L. Kuo, and P. Lewis. 2014. *Bayesian phylogenetics: Methods, algorithms, and applications*. Boca Raton, FL: CRC. [ISBN: 9781466500792]

This edited book summarizes recent developments in Bayesian phylogenetics, in particular on the calculation and use of Bayes factors for model selection.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. *Relaxed phylogenetics and dating with confidence[http://Biologyjournals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040088]*. *PLoS Biology* 4:e88.

BEAST is an MCMC program for Bayesian phylogenetics, an alternative to Mrbayes. BEAST works with rooted trees only under the clock or relaxed-clock models.

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process (with discussion). *Journal of the Royal Statistical Society: Series B* 32:155–174.

This paper, published in a statistics journal, is noteworthy in clarifying the statistical nature of the estimation problem when observed characters for modern species are used to infer the phylogenetic relationships among the species. This may be considered the first attempt to apply Bayesian statistics to phylogeny reconstruction and was the motivation for Rannala and Yang 1996.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17.8: 754–755.

The MrBayes software implements many models of nucleotide and amino acid substitution, originally developed for the maximum likelihood method. Tree-rearrangement algorithms, such as nearest-neighbor-interchange (NNI) and subtree-pruning-regrafting (SPR), are adapted to construct MCMC proposals to allow the Markov chain traverse the tree space. It is largely thanks to MrBayes that the Bayesian method has become one of the dominant analytical frameworks in molecular phylogenetics.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16.6: 750–759.

This paper introduced MCMC proposals (such as LOCAL, a variant of the nearest-neighbor-interchange algorithm) that are still used in modern Bayesian phylogenetic programs. The program DAMBE is more efficient than the MCMCtree program produced in Yang and Rannala 1997, but both were superseded by MrBayes.

Li, S., D. Pearl, and H. Doss. 2000. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of the American Statistical Association* 95.450: 493–508.

This is the paper from the third group working on Bayesian phylogenetics.

Mau, B., and M. A. Newton. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* 6.1: 122–131.

This is the first publication from the second group working on Bayesian phylogenetics. The algorithm works on rooted trees under the molecular clock.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* 43.3: 304–311.

This is a first attempt to apply Bayesian statistics to compare molecular phylogenetic trees. Numerical integration was applied in the computation using small data sets with only 4–5 species. This was motivated by the perceived difficulty of interpreting bootstrap support values. The calculated posterior probability for a phylogeny of the apes was uncomfortably high, and, indeed, unreasonably high posterior probabilities for trees and clades are a major issue in modern Bayesian analysis of phylogenomic data sets.

Yang, Z. 2014. *Molecular evolution: A statistical approach*. Oxford: Oxford Univ. Press. [ISBN: 9780199602605]

Chapter 8 of this book provides a comprehensive coverage of Bayesian MCMC algorithms in molecular phylogenetics.

Yang, Z. 2016. AWF Edwards and the origin of Bayesian phylogenetics. In *AWF Edwards*. Edited by R. G. Winther. Cambridge, UK: Cambridge Univ. Press.
This book chapter provides a synopsis of Edwards 1970 for the biologist reader and discusses the influence of Edwards 1970 on the early efforts to apply Bayesian statistics to phylogeny estimation.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution* 14.7: 717–724.
Like Rannala and Yang 1996, this paper assumes the molecular clock. Node ages (branch lengths) were integrated out using numerical integration (rather than by MCMC) while MCMC was used to move between trees. The resulting algorithm is inefficient.

Species Tree Estimation Despite Gene Tree Conflicts

In modern analysis of multi-locus or phylogenomic data sets, different genes are commonly observed to have different histories among themselves and different from the species phylogeny. Such gene tree-species tree discordance may be due to a number of factors, such as recombination, hybridization/introgression, horizontal gene transfer, gene duplication and loss followed by misidentification of orthologues, and the coalescent process in ancestral species causing so-called incomplete lineage sorting. Evolution is said to be reticulated. Huson, et al. 2011 and Gusfield 2014 are two books that use phylogenetic networks to describe such reticulated evolution. Fundamentally evolution is tree-like: the relationships for a sequence segment in the case of recombination, or for one locus in the case of incomplete lineage sorting, are represented by a binary tree. Thus, networks are useful heuristic and visual summaries but do not accurately represent the biological process. Recent years have seen an explosion of methods that estimate the species tree using the multiple species coalescent model, which naturally accommodates incomplete lineage sorting. Hein, et al. 2005 and Wakeley 2009 are two books on coalescent, of which the multispecies coalescent is a natural extension. Rannala and Yang 2003 describes the probability distribution of gene trees under the multispecies coalescent model. Maddison 1997 is one of the earliest reviews that discuss species tree estimation despite conflicting gene trees, while Edwards, et al. 2016 and Xu and Yang 2016 are two recent reviews. Degnan and Rosenberg 2009 highlights the fact that simple methods that ignore the coalescent process, such as concatenation or the use of the most common gene tree as the estimate of the species, may be statistically inconsistent: they may converge to a wrong species tree when the number of gene loci increases.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 24.6: 332–340.
This influential review highlights the difference between gene trees and the species tree and points out that when the internal branches on the species tree are short and the ancestral populations are large, the most common gene tree may have a different topology from the species tree.

Edwards, S. V., Z. Xi, A. Janke, et al. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution* 94:447–462.

This review discusses many issues surrounding species tree estimation in the presence of gene tree conflicts, such as the impact of recombination, concatenation, and approximate and exact coalescent-based methods.

Gusfield, D. 2014. *ReCombinatorics: The algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. Cambridge, MA: MIT. [ISBN: 9780262027526]
This book summarizes combinatorial algorithms for constructing phylogenetic networks, particularly ancestral recombination graphs. This is written for both population geneticists and phylogeneticists.

Hein, J., M. H. Schieriup, and C. Wiuf. 2005. *Gene genealogies, variation and evolution: A primer in coalescent theory*. Oxford: Oxford Univ. Press. [ISBN: 9780198529965]
This book presents an elementary account of the coalescent theory, which is a central concept in modern population genetics. It includes many examples and illustrations and is ideal for a graduate course in statistics or population genetics. It does not cover inference methods, i.e., methods for estimating parameters and testing hypotheses using genetic sequence data under the coalescent model.

Huson, D. H., R. Rupp, and C. Cornavacca. 2011. *Phylogenetic networks: Concepts, algorithms and applications*. Cambridge, UK: Cambridge Univ. Press. [ISBN: 9780521755962]
This book provides an overview of phylogenetic networks, including algorithms for computing networks from different types of data sets and for drawing networks.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46.3: 523–536.
This is one of the earliest papers that describe the gene tree-species tree conflict.

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164.4: 1645–1656.
This paper gives the probability distribution of gene trees (topologies and coalescent times) under multispecies coalescent when the species tree and parameters are given. The multispecies coalescent is a simple extension to the standard coalescent.

Wakeley, J. 2009. *Coalescent theory: An introduction*. Greenwood Village, CO: Roberts. [ISBN: 9780974707754]
This is a gentle introduction to the coalescent theory, and the author explains the theory with great clarity. The book may be best for biologists who would like to develop insights into the coalescent process, and it is perhaps less ideal for probabilists interested in a rigorous mathematical treatment of the theory. There is little coverage of inference methods under the coalescent model, which typically rely on computation-intensive methods, such as importance sampling, approximate Bayesian computation (ABC) or Markov chain Monte Carlo (MCMC).

Xu, B., and Z. Yang. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204.4: 1353–1368.
This review discusses the statistical nature of the problem of species tree estimation under the multispecies coalescent, and the challenges facing both heuristic methods and full likelihood methods (maximum likelihood and Bayesian inference).

Yang, Z. 2014. *Molecular evolution: A statistical approach*. Oxford: Oxford Univ. Press. [ISBN: 9780199602605]
Chapter 9 produces a systematic introduction to the multispecies coalescent model and its use in inference problems, including estimation of species divergence times and population sizes, species delimitation and species tree estimation.

SOFTWARE

Many software programs have been developed in the field of molecular phylogenetics and evolution. Here we consider only a few commonly used ones. **PHYLIP** and **PAUP** are general packages that include parsimony, distance, and likelihood methods of phylogenetic analysis. **PHYML**, **RAxML**, and **PAML** are maximum likelihood programs for phylogenetic tree search or analysis, while **MrBayes** (and the new version **RevBayes**), **BEAST** (and the version **BEAST2**), and **PhyloBayes** are Bayesian MCMC programs.

BEAST[http://beast.community].

BEAST is a cross-platform Bayesian MCMC program for phylogenetic analysis of molecular sequence data. It uses strict clock or relaxed clock models to infer rooted trees. The MCMC sampler moves between trees, so that evolutionary hypotheses can be tested without conditioning on a single tree topology. A graphical user-interface program (called BEAUTi) can be used to set up standard analyses and a suite of programs (such as Tracer and FigTree) are available for summarizing the MCMC results.

BEAST2[http://www.beast2.org/].

BEAST2 is a redesigned version of BEAST. The new design makes it easier for other researchers to contribute analysis modules. A number of tutorials are provided for specific analyses.

MEGA (for Molecular Evolutionary Genetic Analysis)[http://www.megasoftware.net/].

MEGA is a Windows program for managing genomic sequence data, conducting sequence alignment and phylogenetic analysis using distance, parsimony, and likelihood methods. The graphical user interface makes it one of the most highly cited programs.

MrBayes[http://mrbayes.sourceforge.net/].

MrBayes is a Bayesian MCMC program for phylogenetic inference using nucleotide, amino acid, and codon sequences. It can analyze multiple heterogeneous data sets using partition models.

PAML (Phylogenetic Analysis by Maximum Likelihood)[http://abacus.gene.ucl.ac.uk/software/paml.html].

PAML is a package for likelihood analysis of nucleotide, amino acid, and codon sequences. It can be used to reconstruct ancestral sequences, detect positive selection, and estimate species divergence times under relaxed molecular clock models. It is not good for inferring trees.

PAUP 4 (Phylogenetic Analysis Using Parsimony and other methods)[http://paup.csit.fsu.edu/].

PAUP is a program for phylogenetic analysis of molecular and morphological data using distance, parsimony, and likelihood methods, written by David Swofford.

PHYLIP (Phylogeny Inference Package)[http://evolution.gs.washington.edu/phylip.html].

PHYLIP and PAUP are the two earliest computer program packages for inferring phylogenies. PHYLIP is a collection of C programs for parsimony, distance, and likelihood methods of phylogeny reconstruction, developed by Joseph Felsenstein.

PhyloBayes[http://megasun.bch.umontreal.ca/People/lartillot/www/download.html].

PhyloBayes is a Bayesian MCMC program for phylogenetic reconstruction and molecular clock dating. It includes nonstationary models of nucleotide or amino acid substitution that are particularly suitable for deep phylogenies.

PHYML[http://code.google.com/p/phyml/].

PHYML is a fast maximum likelihood tree search program.

RAxML[http://sco.h-its.org/exelixis/software.html].
RAxML is an extremely efficient maximum likelihood tree search program. The parallel versions implemented using MPI and Pthreads are especially powerful in making use of modern multiprocessor multicore computers to deal with very large data sets.
RevBayes[http://revbayes.github.io/].
RevBayes is a redesigned and rewritten version of MrBayes, released in 2016. It uses probability graphical models to set up the model for Bayesian analysis using MCMC. It has a command-line interface that resembles the popular statistical package R.