Molecular Phylogeny Reconstruction

Sudhir Kumar, *Arizona State University, Tempe, Arizona, USA*

Alan Filipski, *Arizona State University, Tempe, Arizona, USA*

Molecular phylogenetics deals with the inference of evolutionary relationships among individuals, populations, species and higher taxonomic entities using molecular data. By modelling patterns of molecular change in protein and deoxyribonucleic acid (DNA) sequences over time, scientists now routinely reconstruct evolutionary histories of species and evaluate confidence levels of the inferences. Molecular phylogenetic inferences have been not only supportive of traditional phylogenies, but also instrumental in resolving some difficult questions regarding branching orders within many evolutionary lineages. Because of the vast and growing databases of molecular sequence information, this area promises to be an important key to understanding the history and relationships of all life forms on this planet.

Advanced article

Article Contents

Online posting date: 14th March 2008

Introduction

In the second half of the twentieth century, many laboratory techniques became available for examining diversity within and among species by analysis of biologically important molecules. These include methods based on cross-reactivity of antibodies, protein electrophoresis, DNA–DNA (deoxyribonucleic acid) hybridization, restriction fragment length polymorphism and direct sequencing of DNA and proteins (polypeptides). Of these, direct comparisons of DNA sequences have been most informative and powerful (Miyamoto and Cracraft, 1991). Within the last decade, complete DNA sequences of many genomes have been obtained and the public sequence repositories are bulging with sequence information for thousands of genes from diverse species. As of June 2007, complete genomes of 568 organisms, including 49 eukaryotes, have been published and over 2000 more are being currently sequenced (Liolios *et al.*, 2006) (http://www.genomesonline.org). Thousands of virus genomes are already available as well (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html). Sequence repositories such as NCBI Genbank contain over 65 million sequence records (http://www.ncbi.nlm.nih.gov/Genbank/index.html). **See also**: Genome Databases; Molecular Evolution: Techniques

By analogy to classical phylogeny reconstruction using morphology, homologous amino acid or nucleotide sites in different organisms may be thought of as characters, with the identity of the nucleotide or amino acid at that site corresponding to the state of that character for the organism (**Figure 1**). Use of molecular sequence data has several advantages compared to the morphological characters used traditionally. For instance, no subjective appraisal is involved in the determination of character state; laboratory techniques tell us the identity of the nucleotide base at a site. Another advantage is that the same set of states (four bases or 20 amino acids) applies to all organisms, and thus we can directly compare even the most diverse life forms. In addition, the amount of available data is enormous, and it is relatively simple to obtain pertinent data for a given set of species in the laboratory today.

Concurrent availability of low-cost, powerful computers and new software algorithms has led to the routine use of molecular sequences in reconstructing evolutionary histories of organisms at various taxonomic levels. In the following, we discuss the methods of molecular phylogenetic reconstruction for DNA because DNA sequences are used most widely. These discussions also hold true for protein sequence data. Detailed account of methods for these and other types of data can be found elsewhere (Nei, 1987; Yang, 2006; Nei and Kumar, 2000; Felsenstein, 2004). **See also**: Bioinformatics; DNA Sequence Analysis; Genome Sequence Analysis; Protein Sequence Databases

ELS subject area: Evolution and Diversity of Life

How to cite:
Kumar, Sudhir; and, Filipski, Alan (March 2008) Molecular Phylogeny Reconstruction. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0001523.pub2

Methods

Assembling a DNA sequence data set

To infer the evolutionary relationship of a set of organisms using molecular sequence data, we must first ensure that the

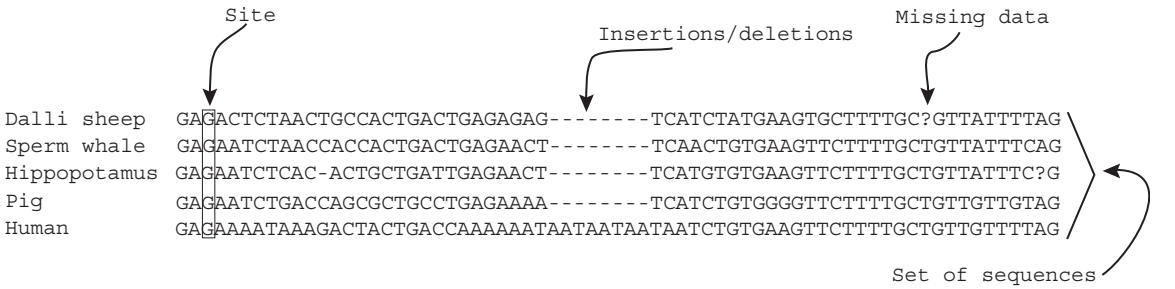

Figure 1 An alignment of a portion of the γ-fibrinogen gene sequence from five mammals. Insertion–deletion mutations predicted by sequence alignment are shown with hyphens (-) and the missing data is shown with question marks (?).

sequences being compared are homologous. We begin by selecting a gene with a homologue in each organism under study. In fact, for this purpose, the chosen sequences must not only be homologous, but need to satisfy the stronger condition of being orthologous, i.e. having diverged by speciation events rather than by gene duplications. Researchers determine sequence orthology by using criteria such as the overall sequence identity and functional similarity, and through the analysis of multigene families to which the sequences belong. One must then decide whether to use the nucleotide sequence of the gene or the amino acid sequence of its protein product, if any. There are many considerations involved. For distantly related organisms, amino acid sequences are often used, because nucleotide sequences evolve much faster than amino acid sequences owing to the redundancy of the genetic code. However, nucleotide sequences can be more informative, for example, by allowing a distinction to be made between nucleotide substitutions that do not alter the amino acid encoded (silent substitutions) and those that do (replacement substitutions). For intraspecific population genetic studies and for closely related interspecies studies in mammals, mitochondrial DNA is often used because parts of it evolve more rapidly than nuclear genes and thus provide more variation for reconstructing evolutionary history. **See also**: Evolutionary Developmental Biology: Homologous Regulatory Genes and Processes; Human Chromosome Evolution; Mitochondria: Origin; Mutations and the Genetic Code

Sequence alignment

The next step is to align the corresponding positions in different sequences. This is not trivial because sequences of a given gene often differ in length in different species as a result of insertion and deletion mutations (reviewed in Kumar and Filipski, 2007). Many computational algorithms and tools are available for this purpose (e.g. Higgins *et al.*, 1996). They work by inserting place holder symbols, usually hyphens, in the sequences to maximize the total similarity at each site, while deducting a cost for each place holder inserted (**Figure 1**). At this point, the sequences are of the same length and can be organized into columns, each representing a homologous site. Alignment of DNA sequences from distantly related species or fast evolving genes is generally more difficult. For this reason, DNA sequences that code for protein products are aligned by first constructing an alignment of corresponding amino acid sequences. The protein sequence alignment is then used as a guide to obtain the alignment of the underlying DNA sequences. **See also**: DNA Sequence Analysis

Inferring the phylogenetic tree

At this point, we are ready to infer the phylogenetic tree. The essential structure of the tree is given by its topology, i.e. which nodes are connected to which others (**Figure 2**). Almost all methods for reconstructing phylogenetic trees produce unrooted trees (**Figure 2b**). In this case, one may

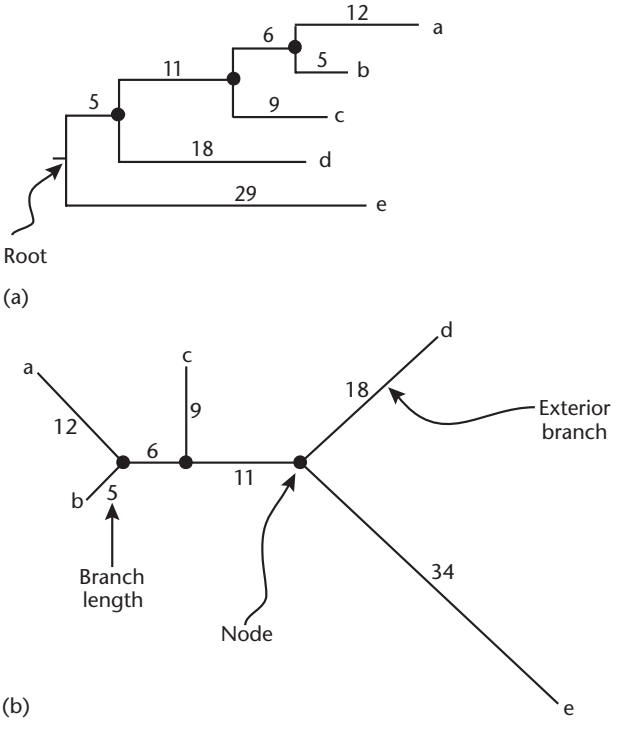

Figure 2 Rooted (a) and unrooted (b) tree of five sequences. Branch lengths are drawn proportional to evolutionary distance, which can be expressed in the units of time or the number of substitutions.

'root' the tree using a known outgroup sequence. In addition to the branching pattern, we are usually also interested in the length of the branches. Elucidation of the branching pattern is more difficult than the estimation of the branch lengths. In fact, once the topology has been established, one can use statistical methods based on least squares or maximum likelihood approaches for determining the branch lengths (reviewed in Nei and Kumar, 2000).

Several different methods are available for reconstructing phylogenetic trees. Most of them use some criterion (optimality principle) for evaluating the fit of a given data set to the topology and then search for the tree that gives the best score in terms of that criterion (see later). If the criterion used is realistic and the data are sufficient, the tree should represent the true phylogenetic relationship of the sequences (and thus the associated organisms). In practical situations, however, this is complicated by the fact that the number of different tree topologies that can be made from a set of sequences increases very rapidly with the number of sequences (**Table 1**), and we must use heuristics to constrain the search to find a potentially optimal tree quickly. Fortunately, the quality of phylogenetic trees produced by quick heuristics is similar to that obtained with extensive (or even exhaustive) searches, while more efficient heuristics are being developed to deal with larger trees (Nei and Kumar, 2000; Felsenstein, 2004; Takahashi and Nei, 2000; Hordijk and Gascuel, 2005; Stamatakis *et al*., 2005).

At present, three commonly used tree-building criteria in molecular phylogenetics are minimum evolution (ME), maximum parsimony (MP) and maximum likelihood (ML). Under the MP criterion, the topology requiring the smallest number of nucleotide changes to fit the observed sequence data is chosen to represent the true tree (Fitch, 1971). In ML methods, the topology with the greatest likelihood (probability of being generated) under a given probabilistic model of nucleotide substitutions is chosen (Felsenstein, 1981). In the ME methods, the sequence data are first transformed into a matrix of distances between sequence pairs. Then, the total branch length needed to fit this matrix to each possible topology is computed, and the topology requiring the smallest total branch length is chosen. The evolutionary distance between a pair of sequences can be estimated in a number of ways. The simplest distance measure between two sequences is the *p*-distance, which is the fraction of sites at which the two sequences differ. The *p*-distance is known to underestimate the true amount of evolution because it does not account for multiple substitutions at the same site. This problem can be remedied by using a more complex model of nucleotide substitution. A detailed explanation for estimating distances under different models of substitutions and guidelines on choosing appropriate distance measures can be found in Nei and Kumar (2000).

Table 1 Number of possible unrooted trees for different numbers of sequences

Sequences	Trees
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
11	654 729 075

The neighbour-joining method (Saitou and Nei, 1987), because of its computational efficiency, has been frequently used in molecular phylogenetics, especially for large-scale data analyses. It is based on, but does not necessarily optimize, the ME criterion. The neighbour-joining method works in a stepwise fashion by minimizing the sum of branch lengths at each step of sequence clustering. The unweighted pair-group method using arithmetic averages (UPGMA) is another distance-matrix-based method, in which pairs of sequences showing the smallest evolutionary distance are clustered first. This method assumes that the evolutionary rate has remained constant throughout the evolutionary history of the given set of organisms. Since this assumption is rarely met in reality, UPGMA should be used cautiously for inferring phylogenetic histories. Many academic software packages are available for computing distances and inferring phylogenetic trees, for example (Kumar *et al*., 2004). **See also**: Biological Computation; Molecular Evolution: Patterns and Rates

Maximum parsimony and maximum likelihood methods are also widely used and each has advantages. MP, for example, is thought to be nonparametric in the sense that it does not require specification of a model of evolutionary change. However, MP is known to be inconsistent under certain conditions, which means that, even with arbitrarily long sequences, certain phylogenetic trees will be reconstructed incorrectly (the so-called long branch attraction problem, discussed later). This is not a fatal flaw, but such potential situations need to be watched for (Felsenstein, 2004; Bergsten, 2005).

Maximum likelihood is generally regarded as the 'gold standard' of molecular phylogenetic reconstruction, in terms of accuracy, but the method does depend on an appropriate model of molecular evolution with accurately estimated parameters. Another disadvantage of ML is that it tends to be very time-consuming in terms of computational effort, and comprehensive searches of tree topology cannot be made except in trivial cases.

In the last few years, Bayesian methods of phylogenetic tree reconstruction have become more popular. Bayesian methods involve similar model-based computations to ML procedures, but they estimate a posterior probability of the model given the observed sequence data. Efficient algorithms (e.g. Markov Chain Monte Carlo) for doing this have enabled Bayesian methods to compete effectively with ML (Huelsenbeck *et al*, 2001; Hall, 2005).

The choice of which tree-building method to use is somewhat arbitrary and often depends on time requirements,

software availability or the philosophical predisposition of the researcher. This is because: (1) no method is uniformly better in reconstructing the true tree when the sequence length is small and (2) all methods tend to perform well given enough data, except when they are inconsistent (Nei and Kumar, 2000; Hall, 2005).

Assessing reliability

The next step in constructing a sequence phylogeny is to assess the reliability of the inferred branching pattern. This is often accomplished by a bootstrap analysis (Felsenstein, 1985). Bootstrap procedures involve construction of new sequence sets by resampling with replacement sites (columns) of the original set, building a tree for each new set, and calculating the percentage of times a cluster reappears in the bootstrap replications. This percentage is called the bootstrap value; clusters with a bootstrap value $\geqslant 95\%$ are widely considered to reflect correct relationships, although some authors have suggested that 70% may be a more realistic cutoff point.

Under Bayesian methods of phylogenetic analysis, information about the reliability of inferred trees is available in terms of the posterior probabilities (credibility intervals) without bootstrapping, but these have been criticized as giving misleadingly high levels of support (Suzuki *et al*., 2002; Douady *et al*., 2003; Alfaro and Holder, 2006; Yang and Rannala, 2005).

Finally, it is important to remember that measures such as bootstrap support and posterior probability refer only to the consistency and reliability of constructing the given tree from a specific sequence data set, and do not take into account possible bias in the data set or incorrect model assumptions. For a detailed explanation of the bootstrap test and information on other types of tests of phylogenetic trees, see Nei and Kumar (2000) and Felsenstein (2004).

Major Software Packages for Building Phylogeny

Hundreds of software packages are available to perform all of the major types of phylogenetic analysis mentioned earlier. Some are command-line oriented while some have graphical user interfaces of varying degrees of sophistication. Most are available on multiple platforms, including Microsoft Windows, Unix/Linux and Apple Macintosh operating systems. Almost all are free and have open source code, but some have a small cost associated with them. Popular programs for MP analysis are PAUP* (Swofford, 2001) and Phylip (Felsenstein, 1993), while ML methods are available in PAML (Yang, 1997), HYPHY (Pond *et al*., 2005) and PHYML (Guindon and Gascuel, 2003). Bayesian methods are implemented in MrBayes (Ronquist and Huelsenbeck, 2003) and Beast (Drummond and Rambaut, 2006), while some programs, such as MEGA (Kumar *et al*., 2004; Tamura *et al*., 2007), integrate several methods under a user-friendly graphic interface. These programs taken together constitute the tools used in the overwhelming majority of published phylogenetic analyses (Kumar and Dudley, 2007). A more complete list of hundreds of phylogeny programs is available from Joseph Felsenstein at the web site http://evolution.genetics.washington.edu/phylip/software.html, and a book by Hall (2007) is an excellent way to get started with molecular phylogenetics.

Impact on Phylogenetics

In general, molecular phylogenetics studies have supported traditional phylogenies constructed on the basis of nonmolecular characters and provided clarification to debates. However, there have been some notable disagreements between molecular and classical phylogenies. One such example involves the identification of the closest living relatives of the hippopotamus (family Hippopotamidae). Before molecular phylogenetic analyses, Hippopotamidae was thought to be most closely related to Suina (pigs and peccaries), within the mammalian order Artiodactyla. Molecular studies using mitochondrial and nuclear DNA sequences have now clearly established that Hippopotamidae is a sister group to Cetacea (containing whales and dolphins), and that this group is more closely related to ruminants (cows, sheep and deer) than to pigs and peccaries (e.g. Gatesy, 1997). Molecular phylogenetics has also resolved the human–chimpanzee–gorilla trichotomy (Satta *et al*., 2000), identified Chimpanzee *Simian immunodeficiency virus* as the closest relative of the *Human immunodeficiency virus type 1* (Paraskevis *et al*., 2003), and provided insights into sister group relationships of animals and fungi (Baldauf and Palmer, 1993). Some other questions still remain unsettled, such as the exact timing and pattern of mammalian ordinal diversification (Springer *et al*., 2003; Murphy *et al*., 2007) and the relationship of certain animal phyla (Wolf *et al*., 2004; Dopazo and Dopazo, 2005; Putnam *et al*., 2007), but these questions only stimulate the collection of data and the development of new techniques of analysis. **See also**: Apes; Artiodactyla (Even-Toed Ungulates Including Sheep and Camels); Cetacea (Whales, Porpoises and Dolphins); Fossils in Phylogenetic Reconstruction; Phylogeny Based on 16S rRNA/DNA

Variable Rates

Molecular phylogenetics would be simpler if all sites in a gene evolved at the same rate (uniform substitution rate among sites) and if all species evolved at the same rate in a given gene (equal rates among lineages). Within a gene, however, certain sites are under stronger natural selection than others because of their functional importance. This variability in evolutionary rates among sites is often accounted for in phylogenetic inference by using a variety of models, including a gamma model of nucleotide substitution

and models that allow for a certain fraction of sites to not change (Yang, 1996; Guindon and Gascuel, 2003).

Variable evolutionary rates among lineages

The observed heterogeneity of evolutionary rates among lineages in a gene is caused partly by the nondeterministic nature of the evolutionary processes, partly by differences in intensity and type of natural selection and partly by unknown factors. Tree-building methods mentioned earlier, with the exception of UPGMA, do not assume constancy of evolutionary rate among lineages (molecular clock) and can thus be used directly. However, some methods are known to produce consistently incorrect results when the evolutionary rates vary significantly among lineages. For instance, Felsenstein (1988) showed that if a four-sequence tree contains two long and two short branches, then the long branches tend to cluster together in the MP trees even if they are distantly related (long-branch attraction problem). One way to avoid this problem is by using a larger number of sequences such that the long branches are broken. Another way to minimize the effect of long-branch attraction is to use ML, ME or Bayesian methods, as described earlier.

See also: Molecular Clocks; Molecular Evolution: Rates

In summary, molecular phylogenetics has become an integral part of research endeavours in diverse areas of molecular biology, population genetics, developmental biology and evolutionary biology and has implications for ecology and medicine. We are only at the beginning of this area of study and rapid growth of data and computational power should lead to the resolution of many long-standing problems in these fields.

References

Alfaro ME and Holder MT (2006) The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **37**: 19–42.

Baldauf SL and Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences of the USA* **90**: 11558–11562.

Bergsten J (2005) A review of long-branch attraction. *Cladistics* **21**: 163–193.

Dopazo H and Dopazo J (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biology* **6**: R41.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF and Douzery EJP (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution* **20**: 248–254.

Drummond AJ and Rambaut A (2006) *BEAST v1.4*, available from http://beast.bio.ed.ac.uk/

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**: 368–376.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Reviews in Genetics* **22**: 521–565.

Felsenstein J (1993) PHYLIP (phylogeny inference package). Version 3.6a. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Felsenstein J (2004) *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Fitch W (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**: 406–416.

Gatesy JC (1997) More DNA support for the Cetacea/Hippopotamidae clade: the blood-clotting protein gene gamma-fibrinogen. *Molecular Biology and Evolution* **14**: 537–543.

Guindon S and Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696–704.

Hall BG (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution* **22**: 792–802.

Hall B (2007) *Phylogenetic Trees Made Easy*, 3rd edn. Sunderland, MA: Sinauer Associates.

Higgins DG, Thompson JD and Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* **266**: 383–402.

Hordijk W and Gascuel O (2005) Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21**: 4338–4347.

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback J (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**: 2310–2314.

Kosakovsky Pond SL, Frost SD and Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Kumar S and Dudley J (2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics* **23**: 1713–1717.

Kumar S and Filipski A (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Research* **17**: 127–135.

Kumar S, Tamura K and Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics* **5**: 150–163. [http://www.megasoftware.net].

Liolios K, Tavernarakis N, Hugenholtz P and Kyrpides NC (2006) The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research* **34**: D332–D334.

Miyamoto MM and Cracraft J (1991) *Phylogenetic analysis of DNA sequences*. New York: Oxford University Press.

Murphy WJ, Pringle TH, Crider TA, Springer MS and Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research* **17**: 413–421.

Nei M (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Paraskevis DP, Lemey P, Salemi M *et al*. (2003) Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using Bayesian inference: implications for the origin of HIV-1. *Molecular Biology and Evolution* **20**: 1986–1996.

Putnam NH, Srivastava M, Hellsten U *et al*. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**: 86–94.

Ronquist F and Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Saitou N and Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **6**: 514–525.

Satta Y, Klein J and Takahata N (2000) DNA archives and our nearest relative: the trichotomy problem revisited. *Molecular Phylogenetics and Evolution* **14**: 259–275.

Springer MS, Murphy WJ, Eizirik E and O'Brian SJ (2003) Placental mammal diversification and the cretaceous-tertiary boundary. *Proceedings of the National Academy of Sciences of the USA* **100**: 1056–1061.

Stamatakis A, Ludwig T and Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**: 456–463.

Suzuki Y, Glazko GV and Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the USA* **99**: 16138–16143.

Swofford DL (2001) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta*. Sunderland, MA: Sinauer Associates.

Takahashi K and Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution* **17**: 1251–1258.

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution* **24**: 1596–1599.

Wolf YI, Rogozin IB and Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Research* **14**: 29–36.

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* **11**: 367–371.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**: 555–556.

Yang Z (2006) *Computational Molecular Evolution*. Oxford: Oxford University Press.

Yang Z and Rannala B (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology* **54**: 455–470.

Further Reading

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis*. Cambridge: Cambridge University Press.

Graur D and Li W-H (1999) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Li W-H (1997) *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Page RDM and Holmes EC (1998) *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science.

Patthy L (1999) *Protein Evolution*. Oxford: Blackwell Science.