VERSITA

CELLULAR & MOLECULAR BIOLOGY LETTERS
http://www.cmbl.org.pl

Received: 16 August 2009
Final form accepted: 01 March 2010
Published online: 19 March 2010

Volume 15 (2010) pp 311-341
DOI: 10.2478/s11658-010-0010-8

Review

MOLECULAR SYSTEMATICS: A SYNTHESIS OF THE COMMON METHODS AND THE STATE OF KNOWLEDGE

DIEGO SAN MAURO[1]* and AINHOA AGORRETA[1, 2]
[1] Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom, [2] Department of Zoology and Ecology, University of Navarra, Irunlarrea s/n, 31008 Pamplona, Spain

Abstract: The comparative and evolutionary analysis of molecular data has allowed researchers to tackle biological questions that have long remained unresolved. The evolution of DNA and amino acid sequences can now be modeled accurately enough that the information conveyed can be used to reconstruct the past. The methods to infer phylogeny (the pattern of historical relationships among lineages of organisms and/or sequences) range from the simplest, based on parsimony, to more sophisticated and highly parametric ones based on likelihood and Bayesian approaches. In general, molecular systematics provides a powerful statistical framework for hypothesis testing and the estimation of evolutionary processes, including the estimation of divergence times among taxa. The field of molecular systematics has experienced a revolution in recent years, and, although there are still methodological problems and pitfalls, it has become an essential tool for the study of evolutionary patterns and processes at different levels of biological organization. This review aims to present a brief synthesis of the approaches and

* Author for correspondence: e-mail: d.san-mauro@nhm.ac.uk, tel.: +44 (0)20 7942 5315, fax: +44 (0)20 7942 5054

Abbreviations used: *actB* – β-actin; AIC – Akaike information criterion; BI – Bayesian inference; BIC – Bayesian information criterion; *cob* – cytochrome *b*; *cox1* – cytochrome *c* oxidase subunit 1; DNA – deoxyribonucleic acid; GTR – General Time-Reversible; HIV – human immunodeficiency virus; HKY – Hasegawa Kishino Yano; hLRT – hierarchical likelihood ratio tests; JTT – Jones Taylor Thornton; LBA – long-branch attraction; LRT – likelihood ratio test; MCMC – Markov chain Monte Carlo; ME – minimum evolution; ML – maximum likelihood; MP – maximum parsimony; mtREV – mitochondrial reversible; NJ – neighbour-joining; PCR – polymerase chain reaction; *rag1* – recombination activating gene 1; rRNA – ribosomal ribonucleic acid

methodologies that are most widely used in the field of molecular systematics today, as well as indications of future trends and state-of-the-art approaches.

Key words: Molecular systematics, Phylogenetic inference, Molecular evolution, Phylogeny, Evolutionary analysis, Evolutionary hypothesis, Divergence time.

INTRODUCTION

In recent years, the outstanding advancement of molecular biology and bioinformatics has supplied researchers with powerful tools for tackling long-unresolved problems in all areas of biology. Molecular systematics can be defined as the use of the information contained in molecular data to reconstruct phylogenetic relationships. A phylogeny, or evolutionary tree, is the pattern of historical relationships among groups (lineages) of elements (e.g. organisms, sequences) [1]. Understanding this pattern of relationships is essential in comparative studies because there are statistical dependencies among elements sharing common ancestry. Phylogenetic analyses used to be restricted to studies of organism evolution, but today, they are a standard tool in broader fields of research, whether related to genomics, protein engineering, conservation biology, or pest control in agriculture. For example, phylogenies were used to study the timing and ancestry of the main pandemic strain of the human immunodeficiency virus (HIV) [2], and more recently to investigate the origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A outbreak [3].

Of the various techniques that can be used for molecular systematics [reviewed in 4 and 5], the analysis of DNA and/or protein sequence variation has become the standard, and has been used in the vast majority of recent phylogenetic studies. Using DNA and amino acid sequences in molecular systematics has several advantages over traditional morphological approaches [6]: the universality of the character types and states (yielding a more objective selection and definition of homology, i.e. a similarity in character states due to their inheritance from a common ancestor); the high number of characters available for analyses (yielding data with a better statistical performance); the high degree of variation in the substitution rates among genes and gene regions (providing different levels of variability for specific questions); our increasingly comprehensive knowledge of the molecular basis underlying sequence evolution and function (allowing the construction of more sophisticated models of the evolutionary process); and the relatively easy collection of the data from different taxa (even from very small tissue samples and by researchers that do not necessarily have taxon-specific expertise). In the last few years, a great amount of sequence data has been generated for higher taxa, and this has definitely boosted the possibilities for comparative and phylogenetic studies.

However, phylogenetic inference from molecular data is not free from methodological problems and pitfalls [6-8]. For example, molecular sequence data has a relatively low character state space (four states in the case of DNA,

20 in the case of amino acids), which may entail a high probability of homoplasy (similarity in character states for reasons other than common ancestry, such as convergence, parallelism, and reversal) due to the saturation of the substitution process (e.g. two sequences might have the same character state at a given position just by chance and not due to common ancestry) [1]. In practice, this problem is particularly important for some methods of phylogenetic inference, such as parsimony. Furthermore, gene phylogenies do not necessarily match those of the organisms due to several evolutionary processes such as horizontal gene transfer, gene duplication and loss, and deep coalescence [9]. Homologous genes that were separated by a speciation event (when a species diverges into two separate species) are termed orthologous, whereas homologous genes that were separated by a gene duplication event and occupy two different positions in the same genome are termed paralogous. One might expect a total match between the gene tree and species tree if the genes used to reconstruct the phylogeny are orthologous, but in practice, different sets of orthologous genes may yield different phylogenies [10] because of analytical limitations [11-13] and differences in the phylogenetic signal:noise ratio due to the unequal action of natural selection or genetic drift [14]. In some cases, molecular data is irretrievable, such as in ancient fossil taxa, although some data has been obtained for 'recently' extinct organisms [15] and in other exceptional cases [16]. Another important problem of molecular data is that it cannot as yet be used on its own to describe new species; however, see [17-19].

Molecular and morphological data is useful and necessary in systematics. These two types of data constitute independent and complementary sources of information for cross-validating hypotheses about evolutionary patterns and processes at different levels of biological organization.

MODELLING SEQUENCE EVOLUTION

A model of sequence evolution provides a statistical description of the process of character state change, i.e. the process of nucleotide or amino acid substitution. In general, nucleotide and amino acid substitution is viewed as a Markov process: a mathematical model of infrequent changes of discrete states over time, in which future events occur by chance and depend only on the current state, and not on the history of how the state was reached [20-22]. This Markov model also assumes that substitution rates do not change over time (time-homogeneous), and that relative frequencies of each character state are in equilibrium (stationary) [23]. The mathematical expression of a substitution model is a table of rates (substitutions per site per unit of evolutionary distance) at which each character state (either of the nucleotide or amino acid) is replaced by each alternative state [20]. As models become more sophisticated, these instantaneous rate matrices grow in complexity, and other parameters can be incorporated [24]. For example, frequency parameters inform about the relative abundance of each character state (each nucleotide in the case of DNA or each

amino acid in the case of protein sequences). Furthermore, substitution rates usually vary across the sites of the DNA or amino acid sequence due to unequal selective pressure, biochemical factors, and/or genetic code constraints [25, 26]. This variation is modelled using among-site rate variation parameters. One of these parameters leaves a proportion of sites incapable of undergoing substitutions (the 'proportion of invariable sites', I), with the remaining positions varying at the same rate [27, 28]. The other parameter uses a gamma distribution (Γ) for modelling rate heterogeneity across sequence sites [29]. The gamma distribution has a shape parameter α, and its mean and variance are 1 and $1/\alpha$, respectively. As α increases, the rate distribution tends to an equal-rates model. The gamma is a continuous distribution, but it is usually implemented in a discrete manner using several categories (commonly four or eight) of equal probability to approximate it [30].

In the last four decades, many models of increasing complexity have been described, both for nucleotide and amino acid sequences [31]. In general, models of sequence evolution are built following two main approaches [21], either using properties calculated through the comparison of large numbers of observed sequences (empirically), or on the basis of the chemical or biological properties of DNA or amino acids (parametrically). Empirical models result in fixed parameter values that are estimated only once and then assumed to be applicable to all datasets. The modelling of amino acid replacement, such as mtREV [32] and JTT [33], has concentrated on this empirical approach. By contrast, parametric models allow the parameter models to be derived from the dataset in each particular analysis. The modelling of nucleotide replacement, such as HKY [28] and GTR [34], has concentrated on this parametric approach. More complex models of sequence evolution have also been described, such as codon-based models that take the genetic code into account when calculating the probability of a change at a site across a branch [35], models attempting to accommodate structural elements of the analyzed molecules [36], or models allowing site-specific rate variation across lineages [37, 38].

A proper characterization of the process of sequence evolution is essential in molecular phylogenetic inference [39], as phylogenetic methods tend to be less accurate or inconsistent (i.e. they may yield spurious phylogenetic relationships) when an incorrect model of sequence evolution is assumed [40, 41]. In general, model selection strategies attempt to find the appropriate level of complexity on the basis of the available data [42]. Increasing model complexity improves the fit to the data, but also increases the error in the estimated parameters [43, 44]. Therefore, the use of overparametrized models should be avoided in order to keep estimates as precise as possible. In recent years, several statistical methods (based on hypothesis testing) have been developed for selecting best-fit models of sequence evolution for a given dataset [23, 43]. These methods use likelihood ratio tests (see below) usually in a hierarchical manner (hLRT) or information criteria such as the AIC [45] or the BIC [46] to contrast the fit to the data of different alternative models. Although both likelihood ratio tests and information

criteria are in wide use, recent studies suggest that the latter are more adequate because they are able to simultaneously compare multiple nested or non-nested models (hLRT can only compare nested models, and the order of tests can influence which model is ultimately chosen), and permit the assessment of model selection uncertainty [47].

ASSEMBLING A SEQUENCE DATA MATRIX

Given a particular group of organisms, the process of phylogenetic estimation (Fig. 1) starts with the collection of homologous sequence data. This sequence data can be obtained anew (using molecular biology techniques), but some can be retrieved from the many gene databases available, such as GenBank [48]. Typically, a few outgroup sequences are included to root the tree, indicating which nodes in the tree are the oldest, and providing clues about ancestral sequence states or ancestral descendent relationships [49]. In general, outgroups are added as a single sister clade (preferably the closest) to the ingroup. The choice of outgroup can strongly affect the chances of obtaining the correct tree, because both topology and unequal rates of molecular evolution between the ingroup and the outgroup affect the ability of tree-building algorithms to find the correct tree [50].

The next step is to align the sequences, because as they diverge from each other, length mutations (insertions and deletions, collectively referred to as *indels*) accumulate [1], and gaps need to be inserted into sequences to increase their similarity. The alignment is an arrangement of the sequences into a matrix so that the character states at each given position (column of the matrix) are related to each other by descent from a common ancestral residue, i.e. there is positional homology. The character states can be either nucleotides or amino acids. The alignment step is critical as the rest of the phylogenetic inference process relies on it [51-53], and many algorithms have been developed for multiple-sequence alignment [54, 55]. In most cases, these algorithms attempt to minimise the total cost of all the possible changes (the combination of length and substitution mutations) in pairwise sequence comparisons [56]. They usually work in a progressive manner (a set of N sequences are aligned by performing N-1 pairwise alignments of pairs of sequences or pairs of intermediate alignments [57]) guided by a phylogenetic tree connecting the sequences (progressive tree alignment). The main problem of multiple alignments is that the costs are difficult to define and interpret biologically [58]. Ambiguously aligned positions are usually excluded from the dataset before analysis.

The accuracy of multiple-sequence alignment tools has greatly improved in recent years [54]. Future improvements will likely be related to the incorporation of additional biological information in the alignment algorithm, such as the secondary structure in the case of rRNA sequences [59, 60], the combination of sequence alignment algorithms with the statistical methods applied to the

analysis of genomic data [61], and the simultaneous co-estimation of sequence alignments and phylogenetic trees in a probabilistic framework [62, 63].

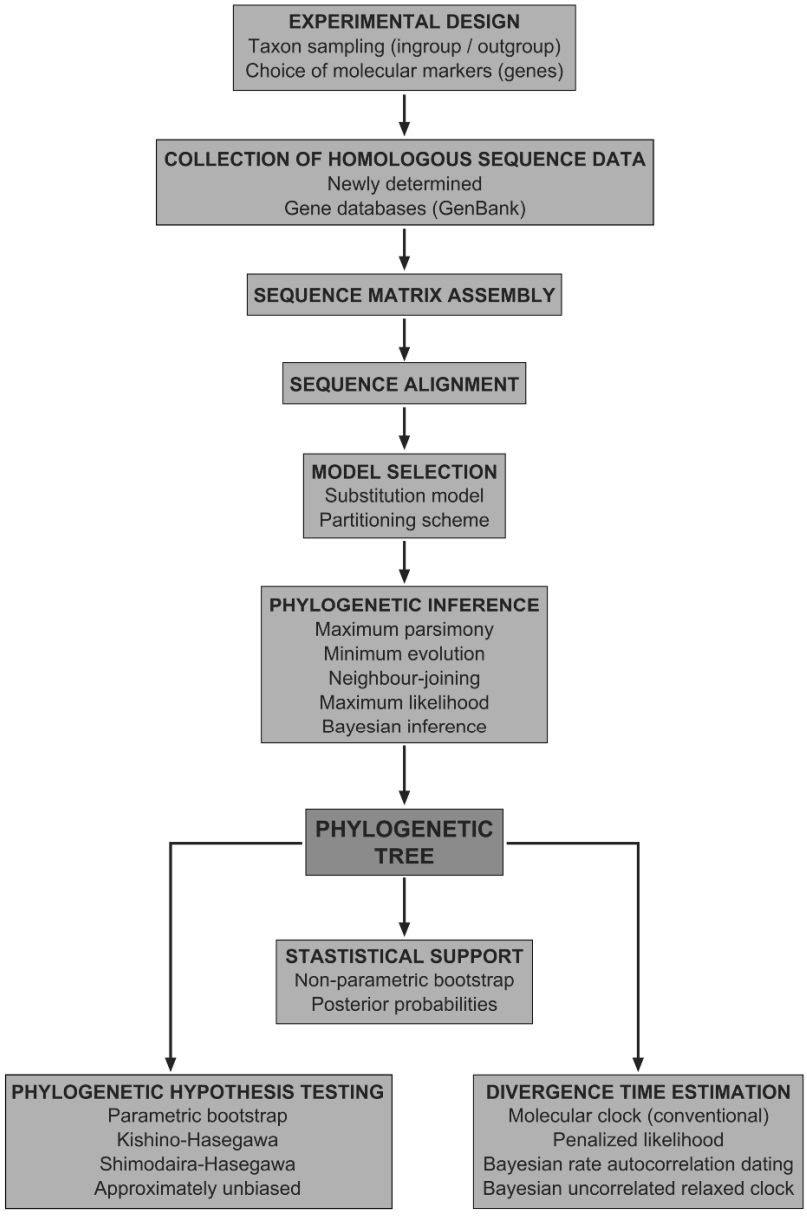

Fig. 1. A flowchart of the process of phylogenetic estimation. Some steps have been omitted or condensed for simplicity.

In addition to the sequence data, a model of sequence evolution must be chosen, as the methods used in molecular phylogenetics are based on a series of assumptions about how the substitution process works (see above). These assumptions can be implicit, as in parsimony methods, or explicit, as in likelihood-based methods [43].

METHODS OF PHYLOGENETIC INFERENCE

Of the various methods developed to reconstruct phylogenetic relationships [see reviews in 20, 21, 31 and 42], there are four that have dominated molecular systematic studies: maximum parsimony, neighbour-joining, maximum likelihood, and Bayesian inference.

Maximum parsimony

Maximum parsimony is one of the earliest inference methods [64, 65] that arose from Hennig's phylogenetic systematics [66]. Unlike distance methods, it directly uses character states, and is based on an optimality criterion, which is a rule to decide which of the alternative trees is the best: it selects the tree or trees requiring the fewest character state changes, thus attempting to minimize homoplasy. The length of an unrooted tree can be directly calculated using Fitch's algorithm [64], which moves along the tree assigning one or more states to each of the internal nodes. In this method, the tree space (the theoretically possible tree topologies for a given number of taxa) is usually searched using heuristic searches or, when the number of sequences is small (<12), exact searches [31]. Exact searches are those that evaluate all possible trees (exhaustive searches) or parts thereof (in a way that ensures that the optimal tree will not be missed from the evaluation, such as 'branch and bound' searches), and thus guarantee that the tree found is optimal. Heuristic searches are those that do not evaluate all the possible trees and cannot guarantee that the tree found is the optimal one. To maximize the chances of success, several independent searches are performed starting from different regions of the tree space, giving a good, albeit not definitive indication that the optimal tree has been found if all the searches find the same tree.

The advantage of maximum parsimony is that it is fast enough for the analysis of large datasets containing many sequences, and it is robust if the branches of the tree are short, whether this is because the sequences are closely related or because the taxon sampling is dense. However, maximum parsimony can perform poorly, and even be seriously misleading, if there is substantial variation in the rates of evolution among the taxa. In this case, the taxa with the fastest substitution rates appear in the tree as long branches, and tend to artefactually group together. This phenomenon is called long-branch attraction (LBA) [67], and parsimony is particularly affected by it [20, 68]. Unweighted parsimony lacks an explicit model of sequence evolution [69], so it is difficult for this method to deal with a high degree of homoplasy (i.e. parallel, convergent,

reversed, or superimposed changes) when markedly divergent sequences are analysed. In such a case, parsimony analyses can be weighted through the use of step matrices to incorporate prior assumptions about the costs of character state change (e.g. transitions and transversions) [20].

Neighbour-joining and minimum evolution

These are pairwise distance methods [70, 71] based on the assumption that dissimilarity between two sequences is directly related to their phylogenetic relationship. Such dissimilarity arises from the number of changes that have occurred along the branches, i.e. the evolutionary distance. Distance methods comprise both clustering methods such as neighbour-joining, and optimality methods such as minimum evolution. Clustering methods were originally developed to detect similarities rather than to estimate evolutionary relationships [72]. In neighbour-joining [73], the DNA or amino acid sequences are first converted into a distance matrix that is then used to reconstruct a phylogenetic tree. Optimality methods calculate the score of a tree based on the squared deviation of the pairwise observed distances between each pair of taxa, estimated from the data matrix, and the distance separating those taxa on the tree [71]. In minimum evolution [74], the optimality criterion is the sum of branch lengths optimized according to the least-squares criterion above (the minimum evolution score).

The main advantage of distance methods is that they are relatively rapid compared to all the other methods available, and they perform well when the divergence between the sequences is low. The disadvantages of distance methods include the loss of information when the sequences are converted to distances, and the difficulty in obtaining reliable estimates of pairwise distances for highly divergent sequences. Both neighbour-joining and minimum evolution can incorporate models of evolution to correct pairwise genetic distances for multiple substitutions at the same site [75].

Maximum likelihood

This method is one of the standard tools of statistics [76], and was first applied to phylogenetics several decades ago [22, 77, 78]. In the context of molecular systematics, the likelihood of a phylogenetic tree is the probability of observing the data (the set of sequences being analyzed) given the tree and the model of evolution. This is also an optimality method: the best tree is the one that renders the observed sequences most likely to have evolved under the assumed evolutionary model. The likelihood of a site is the probability of the observed states at that site given all the possible combinations of states at the internal nodes of the tree (ancestral states). The likelihood of a tree is the product of the likelihoods for each site of the alignment. Because likelihood values are often very small, they are usually expressed as log likelihoods, ln L (which are computationally easier to handle).

As with parsimony, the tree space is usually explored using heuristic searches. The great advantage of maximum likelihood is that it allows the inference of

phylogenetic trees using complex models of sequence evolution, including the ability to estimate model parameters, thus allowing simultaneous inference of patterns and processes of molecular evolution, and provides a powerful statistical framework for hypotheses testing (see below). The models of sequence evolution can be implemented for the whole sequence dataset [79], and for different partitions (subsets treated independently) of it, such as the different codon positions of a protein-coding gene [80-82]. The strong statistical foundations of likelihood-based methods probably make them the most robust way for estimating molecular phylogenies and understanding sequence evolution [21]. However, there are also criticisms of this method related to the fact that the result may be especially dependent on the correctness of the employed model of sequence evolution [21, 42; but see also 83]. Another criticism is that maximum likelihood can be prohibitively slow and computationally demanding, particularly when there is a large number of sequences (terminal taxa) to be analyzed. However, the exponential increase in computer power and the development of faster algorithms [81, 84] have largely solved this problem.

Bayesian inference

This is the most recently developed of all the phylogenetic inference methods [85-87]. The field of Bayesian statistics is closely allied with maximum likelihood: the optimal hypothesis is the one that maximizes the posterior probability. According to Bayes' theorem, the posterior probability for a hypothesis is proportional to the likelihood multiplied by the prior probability of that hypothesis. Like maximum likelihood, Bayesian analysis allows complex models of sequence evolution, generally the same ones used in maximum likelihood, to be implemented for the whole sequence dataset, and for different partitions of it. This method involves specifying a model and a prior distribution (the probability distribution of parameter values before observing the data) and then integrating the product of these quantities over all possible parameter values to determine the posterior probability for each tree. However, the likelihood functions for phylogenetic models are currently too complex to integrate analytically, so Bayesian approaches rely on Markov chain Monte Carlo (MCMC) procedures [88-90]. This algorithm works by taking sample trees from the distribution of posterior probabilities. Unlike maximum likelihood, which searches for a single most likely tree, Bayesian MCMC searches for the 'best set of trees' in the landscape of possible trees. The initial state of the chain is a tree with a combination of branch lengths and parameters of the substitution model. Given this initial state, a probability can be calculated from each site along the alignment. A new state of the chain is then proposed, changing a parameter of the model, moving a branch and/or varying a branch length to create a modified tree, and the ratio of likelihoods of the states is calculated. If that ratio is higher than a number randomly drawn between 0 and 1, the new state is accepted. Otherwise, it remains the same. In general, if the new tree is more likely than the preceding tree (given the data and substitution model), it is more likely to be

accepted. These steps constitute an MCMC generation. Sequential values (new states) are simulated from the chain until it converges (i.e. the simulated variables stay on values with high probability in the stationary chain), and then the states (tree and model parameters) are sampled at intervals from the chain, thus constituting independent samples from the stationary distribution. As the number of generations increases, the process does an approximation of the landscape of possible states (tree and parameters); the longer the MCMC runs, the closer the approximation becomes. Current Bayesian phylogenetic procedures use a variant of MCMC called Metropolis-coupled MCMC, which is less prone to entrapment in local optima. This MCMC variant involves running several independent chains simultaneously (typically four). Each generation, every chain proposes and accepts/rejects moves independently, and a swap of the states is attempted between two randomly chosen chains. States are only sampled from one of the chains, designated as the 'cold' chain (the rest are 'heated' chains). If the cold chain gets stuck in local optima (a low probability hill in the posterior density landscape), it has a chance to escape by swapping with another chain that may be on a higher hill.

Bayesian inference has the advantage of a strong connection with the likelihood framework and its powerful statistical foundations. Moreover, as a result of the MCMC process, there is a posterior probability associated to each node on the inferred Bayesian tree (the fraction of times a clade occurs among the sampled trees) that can be used as a measure of support for that node. The disadvantages of Bayesian methods stem from the fact that prior distributions for parameters must be specified, and that it can be difficult to determine if the MCMC approximation has run for a sufficient number of cycles, meaning that the chains have converged [42], and thus if the tree space has been adequately searched. Long Bayesian runs (millions of generations, and starting from different initial states) are typically required to reach convergence and ensure an adequate search of the tree space [91]. There is a myth that the Bayesian MCMC is faster and computationally less demanding than maximum likelihood using equally (or even more) complex models of sequence evolution. However, each procedure is trying to do rather different things: the Bayesian approach explores the entire posterior distribution of the tree and all the parameters, while maximum likelihood just searches for a single tree and set of parameters that maximise the likelihood. The amount of computation in each case can vary greatly.

STATISTICAL SUPPORT OF PHYLOGENETIC TREES

Other than Bayesian inference, which yields a tree with support values for each node, measured as posterior probabilities, the methods of phylogenetic reconstruction produce only point estimates of the phylogeny. However, an important issue is to know how strongly the data supports each of the relationships depicted in the tree. Several methods for assessing confidence exist [92], but this issue has traditionally been tackled by bootstrapping, which was

first applied to phylogenetics by Felsenstein [93]. This is a statistical resampling technique by which distributions that are difficult to calculate exactly can be estimated by the repeated creation and analysis of artificial datasets. To assess node support in phylogenetics, non-parametric bootstrapping is used: new datasets are created by sampling randomly and with replacement from the original data (these new bootstrap datasets are of the same size as the original); a desired quantity of bootstrap datasets is computed (typically between 500 and 2000 [94, 95]); and the resulting distribution is used to estimate the variation that would be expected if the same number of new independent datasets had been collected. The exact interpretation of the statistical significance of bootstrap proportions is elusive, but several authors [95, 96] have proposed that they are conservative measures of support, so a value of 70% or greater might indicate substantial confidence for a group. In Bayesian inference, some criticisms are related to the putative overconfidence of posterior probability measures of node support [97], and the general recommendation is that posterior probabilities should only be considered reliable (strong support) if greater than 0.95 [98-100]. In general, these misleadingly high posteriors are associated with arbitrary resolutions of hard polytomies [101], inappropriate prior choice, and failure to allow convergence [91].

HYPOTHESIS TESTING IN PHYLOGENETICS

One of the most appealing topics in molecular systematics is the availability of methods for the statistical testing of competing phylogenetic hypotheses. These methods are available almost exclusively within the likelihood framework, although some tests have also been developed for other frameworks, such as parsimony [102]. They allow assessment of which model provides the best fit for a given dataset, and the degree of confidence we have in any given topology being the true topology.

One of the methods to compare two competing hypothesis is the likelihood ratio test (LRT) [22, 44, 103], which has been extensively used for selecting competing best-fit models of sequence evolution for a given dataset, and for testing deviations from clock-like evolution (the global molecular clock hypothesis). Competing hypotheses are compared using a statistic, 2δ (calculated as the ratio of the likelihood scores of the alternative hypothesis to the null hypothesis), which measures how much better an explanation of the data the alternative hypothesis gives. In order to perform a significance test, the distribution of 2δ values expected under the simpler hypothesis is required. If the two competing hypotheses are nested (i.e. the null hypothesis is a special case of the alternative hypothesis), then the 2δ distribution is asymptotically distributed as χ^2 with the number of degrees of freedom equal to the difference in the number of parameters between the two models.

When the hypotheses being compared are not nested, the χ^2 approximation may perform poorly. In this case, the null distribution of the LRT statistic can be

approximated by parametric bootstrapping [104-106]. Unlike the non-parametric bootstrap (where datasets are generated by resampling from the original data), the parametric bootstrap uses Monte Carlo simulation to generate the data. Replicate datasets of the same size as the original (usually 200-1000) are simulated according to the null hypothesis being tested. For each replicate dataset, the likelihoods according to both the null and alternative hypotheses are estimated, and the LRT statistic is calculated. These simulated 2δ values form the null distribution of the LRT statistic, allowing implementation of a significance test. The main disadvantage of parametric bootstrapping is that it is computationally demanding, and even unfeasible when large datasets are considered.

Apart from parametric bootstrapping, there are several non-parametric likelihood-based tests. These tests intend to determine whether the difference in fit of two or more alternative tree topologies (always non-nested hypotheses) to the data is significantly greater than expected under the null hypothesis of random sampling error. Of the various methods of this kind, the most widely used are the Kishino-Hasegawa [107], the Shimodaira-Hasegawa [108], and the approximately unbiased [109] tests. They all are based on the estimation of LRT statistics, and use different non-parametric bootstrapping procedures to assess their variance and obtain an estimation of their distribution, thus permitting significance tests. The Kishino-Hasegawa test is valid in the case that the competing hypotheses, each consisting of specific tree topologies, are chosen *a priori*, so that they are not derived from the same data, but this constraint is usually overlooked [92]. The Shimodaira-Hasegawa test can be used to evaluate multiple trees chosen *a posteriori*, allowing a proper multiple comparison even with topologies derived from the same data, but it requires the inclusion of all 'reasonable' trees to be valid because different sets of alternative hypotheses can vary the results [110]. However, it is unclear how the set of reasonable trees can be selected. The approximately unbiased test uses a multiscale bootstrap approach to control for Type 1 errors, while reducing the overly conservative tree selection biases of other non-parametric tests, such as the Shimodaira-Hasegawa test, often accused of being very conservative [92]. However, information about the actual power and appropriateness of this test in empirical cases is still limited. Several concerns exist regarding robustness to deviations from some of its basic assumptions (such as the breakdown of the asymptotic theory [109]), model misspecification [111], and heterogeneity in the rates of sequence evolution (such as the effect of unequal evolutionary rates among taxa [112]). Interestingly, several studies [109, 113] have indicated a good correlation between results from the approximately unbiased and Kishino-Hasegawa tests.

Empirical comparisons of non-parametric and parametric bootstrapping tests appear to indicate that the former tend to be conservative (i.e. unwilling to reject topologies as untrue) because of multiple comparisons and deviations from some of their basic assumptions, and that the latter tend to be liberal (i.e. willing to reject topologies as untrue) because of the use of oversimplified models of

sequence evolution to construct the null distribution [92, 110, 114]. There is no easy way to overcome this conflict, but the trend is to perform both parametric and non-parametric tests, and use the resulting significances to assign 'credibility ranks' for each alternative phylogenetic hypothesis, depending on the concordance or disparity of the two types of test.

ESTIMATION OF DIVERGENCE TIMES

A key feature of molecular phylogenies is that not only can relationships be reconstructed, but divergence events can also be dated using various models of the expected rate of accumulation of substitutions in the sequences over time. The idea of dating evolutionary divergences using calibrated sequence distances was first proposed by Zuckerkandl and Pauling [115], who postulated that the amount of difference between the DNA molecules of two species is a function of the time since their evolutionary separation. This was termed a 'molecular clock' and was shown by comparing amino acid substitution rates with ages estimated from fossils. The central assumption of the molecular clock is that all branches of a phylogenetic tree evolve at the same global substitution rate, i.e. there is rate constancy. A clock-like tree is ultrametric (i.e. the total distance between the root and every tip is constant), so nodal depths can be easily dated if the divergence time for at least one node is known (calibration point). The global rate of substitution can be calculated and, based on it, divergence times for all nodes are estimated by linear regression of the molecular distances [116, 117]. If several calibration points are used, then a regression line is built, the slope of which is an estimate of the global substitution rate, and the divergence times for the unknown nodes are interpolated (or extrapolated). The molecular clock hypothesis is in perfect agreement with the neutral theory of evolution that postulates that the majority of substitutions in genes are the result of the random fixation of selectively neutral mutations [118, 119].

Molecular clocks constructed in this way (conventional) suffer from several limitations that lead to overestimation biases [120, 121]. There is increasing evidence that the assumption of rate constancy is often violated, and that DNA and amino acid sequences of even closely related species can evolve at different rates [122, 123]. The reasons given for these deviations from the clock-like model of sequence evolution are related to generation time [124], metabolic rate [125], mutation rate [126], and the effect of effective population size on the rate of fixation of mutations [124]. In practice, clock-like behaviour of the data can be tested using several methods, the LRT statistic (see above) being the most widely used. If the null hypothesis of a constant rate is rejected, methods that try to model rate changes over the tree, so-called 'relaxed clock methods', are necessary. There are many such methods that use different approaches to either correct or incorporate rate heterogeneity in the dating process on the basis of specific rate change models (for a list see [127]). The biological factors affecting the violation of the mutation rate constancy (see above) are modelled in

a function describing the behaviour of rates throughout the tree; this function can be an algorithm to minimize the rate changes between adjacent branches, or an explicit model of rate variation in which substitution rates can change along branches [128]. Many of these methods are based on the idea of rate autocorrelation: the assumption that evolutionary rates among closely related lineages are similar. Simulations indicate that rate autocorrelation methods are less prone to overestimation than conventional molecular clocks [129]. The most widely-used relaxed clock methods are: penalized likelihood [130, 131], Bayesian rate autocorrelation dating [132-134], and Bayesian uncorrelated relaxed clock [135, 136].

Penalized likelihood is a semi-parametric technique that attempts to simultaneously estimate unknown divergence times and smooth the rapidity of change along lineages. To smooth rate variation, a nonparametric function is used that penalizes rates that change too fast from branch to neighbouring branch, thus reflecting an idea of autocorrelation of rates. Because the penalty function includes unknown times, an optimality criterion based on this penalty permits an estimation of the divergence times [131]. The use of a likelihood framework permits the specification of the relative contribution of the rate smoothing and the data-fitting parts of the estimation procedure. The optimal level of smoothing can be estimated by running a cross-validation procedure [130]. Penalized likelihood provides confidence intervals on the estimated parameters by calculating an age distribution based on chronograms generated from bootstrapped datasets, and allows multiple calibration constraints to permit scaling of rates and times to real units [137].

The Bayesian dating methods use a fully probabilistic and high parametric model to describe the change in evolutionary rate over time, and use MCMC approximation to derive the posterior distribution of rates and times from a prior distribution. In Bayesian rate autocorrelation dating, rates are drawn from a log normal distribution and assigned to different branches in the tree, and a parameter called the Brownian motion constant describes the amount of autocorrelation [132-134]. In order to scale rates and times, the prospective age of the root node must be specified *a priori*. This method provides Bayesian credibility intervals for estimated divergence times and substitution rates, and allows multiple calibration constraints on nodes, specified as prior age intervals. Other approaches of the Bayesian autocorrelated model have also been described [138]. In contrast to penalized likelihood, the Bayesian rate autocorrelation dating method is able to account for multiple genes/loci, or dataset partitions in general, with different evolutionary behaviours. This simultaneous analysis of multiple genes may yield more accurate estimates of divergence times [134]. More recently, the Bayesian uncorrelated relaxed clock method has been developed [135, 136]. This method, which is also highly parametric, assumes no *a priori* correlation of the rates on adjacent branches of the phylogenetic tree, but uses a model in which the rate on each branch is drawn independently and identically from an underlying rate distribution. Although computationally

demanding, this uncorrelated dating method is able to co-estimate phylogeny and divergence times, and it allows the implementation of several models of rate change for the analysis depending on the user's assumptions about how rates change over time.

Even when using similar input topologies and calibration constraints, the results of the penalized likelihood and Bayesian dating methods may look quite different, and this is related to their different assumptions about rate change, their different implementations of models of sequence evolution, branch length estimation, the use of prior information, and the different ways in which confidence intervals are calculated [130, 132, 134, 135]. An important issue in molecular dating is the choice of appropriate calibrations [139-142]. In a conventional molecular clock, they were merely points used for linear interpolation, but the relaxed clock methods have introduced more sophisticated approaches of incorporating calibrations (see [143] for a recent comparison of the various calibration techniques): minimum and/or maximum hard bounds on the age of internal nodes (implemented in penalized likelihood and Bayesian rate autocorrelation dating), soft bounds [144], and prior age distributions for selected nodes in the tree (implemented in the Bayesian uncorrelated relaxed clock method). In recent years, Bayesian methods, particularly the Bayesian uncorrelated relaxed clock method, have been attracting more attention because of their relative novelty compared to penalized likelihood, and because their highly parametric framework, albeit not necessarily more parametric than maximum likelihood methods, is often interpreted as a grade of sophistication that permits the extraction of more information about the evolutionary processes that generated the observed variation.

MOLECULAR MARKERS

Of the various molecular techniques that have been employed in phylogenetic studies [4, 5, 145], the analysis of DNA and/or amino acid sequences has become the most widely used by far nowadays, particularly since the advent of the PCR [146]. It is now possible and relatively easy to determine the precise nucleotide sequence of specific genes or sets of genes for entire groups of organisms, and to use that information to investigate their phylogenetic relationships and molecular evolution. However, the choice of specific genes and taxa that are most appropriate for the phylogenetic question at hand is a crucial step, as the results of the study are largely dependent on the choice [12, 147, 148]. There is an ongoing debate on whether it is better to add more genes or more taxa to increase phylogenetic accuracy and robustness [149, 150]. It is generally accepted that dense taxon sampling improves overall phylogenetic accuracy [12, 151-156]. By contrast, some studies [148, 157, 158] have indicated that dense character (e.g. gene, nucleotide) sampling increases phylogenetic robustness.

In general, the use of 'favourite' genes or genomic regions in phylogenetic studies is more commonly related to the technical ease with which their sequences can be determined, and their 'success' in previous similar-level studies (see [147] for a review). Over the years, ribosomal genes (particularly mitochondrial ones) have been used in animal phylogenetic studies at various taxonomic levels. Also, some mitochondrial protein-coding genes, such as *cob* and *cox1*, have become particularly popular. The growth in popularity of all of these genes was often due to the early availability of 'universal' PCR primers for them [159, 160], but in most cases, only short partial fragments of 300-600 base pairs (bp) of these genes are sequenced. Most of the 'favourite' genes are encoded by mitochondrial DNA, because some of its features (lack of introns, maternal inheritance, practical absence of recombination, and haploidy) have made it particularly suitable for estimating animal molecular phylogenies [145, 161]. Several studies have demonstrated the need to establish high-level phylogenetic inferences based on rather large sequence datasets in order to achieve statistical confidence [11, 162, 163], and in recent years, there has been a growing trend to sequence large genomic regions (or even complete genomes) to tackle phylogenetic problems [10, 164, 165]. In addition, recent studies have also indicated that some orthologous nuclear protein-coding genes, such as *rag1* and *actB*, may outperform mitochondrial sequences in reconstructing ancient phylogenetic relationships [166, 167].

MOLECULAR SYSTEMATICS FOR THE FUTURE

In the coming years, an enormous amount of sequence information, including that of entire genomes [168], will probably become available for a vast array of taxa, particularly with the advent of research initiatives aiming to assemble a single, inclusive tree of all life on Earth, such as the 'Assembling the Tree of Life' program [169]. Therefore, the new challenges to the field of molecular systematics will be mainly related to the handling of very large datasets, and the integration of different levels of genomic information [10, 164, 165, 170, 171]. Several online tools are already available, such as the PhyLoTA browser [172], intended to systematize GenBank information for large-scale molecular phylogenetics analysis. Models of sequence evolution will soon become more realistic and probabilistic algorithms, particularly Bayesian ones [173], will become more sophisticated, while computing times will be reduced with developments in supercomputing and parallel processing technologies (such as the 'Cyberinfrastructure for Phylogenetic Research' project [174]), and the general improvements in processor and other hardware technology. Also, the use of techniques from information theory [175-177] will allow the design of more efficient phylogenetic studies. These techniques can provide quantitative comparisons of the phylogenetic information content of different datasets (genes, data partitions) across a tree or part thereof, which can then be used to

conquer' strategy [191, 192] to explore the huge tree space derived from the analysis of thousands or even millions of taxa. Other approaches that are essentially variants of the supermatrix approach are also being developed [193, 194], but only time will tell if they become essential tools for the new era of molecular systematics.

PHYLOGENETIC SOFTWARE

The number of computer programs available for phylogenetic analysis has greatly increased in recent years, but the reader can find an excellent, comprehensive, well-organized, and up-to-date compilation of most (if not all) of the available programs at Joseph Felsenstein's website [195]. Tab. 1 shows a selection of these programs, implementing most of the methods described in this review.

Tab. 1. An overview of the commonly used programs implementing most of the methods described in this review.

Program	Method	Reference
CLUSTALW	Multiple sequence alignment	[196]
MAFFT	Multiple sequence alignment	[197, 198]
T-COFFEE	Multiple sequence alignment	[199]
GBLOCKS	Selection of conserved blocks from multiple alignments	[200]
MODELTEST	Substitution model selection (DNA)	[201, 202]
PROTTEST	Substitution model selection (proteins)	[203]
PAUP*	Phylogenetic inference (MP, ME, NJ, ML)	[79]
PHYLIP	Phylogenetic inference (MP, ME, NJ, ML)	[204]
MEGA	Phylogenetic inference (MP, ME, NJ)	[205]
RAxML	Phylogenetic inference (fast ML)	[81]
GARLI	Phylogenetic inference (fast ML)	[206]
PHYML	Phylogenetic inference (fast ML)	[84]
MRBAYES	Phylogenetic inference (BI)	[207, 208]
CONSEL	Tree comparison tests	[209]
MACCLADE	Character evolution (MP) / Sequence alignment editor	[210]
MESQUITE	Character evolution (MP, ML)	[211]
PAML	Phylogenetic analysis (ML)	[82, 212]
P4	Phylogenetic analysis (ML, BI)	[213]
R8S	Penalized likelihood dating	[137]
MULTIDIVTIME	Bayesian relaxed-clock dating	[214]
BEAST	Bayesian inference and relaxed-clock dating	[136]
TREEVIEW	Tree visualization and editing	[215]
FIGTREE	Tree visualization and editing	[216]

Acknowledgements. We would like to thank three anonymous reviewers for their insightful comments on an earlier version of this manuscript. Diego San Mauro and Ainhoa Agorreta were respectively supported by a post-doctoral (MEC/Fulbright 2007-0448) and a pre-doctoral (FPU AP2006-00608) fellowship of the Ministry of Science and Innovation of Spain.

REFERENCES

1. Page, R.D.M. and Holmes, E.C. **Molecular evolution: a phylogenetic approach**, Blackwell Science, Oxford, 1998.
2. Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S. and Bhattacharya, T. Timing the Ancestor of the HIV-1 Pandemic Strains. **Science** 288 (2000) 1789-1796.
3. Smith, G.J.D., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S., Peiris, J.S.M., Guan, Y. and Rambaut, A. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. **Nature** 459 (2009) 1122-1125.
4. Rokas, A. and Holland, P.W.H. Rare genomic changes as a tool for phylogenetics. **Trends Ecol. Evol.** 15 (2000) 454-459.
5. Hillis, D.M., Moritz, C. and Mable, B.K., Eds. **Molecular systematics**. Sinauer Associates, Inc., Sunderland, MA, 1996.
6. Hillis, D.M. and Wiens, J.J. Molecules versus morphology in systematics: conflicts, artifacts, and misconceptions. in: **Phylogenetic analysis of morphological data** (Wiens, J.J., Ed.), Smithsonian Institution Press, Washington, DC, 2000, 1-19.
7. Maley, L.E. and Marshall, C.R. The coming of age of molecular systematics. **Science** 279 (1998) 505-506.
8. Stevens, J.R. and Schofield, C.J. Phylogenetics and sequence analysis - some problems for the unwary. **Trends Parasitol.** 19 (2003) 582-588.
9. Doyle, J.J. Gene trees and species trees: molecular systematics as one-character taxonomy. **Syst. Bot.** 17 (1992) 144-163.
10. Rokas, A., Williams, B.L., King, N. and Carroll, S.B. Genome-scale approaches to resolving incongruence in molecular phylogenies. **Nature** 425 (2003) 798-804.
11. Cummings, M.P., Otto, S.P. and Wakeley, J. Sampling properties of DNA sequence data in phylogenetic analysis. **Mol. Biol. Evol.** 12 (1995) 814-822.
12. Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? **Syst. Biol.** 47 (1998) 9-17.
13. Huelsenbeck, J.P. Performance of phylogenetic methods in simulation. . **Syst. Biol.** 44 (1995) 17-48.
14. Maddison, W.P. Gene trees in species trees. **Syst. Biol.** 46 (1997) 523–536.
15. Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. and Hofreiter, M. Genetic analyses from ancient DNA. **Annu. Rev. Genet.** 38 (2004) 645-679.

16. Organ, C.L., Schweitzer, M.H., Zheng, W., Freimark, L.M., Cantley, L.C. and Asara, J.M. Molecular phylogenetics of mastodon and *Tyrannosaurus rex*. **Science** 320 (2008) 499.
17. Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. A plea for DNA taxonomy. **Trends Ecol. Evol.** 18 (2003) 70-74.
18. Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. and Vogler, A.P. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. **Syst. Biol.** 55 (2006) 595-609.
19. Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. Biological identifications through DNA barcodes. **Proc. R. Soc. Lond. B** 270 (2003) 313-321.
20. Swofford, D.L., Olse, G.J., Waddell, P.J. and Hillis, D.M. Phylogenetic inference. in: **Molecular systematics** (Hillis, D.M., Moritz, C. and Mable, B.K., Eds.), Sinnauer Associates, Sunderland, MA, 1996, 407-514.
21. Whelan, S., Liò, P. and Goldman, N. Molecular phylogentics: state-of-the-art methods for looking into the past. **Trends Genet.** 17 (2001) 262-272.
22. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. **J. Mol. Evol.** 17 (1981) 368-376.
23. Posada, D. Selecting models of evolution. in: **The phylogenetic handbook** (Salemi, M. and Vandamme, A.-M., Eds.), Cambridge University Press, Cambridge, 2003, 256-282.
24. Yang, Z. Estimating the pattern of of nucleotide substitution. **J. Mol. Evol.** 39 (1994) 105-111.
25. Fitch, W.M. and Margoliash, E. A method for estimating the number of invariant amino acid coding positions in a gene, using cytochrome c as a model case. **Biochem. Genet.** 1 (1967) 65-71.
26. Wakeley, J. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. **J. Mol. Evol.** 37 (1993) 613-623.
27. Reeves, J.H. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. **J. Mol. Evol.** 35 (1992) 17-31.
28. Hasegawa, M., Kishino, H. and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. **J. Mol. Evol.** 22 (1985) 160-174.
29. Yang, Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. **Mol. Biol. Evol.** 10 (1993) 1396-1401.
30. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. **J. Mol. Evol.** 39 (1994) 306-314.
31. Felsenstein, J. **Inferring phylogenies.** Sinauer Associates, Inc., Sunderland, MA, 2004.
32. Adachi, J. and Hasegawa, M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. **J. Mol. Evol.** 42 (1996) 459-468.

33. Jones, D.T., Taylor, W.R. and Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. **Comp. Appl. Biosci.** 8 (1992) 275-282.
34. Rodríguez, F., Oliver, J.F., Marín, A. and Medina, J.R. The general stochastic model of nucleotide substitution. **J. Theor. Biol.** 142 (1990) 485-501.
35. Ren, F., Tanaka, H. and Yang, Z. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. **Syst. Biol.** 54 (2005) 808-818.
36. Tavaré, S., Adams, D.C., Fedrigo, O. and Naylor, G.J.P. A model for phylogenetic inference using structural and chemical covariates. in: **Pacific Symposium on Biocomputing** (Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K. and Klein, T.E., Eds.), World Scientific, Singapore, 2001, 215-225.
37. Galtier, N. Maximum-likelihood phylogenetic analysis under a covarion-like model. **Mol. Biol. Evol.** 18 (2001) 866-873.
38. Huelsenbeck, J.P. Testing a covariotide model of DNA substitution. **Mol. Biol. Evol.** 19 (2002) 698-707.
39. Cunningham, C.W., Zhu, H. and Hillis, D.M. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. **Evolution** 52 (1998) 978-987.
40. Bruno, W.J. and Halpern, A.L. Topological bias and inconsistency of maximum likelihood using wrong models. **Mol. Biol. Evol.** 16 (1999) 564-566.
41. Huelsenbeck, J.P. and Hillis, D.M. Success of phylogenetic methods in the four-taxon case. **Syst. Biol.** 42 (1993) 247-264.
42. Holder, M. and Lewis, P.O. Phylogeny estimation: traditional and Bayesian approaches. **Nat. Rev. Genet.** 4 (2003) 275-284.
43. Posada, D. and Crandall, K.A. Selecting the best-fit model of nucleotide substitution. **Syst. Biol.** 50 (2001) 580-601.
44. Huelsenbeck, J.P. and Crandall, K.A. Phylogeny estimation and hypothesis testing using maximum likelihood. **Ann. Rev. Ecol. Syst.** 28 (1997) 437-466.
45. Akaike, H. Information theory as an extension of the maximum likelihood principle. in: **Second international symposium of information theory** (Petrov, B.N. and Csaki, F., Eds.), Akademiai Kiado, Budapest, Hungary, 1973.
46. Schwarz, G. Estimating the dimensions of a model. **Ann. Stat.** 6 (1978) 461-464.
47. Posada, D. and Buckley, T.R. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. **Syst. Biol.** 53 (2004) 793-808.
48. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. GenBank. **Nucleic Acids Res.** 35 (2007) D21-D25.
49. Kitching, I.L., Forey, P.L., Humphries, C.J. and Williams, D.M. **Cladistics. ed. 2. The theory and practice of parsimony analysis**, Oxford University Press, Oxford, 1998.

50. Smith, A.B. Rooting molecular trees: problems and strategies. **Biol. J. Linn. Soc.** 51 (1994) 279-292.
51. Phillips, A., Janies, D. and Wheeler, W. Multiple sequence alignment in phylogenetic analysis. **Mol. Phylogenet. Evol.** 16 (2000) 317-330.
52. Goldman, N. Effects of sequence alignment procedures on estimates of phylogeny. **BioEssays** 20 (1998) 287-290.
53. Ogden, T.H. and Rosenberg, M.S. Multiple sequence alignment accuracy and phylogenetic inference. **Syst. Biol.** 55 (2006) 314-328.
54. Edgar, R.C. and Batzoglou, S. Multiple sequence alignment. **Curr. Opin. Struct. Biol.** 16 (2006) 368-373.
55. Notredame, C. Recent evolutions of multiple sequence alignment algorithms. **PLoS Comput. Biol.** 3 (2007) e123.
56. Thompson, J.D., Plewniak, F. and Poch, O. A comprehensive comparison of multiple sequence alignment programs. **Nucleic Acids Res.** 7 (1999) 2682-2690.
57. Feng, D.F. and Doolittle, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **J. Mol. Evol.** 25 (1987) 351-361.
58. Morrison, D.A. Why would phylogeneticists ignore computerized sequence alignment? **Syst. Biol.** 58 (2009) 150-158.
59. Hickson, R.E., Simon, C. and Perrey, S.W. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. **Mol. Biol. Evol.** 17 (2000) 530-539.
60. Hofacker, I.L. Vienna RNA secondary structure server. **Nucleic Acids Res.** 31 (2003) 3429-3431.
61. Wong, K.M., Suchard, M.A. and Huelsenbeck, J.P. Alignment uncertainty and genomic analysis. **Science** 319 (2008) 473-476.
62. Löytynoja, A. and Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. **Science** 320 (2008) 1632-1635.
63. Thorne, J.L., Kishino, H. and Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. **J. Mol. Evol.** 33 (1991) 114-124.
64. Fitch, W.M. Toward defining the course of evolution: minimal change for a specific tree topology. **Syst. Zool.** 20 (1971) 406-416.
65. Farris, J.S. The logical basis of phylogenetic systematics. in: **Advances in Cladistics** (Platnick, N.I. and Funk, V.A., Eds.), Columbia University Press, New York, 1983, 7-36.
66. Hennig, W. **Grundzüge einer theorie der phylogenetischen systematik**, Deutsche Zentral Verlag, Berlin, 1950.
67. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. **Syst. Zool.** 27 (1978) 401-410.
68. Huelsenbeck, J.P. Is Felsenstein zone a fly trap? **Syst. Biol.** 46 (1997) 69-74.
69. Goldman, N. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. **Syst. Zool.** 39 (1990).

70. Cavalli-Sforza, L.L. and Edwards, A.W.F. Phylogenetic analysis: Models and estimation procedures. **Evolution** 21 (1967) 550-570.
71. Fitch, W.M. and Margoliash, E. Construction of phylogenetic trees. **Science** 155 (1967) 279-284.
72. Sneath, P.H.A. and Sokal, R.R. **Numerical taxonomy**, W.H. Freeman, San Francisco, 1973.
73. Saitou, N. and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. **Mol. Biol. Evol.** 4 (1987) 406-425.
74. Rzhetsky, A. and Nei, M. A simple method for estimating and testing minimum-evolution trees. **Mol. Biol. Evol.** 9 (1992) 945-967.
75. Nei, M. and Kumar, S. **Molecular evolution and phylogenetics**, Oxford University Press, Oxford, 2000.
76. Edwards, A.W.F. **Likelihood**, Cambridge University Press, Cambridge, 1972.
77. Edwards, A.W.F. and Cavalli-Sforza, L.L. Reconstruction of evolutionary trees. in: **Phenetic and phylogenetic classification** (Heywood, V.H. and McNeill, J., Eds.), Systematics Association Publ. No. 6, London, 1964, 67-76.
78. Neyman, J. Molecular studies of evolution: a source of novel statistical problems. in: **Statistical decision theory and related topics** (Gupta, S.S. and Yackel, J., Eds.), Academic Press, New York, 1971, 1-27.
79. Swofford, D.L. **PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0**, Sinauer Associates, Inc., Sunderland, MA, USA, 1998.
80. Kosakovsky Pond, S.L., Frost, S.D.W. and Muse, S.V. HyPhy: hypothesis testing using phylogenies. **Bioinformatics** 21 (2005) 676-679.
81. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. **Bioinformatics** 22 (2006) 2688-2690.
82. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. **Mol. Biol. Evol.** 24 (2007) 1586-1591.
83. Yang, Z. How often do wrong models produce better phylogenies? **Mol. Biol. Evol.** 14 (1997) 105-108.
84. Guindon, S. and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. **Syst. Biol.** 52 (2003) 696-704.
85. Huelsenbeck, J.P., Ronquist, F.R., Nielsen, R. and Bollback, J.P. Bayesian inference of phylogeny and its impact on evolutionary biology. **Science** 294 (2001) 2310-2314.
86. Rannala, B. and Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. **J. Mol. Evol.** 43 (1996) 304-311.
87. Larget, B. and Simon, D.L. Markov chain Monet Carlo algorithms for the Bayesian analysis of phylogenetic trees. **Mol. Biol. Evol.** 16 (1999) 750-759.
88. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., Eds. **Markov Chain Monte Carlo in Practice**. Chapman & Hall, London, 1996.

89. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika** 57 (1970) 97-109.
90. Metropolis, N., Rosenbluth, A.W., Teller, A.H. and Teller, E. Equations of state calculations by fast computing machines. **J. Chem. Phys.** 21 (1953) 1087-1091.
91. Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L. and Swofford, D.L. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. **Bioinformatics** 24 (2008) 581-583.
92. Goldman, N., Anderson, J.P. and Rodrigo, A.G. Likelihood-based tests of topologies in phylogenetics. **Syst. Biol.** 49 (2000) 652-670.
93. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. **Evolution** 39 (1985) 783-791.
94. Hedges, S.B. The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. **Mol. Biol. Evol.** 9 (1992) 366-369.
95. Zharkikh, A. and Li, W.-H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. **J. Mol. Evol.** 35 (1992) 356-366.
96. Hillis, D.M. and Bull, J.J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. **Syst. Biol.** 42 (1993) 182-192.
97. Suzuki, Y., Glazko, G.V. and Nei, M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. **Proc. Natl. Acad. Sci. USA** 99 (2002) 16138-16143.
98. Huelsenbeck, J.P. and Rannala, B. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. **Syst. Biol.** 53 (2004) 904-913.
99. Erixon, P., Svennblad, B., Britton, T. and Oxelman, B. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. **Syst. Biol.** 52 (2003) 665-673.
100. Alfaro, M.E., Zoller, S. and Lutzoni, F. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. **Mol. Biol. Evol.** 20 (2003) 255-256.
101. Lewis, P.O., Holder, M.T. and Holsinger, K.E. Polytomies and Bayesian phylogenetic inference. **Syst. Biol.** 54 (2005) 241-253.
102. Templeton, A.R. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of human and the apes. **Evolution** 37 (1983) 221-244.
103. Wilks, S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. **Ann. Math. Statist.** 9 (1938) 60-62.
104. Huelsenbeck, J.P., Hillis, D.M. and Jones, R. Parametric bootstrapping in molecular phylogenetics: Applications and performance. in: **Molecular Zoology: Advances, Strategies, and Protocols** (Ferarris, J.D. and Palumbi, S.R., Eds.), Wiley-Liss, New York, 1996, 19-45.

105. Goldman, N. Statistical tests of models of DNA substitution. **J. Mol. Evol.** 36 (1993) 182-198.
106. Efron, B. Bootstrap confidence intervals for a class of parametric problems. **Biometrika** 72 (1985) 45-58.
107. Kishino, H. and Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. **J. Mol. Evol.** 29 (1989) 170-179.
108. Shimodaira, H. and Hasegawa, M. Multiple comparisons of Log-likelihoods with applications to phylogenetic inference. **Mol. Biol. Evol.** 16 (1999) 1114-1116.
109. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. **Syst. Biol.** 51 (2002) 492-508.
110. Buckley, T.R. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. **Syst. Biol.** 51 (2002) 509-523.
111. Aris-Brosou, S. Least and most powerful tests to elucidate the origin of seed plants in the presence of conflicting signals under misspecified models. **Syst. Biol.** 52 (2003) 781-793.
112. Gissi, C., San Mauro, D., Pesole, G. and Zardoya, R. Mitochondrial phylogeny of Anura (Amphibia): A case study of congruent phylogenetic reconstruction using amino acid and nucleotide characters. **Gene** 366 (2006) 228-237.
113. San Mauro, D., Gower, D.J., Oommen, O.V., Wilkinson, M. and Zardoya, R. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. **Mol. Phylogenet. Evol.** 33 (2004) 413-427.
114. Strimmer, K. and Rambaut, A. Inferring confidence sets of possible misspecified gene trees. **Proc. R. Soc. London B** 269 (2001) 137-142.
115. Zuckerkandl, E. and Pauling, L. Evolutionary divergence and convergence in proteins. in: **Evolving genes and proteins** (Bryson, V. and Vogel, H., Eds.), Academic Press, New York, 1965, 97-166.
116. Li, W.-H. and Graur, D. **Fundamentals of Molecular Evolution**, Sinauer, Sunderland, MA., 1991.
117. Nei, M. **Molecular evolutionary genetics**, Columbia University Press, New York, 1987.
118. Kimura, M. **The neutral theory of molecular evolution**, Cambridge University Press, Cambridge, 1983.
119. Kimura, M. Evolutionary rate at the molecular level. **Nature** 217 (1968) 624-626.
120. Benton, M.J. and Ayala, F.J. Dating the tree of life. **Science** 300 (2003) 1698-1700.
121. Rodríguez-Trelles, F., Tarrío, R. and Ayala, F.J. A methodological bias toward overstimation of molecular evolutionary time scales. **Proc. Natl. Acad. Sci. USA** 99 (2002) 8112-8115.

122. Bromham, L. and Penny, D. The modern molecular clock. **Nat. Rev. Genet.** 4 (2003) 216-224.
123. Wu, C.I. and Li, W.H. Evidence for higher rates of nucleotide substitution in rodents than in man. **Proc. Natl. Acad. Sci. USA** 82 (1985) 1741-1745.
124. Ohta, T. Near-neutrality in evolution of genes and in gene regulation. **Proc. Natl. Acad. Sci. USA** 99 (2002) 16134-16137.
125. Martin, A.P. and Palumbi, S.R. Body size, metabolic rate, generation time and the molecular clock. **Proc. Natl. Acad. Sci. USA** 90 (1993) 4087-4091.
126. Ota, R. and Penny, D. Estimating changes in mutational mechanisms of evolution. **J. Mol. Evol.** 57 (2003) S233-S240.
127. Welch, J.J. and Bromham, L. Molecular dating when rates vary. **Trends Ecol. Evol.** 20 (2005) 320-327.
128. Ho, S.Y.W. An examination of phylogenetic models of substitution rate variation among lineages. **Biol. Lett.** 5 (2009) 421-424.
129. Douzery, E.J.P., Snell, E.A., Baptese, E., Delsuc, F. and Philippe, H. The timing of eukaryotic evolution: Does a realxed molecular clock reconcile proteins and fossils? **Proc. Natl. Acad. Sci. USA** 101 (2004) 15386-15391.
130. Sanderson, M.J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. **Mol. Biol. Evol.** 19 (2002) 101-109.
131. Sanderson, M.J. A nonparametric approach to estimating divergence times in the absence of rate constancy. **Mol. Biol. Evol.** 14 (1997) 1218-1231.
132. Kishino, H., Thorne, J.L. and Bruno, W.J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. **Mol. Biol. Evol.** 18 (2001) 352-361.
133. Thorne, J.L., Kishino, H. and Painter, I.S. Estimating the rate of evolution of the rate of molecular evolution. **Mol. Biol. Evol.** 15 (1998) 1647-1657.
134. Thorne, J.L. and Kishino, H. Divergence time and evolutionary rate estimation with multilocus data. **Syst. Biol.** 51 (2002) 689-702.
135. Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. Relaxed phylogenetics and dating with confidence. **PLoS Biology** 4 (2006) 699-710.
136. Drummond, A.J. and Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evol. Biol.** 7 (2007) 214.
137. Sanderson, M.J. R8S: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. **Bioinformatics** 19 (2003) 301-302.
138. Lepage, T., Bryant, D., Philippe, H. and Lartillot, N. A general comparison of relaxed molecular clock models. **Mol. Biol. Evol.** 24 (2007) 2669-2680.
139. Donoghue, P.C. and Benton, M.J. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. **Trends Ecol. Evol.** 22 (2007) 424-431.
140. Graur, D. and Martin, W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. **Trends Genet.** 20 (2004) 80-86.

141. Hedges, S.B. and Kumar, S. Precision of molecular time estimates. **Trends Genet.** 20 (2004) 242-247.
142. Ho, S.Y.W. Calibrating molecular estimates of substitution rates and divergence times in birds. **J. Avian Biol.** 38 (2007) 409-414.
143. Ho, S.Y.W. and Phillips, M.J. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. **Syst. Biol.** DOI:10.1093/sysbio/syp035 (2009).
144. Yang, Z. and Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. **Mol. Biol. Evol.** 23 (2006) 212-226.
145. Avise, J.C. **Molecular Markers, Natural History and Evolution**, Chapman & Hall, New York, 1994.
146. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. **Science** 239 (1988) 487-491.
147. Cummings, M.P. and Meyer, A. Magic bullets and golden rules: data sampling in molecular phylogenetics. **Zoology** 108 (2005) 329-336.
148. Rokas, A. and Carroll, S.B. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. **Mol. Biol. Evol.** 22 (2005) 1337-1344.
149. Wiens, J.J. Missing data, incomplete taxa, and phylogenetic accuracy. **Syst. Biol.** 52 (2003) 528-538.
150. Wiens, J.J. Missing data and the design of phylogenetic analyses. **J. Biomed. Inform.** 39 (2006) 34-42.
151. Hillis, D.M. Taxonomic sampling, phylogenetic accuracy, and investigatior bias. **Syst. Biol.** 47 (1998) 3-8.
152. Poe, S. and Swofford, D.L. Taxon sampling revisited. **Nature** 398 (1999) 299-300.
153. Pollock, D.D. and Bruno, W.J. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. **Mol. Biol. Evol.** 17 (2000) 1854-1858.
154. Pollock, D.D., Zwickl, D.J., McGuire, J.A. and Hillis, D.M. Increased taxon sampling is advantageous for phylogenetic inference. **Syst. Biol.** 51 (2002) 664-671.
155. Rannala, B., Huelsenbeck, J.P., Yang, Z. and Nielsen, R. Taxon sampling and the accuracy of large phylogenies. **Syst. Biol.** 47 (1998) 702-710.
156. Zwickl, D.J. and Hillis, D.M. Increased taxon sampling greatly reduces phylogenetic error. **Syst. Biol.** 51 (2002) 588-598.
157. Kim, J. Large-scale phylogenies and measuring the performance of phylogenetic estimators. **Syst. Biol.** 47 (1998) 43-60.
158. Rosenberg, M.S. and Kumar, S. Incomplete taxon sampling is not a problem for phylogenetic inference. **Proc. Natl. Acad. Sci. USA** 98 (2001) 10751-10756.

159. Palumbi, S.R., Martin, A., Romano, S., Owen MacMillan, W., Stice, L. and Grabowski, G. **The simple fool's guide to PCR**, Department of Zoology, University of Hawaii, Honolulu, 1991.
160. Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Pääbo, S., Villablanca, F.X. and Wilson, A.C. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. **Proc. Natl. Acad. Sci. USA** 86 (1989) 6196-6200.
161. Ballard, J.W.O. and Rand, D.M. The population biology of mitochondrial DNA and its phylogenetics implications. **Annu. Rev. Ecol. Evol. Syst.** 36 (2005) 621-642.
162. Russo, C.A.M., Takezaki, N. and Nei, M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. **Mol. Biol. Evol.** 13 (1996) 525-536.
163. Zardoya, R. and Meyer, A. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. **Mol. Biol. Evol.** 13 (1996) 933-942.
164. Delsuc, F., Brinkmann, H. and Philippe, H. Phylogenomics and the reconstruction of the tree of life. **Nat. Rev. Genet.** 6 (2005) 361-375.
165. Philippe, H., Delsuc, F., Brinkmann, H. and Lartillot, N. Phylogenomics. **Annu. Rev. Ecol. Evol. Syst.** 36 (2005) 541-562.
166. Springer, M.S., DeBry, R.W., Douady, C.J., Amrine, H.M., Madsen, O., deJong, W.W. and Stanhope, M.J. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. **Mol. Biol. Evol.** 18 (2001) 132-143.
167. Groth, J.G. and Barrowclough, G.F. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. **Mol. Phylogenet. Evol.** 12 (1999) 115-123.
168. Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyrpides, N.C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. **Nucl. Acids. Res.** 34 (2006) D332-D334.
169. AToL initiative. (Assembling the Tree of Life). http://atol.sdsc.edu/.
170. Boore, J.L. The use of genome-level characters for phylogenetic reconstruction. **Trends Ecol. Evol.** 21 (2006) 439-446.
171. Rannala, B. and Yang, Z. Phylogenetic inference using whole genomes. **Annu. Rev. Genomics Hum. Genet.** 9 (2008) 217-231.
172. Sanderson, M.J., Boss, D., Chen, D., Cranston, K.A. and Wehe, A. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. **Syst. Biol.** 57 (2008) 335-346.
173. Beaumont, M.A. and Rannala, B. The Bayesian revolution in genetics. **Nat. Rev. Genet.** 5 (2004) 251-261.
174. Cyberinfrastructure for Phylogenetic Research (CIPRES). Available at http://www.phylo.org/.
175. Goldman, N. Phylogenetic information and experimental design in molecular systematics. **Proc. R. Soc. Lond. B** 265 (1998) 1779-1786.

176. Geuten, K., Massingham, T., Darius, P., Smets, E. and Goldman, N. Experimental design criteria in phylogenetics: where to add taxa. **Syst. Biol.** 56 (2007) 609-622.
177. San Mauro, D., Gower, D.J., Massingham, T., Wilkinson, M., Zardoya, R. and Cotton, J.A. Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear *rag1*. **Syst. Biol.** 58 (2009) 425-438.
178. Cotton, J.A. and Page, R.D.M. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. in: **Phylogenetic supertrees: combining information to reveal the Tree of Life** (Bininda-Emonds, O.R.P., Ed.), Kluwer Academic, Dordrecht, the Netherlands, 2004, 107-125.
179. Maddison, W.P. and Knowles, L.L. Inferring phylogeny despite incomplete lineage sorting. **Syst. Biol.** 55 (2006) 21-30.
180. de Queiroz, A. and Gatesy, J. The supermatrix approach to systematics. **Trends Ecol. Evol.** 22 (2007) 34-41.
181. Kearney, M. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. **Syst. Biol.** 51 (2002) 369–381.
182. Campbell, V. and Lapointe, F.-J. The use and validity of composite taxa in phylogenetic analysis. **Syst. Biol.** 58 (2009) 560-572.
183. Hartmann, S. and Vision, T.J. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? **BMC Evol. Biol.** 8 (2008) 95.
184. Wheeler, W.C. Search-based optimization. **Cladistics** 19 (2003) 348-355.
185. Wheeler, W.C. Homology and the optimization of DNA sequence data. **Cladistics** 17 (2001) S3-S11.
186. Simmons, M.P. Independence of alignment and tree search. **Mol. Phylogenet. Evol.** 31 (2004) 874-879.
187. Bininda-Emonds, O.R.P. The evolution of supertrees. **Trends Ecol. Evol.** 19 (2004) 315-322.
188. Sanderson, M.J., Purvis, A. and Henze, C. Phylogenetic supertrees: assembling the trees of life. **Trends Ecol. Evol.** 13 (1998) 105-109.
189. Gatesy, J., Matthee, C., DeSalle, R. and Hayashi, C. Resolution of a supertree/supermatrix paradox. **Syst. Biol.** 51 (2002) 652-664.
190. Wilkinson, M., Cotton, J.A., Creevey, C., Eulenstein, O., Harris, S.R., Lapointe, F.-J., Levasseur, C., McInerney, J.O., Pisani, D. and Thorley, J.L. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. **Syst. Biol.** 54 (2005) 419-431.
191. Roshan, U., Moret, B.M.E., Williams, T.L. and Warnow, T. Performance of supertree methods on various data set decompositions. in: **Phylogenetic supertrees: Combining information to reveal the Tree of Life** (Bininda-Emonds, O.R.P., Ed.), Kluwer Academic, Dordrecht, The Netherlands, 2004, 301-328.

192. Wilkinson, M. and Cotton, J.A. Supertree methods for building the Tree of Life: Divide-and-conquer approaches to large phylogenetic problems. in: **Reconstructing the Tree of Life. Taxonomy and systematics of species rich taxa** (Hodkinson, T.R. and Parnell, J.A.N., Eds.), The Systematics Association and CRC Press, London, 2007, 61-75.
193. Ren, F., Tanaka, H. and Yang, Z. A likelihood look at the supermatrix–supertree controversy. **Gene** 441 (2009) 119-125.
194. Smith, S.A., Beaulieu, J.M. and Donoghue, M.J. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. **BMC Evol. Biol.** 9 (2009) 37.
195. http://evolution.gs.washington.edu/phylip/software.html.
196. Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. **Nucleic Acids Res.** 22 (1994) 4673-4680.
197. Katoh, K., Kuma, K., Toh, H. and Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. **Nucleic Acids Res.** 33 (2005) 511-518.
198. Katoh, K., Misawa, K., Kuma, K. and Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Res.** 30 (2002) 3059-3066.
199. Notredame, C., Higgins, D.G. and Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. **J. Mol. Biol.** 302 (2000) 205-217.
200. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. **Mol. Biol. Evol.** 17 (2000) 540-552.
201. Posada, D. and Crandall, K.A. MODELTEST: testing the model of DNA substitution. **Bioinformatics** 14 (1998) 817-818.
202. Posada, D. jModelTest: phylogenetic model averaging. **Mol. Biol. Evol.** 25 (2008) 1253-1256.
203. Abascal, F., Zardoya, R. and Posada, D. ProtTest: Selection of best-fit models of protein evolution. **Bioinformatics** 21 (2005) 2104-2105.
204. Felsenstein, J. PHYLIP - Phylogeny inference package (Version 3.2.). **Cladistics** 5 (1989) 164-166.
205. Tamura, K., Dudley, J., Nei, M. and Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. **Mol. Biol. Evol.** 24 (2007) 1596-1599.
206. Zwickl, D.J. (2006) Garli. Available from the author at http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html.
207. Ronquist, F. and Huelsenbeck, J.P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics** 19 (2003) 1572-1574.
208. Huelsenbeck, J.P. and Ronquist, F.R. MRBAYES: Bayesian inference of phylogenetic trees. **Bioinformatics** 17 (2001) 754-755.

209. Shimodaira, H. and Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. **Bioinformatics** 17 (2001) 1246-1247.
210. Maddison, W.P. and Maddison, D.R. **MacClade: analysis of phylogeny and character evolution**, Sinauer Associates Inc., Sunderland, Massachusetts, USA, 1992.
211. Maddison, W.P. and Maddison, D.R. (2009) Mesquite: a modular system for evolutionary analysis. Available from the authors at http://mesquiteproject.org.
212. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. **Comput. Appl. Biosci.** 13 (1997) 555-556.
213. Foster, P.G. (2009) P4. Available from the author at http://www.bmnh.org/~pf/p4.html.
214. Thorne, J.L. and Kishino, H. (2003) Multidivtime. Available from the authors at http://statgen.ncsu.edu/thorne/multidivtime.html.
215. Page, R.D.M. TREEVIEW: An application to display phylogenetic trees on personal computers. **Comp. Appl. Biosci.** 12 (1996) 357-358.
216. Rambaut, A. (2006) FigTree. Available from the author at http://tree.bio.ed.ac.uk/software/figtree/.