J Appl Genet 49(1), 2008, pp. 49–67

Review article

Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods

Borys Wróbel

Department of Marine Genetics and Biotechnology, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland

Abstract. In recent years, the emphasis of theoretical work on phylogenetic inference has shifted from the development of new tree inference methods to the development of methods to measure the statistical support for the topologies. This paper reviews 3 approaches to assign support values to branches in trees obtained in the analysis of molecular sequences: the bootstrap, the Bayesian posterior probabilities for clades, and the interior branch tests. In some circumstances, these methods give different answers. It should not be surprising: their assumptions are different. Thus the interior branch tests assume that a given topology is true and only consider if a particular branch length is longer than zero. If a tree is incorrect, a wrong branch (a low bootstrap or Bayesian support may be an indication) may have a non-zero length. If the substitution model is oversimplified, the length of a branch may be overestimated, and the Bayesian support for the branch may be inflated. The bootstrap, on the other hand, approximates the variance of the data under the real model of sequence evolution, because it involves direct resampling from this data. Thus the discrepancy between the Bayesian support and the bootstrap support may signal model inaccuracy. In practical application, use of all 3 methods is recommended, and if discrepancies are observed, then a careful analysis of their potential origins should be made.

Keywords: Bayesian support values, bootstrap support, interior branch test, phylogenetic inference, topology testing.

Introduction

The tree inference problem

The inference of phylogenetic trees is the most basic tool in the analysis of evolution of genes and the assemblies of cooperating genes (genomes), which characterize the species. From the early 1960s (e.g. Edwards and Cavalli-Sforza 1963), several thousand papers on methods for inferring phylogenies have been published (for a review, see Holder and Lewis 2003 or Felsenstein 2004). These methods either use a greedy approach or search the tree space, possibly heuristically (Figure 1). The output is a single tree (or a set of closely related trees) optimal under the assumptions of the particular method (Figure 2).

For the methods that take as input a matrix of character data (e.g. morphological characters, molecular sequences or gene order data), such an optimality criterion can be maximum parsimony, maximum likelihood or maximum posterior probability (for the Bayesian tree inference methods). Another group of methods, the distance methods, take as input a distance matrix, e.g. DNA hybridization data. The distance matrix is often derived from the character data, which can be, again, molecular sequences (Figure 2) or presence/absence of restriction sites. A special case are genome-rearrangement distances. For the distance methods, the optimality criteria may be minimum evolution or least squares; these criteria are related (Desper and Gascuel 2004). One can also view the minimum evolution criterion (in which the tree with the shortest overall branch lengths is optimal) as conceptually related to maximum parsimony (in which the optimal tree is the one with the smallest amount of character changes). The least

Received: September 13, 2007. Revised: December 18, 2007. Accepted: January 12, 2008.
Correspondence: B. Wróbel, Department of Marine Genetics and Biotechnology, Institute of Oceanology, Polish Academy of Sciences, Powstanców Warszawy 55, 81–712 Sopot, Poland; e-mail: bwrobel@iopan.gda.pl

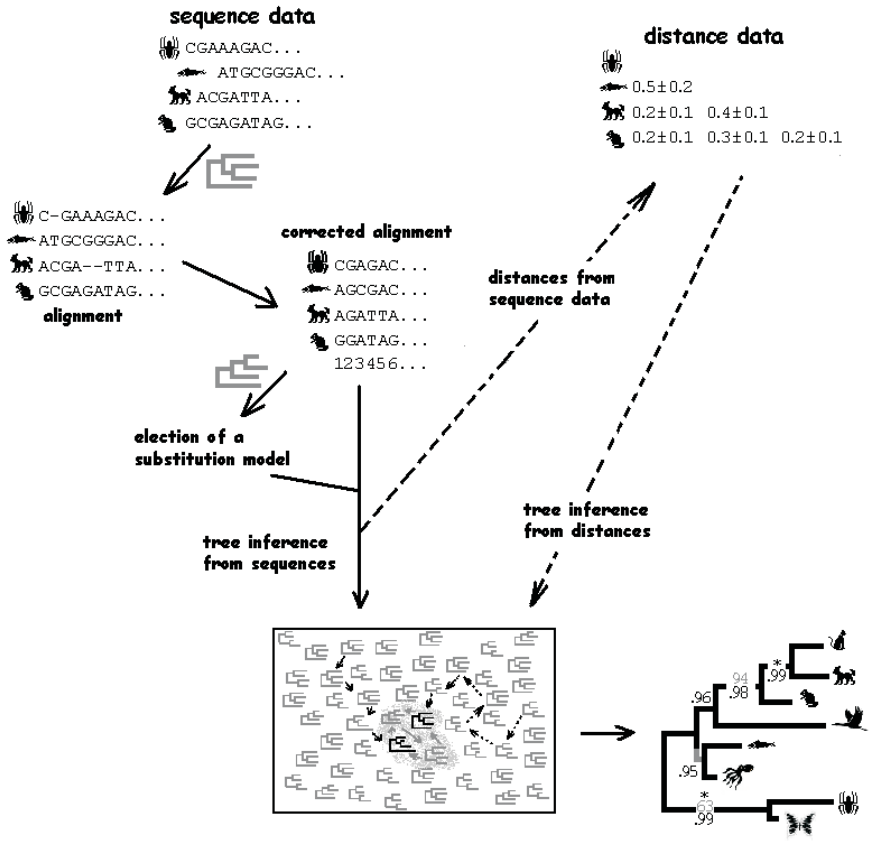

Figure 1. Sources of uncertainty in phylogenetic inference. If molecular sequences (here: nucleotide sequences) are to be used to infer trees, the sequences first have to be aligned properly, so that homologous characters in different sequences are placed in the same columns of the alignment. For multiple sequence alignment with some methods (e.g. progressive algorithms) and for the election of the substitution model, a tree is required. Such a tree may be found by using greedy methods and need not be optimal (grey). For likelihood-based methods, parameters in the model may describe the substitution matrix, character frequencies, and heterogeneity of the substitution rate. Thus the main sources of uncertainty in phylogenetic inference from sequence data are: sequence sampling, alignment errors, and the choice of the model (more parameters than necessary may result in more uncertainty in the value of a given parameter). The actual tree inference involves a search in the universe of possible tree inference (small rectangle; see Figure 2 for details) or else a greedy method can be used. The greedy methods may use a distance matrix derived from the sequences (e.g. maximum likelihood distances by using the elected model). Another type of distance data may be, for example, distances derived from DNA hybridization experiments.

squares solution, on the other hand, is the one that corresponds to the maximum probability of observing the distance matrix (in other words, maximum likelihood; Figure 3).

Some greedy approaches to tree inference can be viewed as implicitly based on optimality criteria. For instance, there exists a relationship between the popular neighbour-joining algorithm (and its various improvements) and the minimum evolution and least squares (Saitou and Nei 1987; Bulmer 1991; Rzhetsky and Nei 1992b, 1993; Gascuel 1997a,b; Gascuel and Steel 2006).

Tuffley and Steel (1997) have shown that for character data the maximum parsimony method is equivalent to maximum likelihood when no common mechanism for the evolution of sites in the sequences is assumed (see also Holmes 2003a and Brandley et al. 2006). When analysing sequence data, distance-based methods, maximum likelihood and the Bayesian methods assume that the nucleotide sites in the analysed sequences evolve according to a certain model of sequence evolution. Commonly it is also assumed that the same model can be used even though the sequences may be gathered from diverse taxa, and – increasingly often – various genomic regions. This is a simplified view. Broadly speaking, the evolution of nucleotide sites is the combination of 2 processes. The first consists of the mutational events that may very well depend on the short-range or long-range neighbourhood in the nucleic acid polymer and the environment in the cell (which may be different in different species). The second process is selection, and the selection pressures at various sites differ. The difference may originate at the level of the informational polymer itself (e.g. sequences that

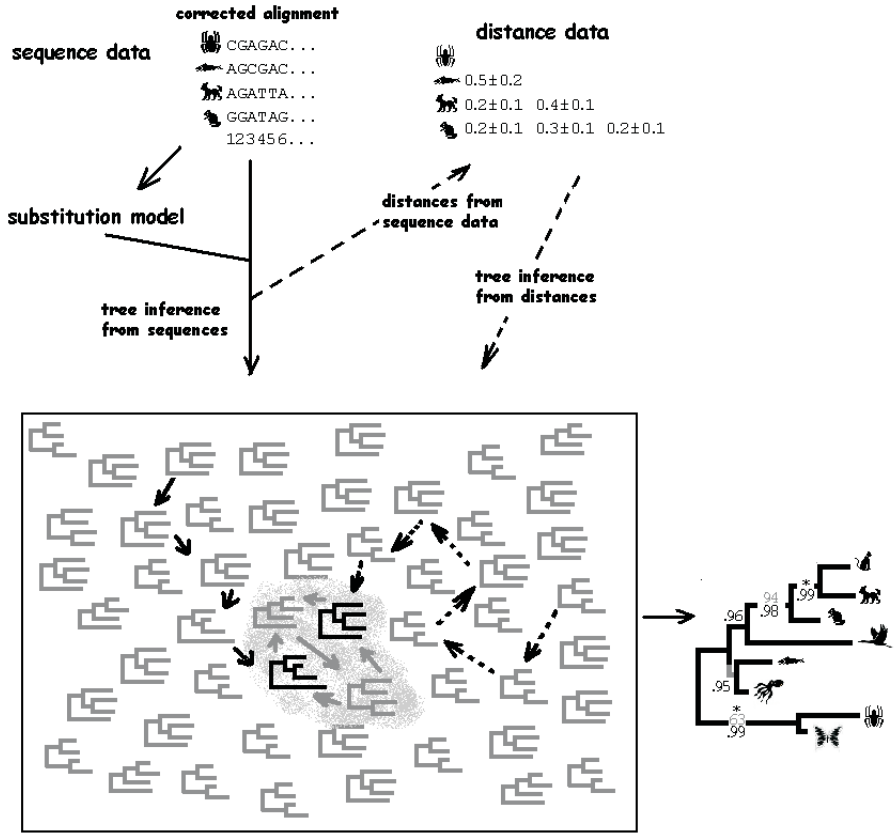

Figure 2. Phylogenetic inference from distance data and molecular sequences. The actual tree inference may involve a greedy tree-building algorithm (not shown) or a search in the treespace (rectangle). The number of possible trees may be very large, and it may be necessary to start from a suboptimal tree (which may be the result of a greedy algorithm) rather than from a random tree. If heuristic methods are used, the moves in the space correspond to topological changes. After each change, a new tree is evaluated under the optimality criterion chosen (i.e. the heuristic). This criterion may be, for example, likelihood (Figure 3), and if so, the value of the substitution model parameters remain fixed at the maximum likelihood values estimated by using a suboptimal tree. The search can be more or less effective, depending on the number of the changes introduced during the search, the extent of the changes, and the permitted decrease in the value of the heuristic. For some data, different starting points (solid and dashed arrows in the rectangle) may result in different trees, perhaps from a large set of trees almost equally well supported by the data (the grey area; these trees may be close to each other, according to the distance measures for trees, or may come from different regions of the tree universe). If instead of character data, distance data are used, the criteria in the heuristic search may be minimum evolution or least squares. If the Markov chain Monte Carlo method is used (for character data), the steps in the chain correspond to the topological changes or to changes in the values of the parameters in the substitution model. In other words, the values of the parameters of the substitution model are not fixed, but the number of the parameters usually remains constant. These moves are always accepted if they result in increased likelihood. If they decrease the likelihood, they are accepted with a probability depending on the magnitude of the decrease. If the method works properly, after a certain number of steps the trees will be sampled from the posterior distribution of trees. The tree sampled most often is the maximum *a posteriori* probability (MAP) tree. The Bayesian branch support is also derived from the same distribution (see Figure 4).

may form particular atypical structures may be selected against) or stem from the effect on the structure and function of the molecules (protein, RNA) that a particular region codes for.

Uncertainty in the phylogenetic inference problem

It can be argued that as for many other quantitative problems, the inference of phylogenies from sequence data is the most powerful in a model-based framework. The situation is in principle similar to the estimation problem more familiar from a basic course in statistics, set in the Euclidean space. Thus, when considering evolutionary trees, it is not only important to obtain an estimate for a parameter or a variable. It is also crucial to have a measure of uncertainty about such an estimate, e.g. the confidence interval or the region of the highest posterior probability density. In other words, it is not only important to obtain the optimal tree but also to capture the uncertainty about the reconstruction.

The uncertainty in estimating the phylogenetic trees originates from the limitations in data sampling and the knowledge about the underlying

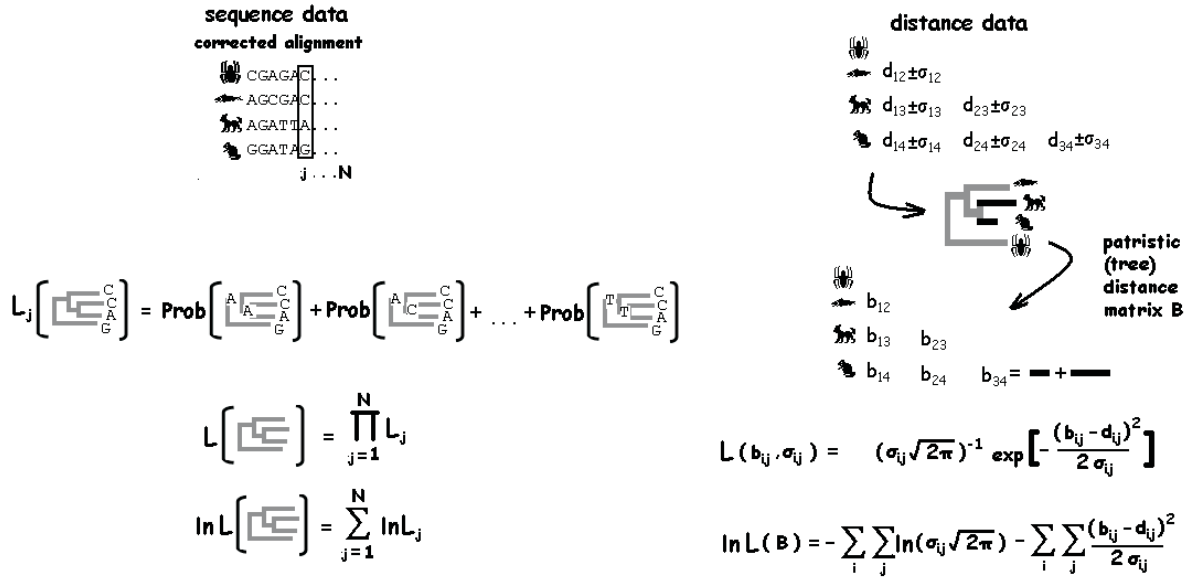

Figure 3. Likelihood of a tree and least squares. In likelihood methods for tree inference from character data, all the sites (the columns in the alignment) are considered to evolve independently. For each column, the likelihood of a topology is the probability of observing a particular character pattern, i.e. the sum over all possible character patterns for the interior nodes, considering the branch lengths and the estimated parameters of the substitution model. Since the sites are independent, the overall likelihood is the product of likelihoods for each site (and the log-likelihood is the sum of the log-likelihoods for each site). The least squares method, on the other hand, considers the likelihood of observing particular distances between the taxa measured on the tree (the patristic distances; the distance between 2 mammals – black on the tree – is shown as an example). The patristic distances are expectations of the distribution of the observed distances. It is assumed that all the distances are normally distributed, and the additional assumption that the distances are independent, simplifies the calculations (first equation on the right; this is the weighted least squares approach, see Sanjuan and Wróbel 2005 for a discussion). Thus the topology for which patristic distances minimize the second sum in the second equation on the right (which shows the sum over all matrix entries) has the highest likelihood (i.e. gives the highest probability of observing the distance data). A tree with a trifurcation (Figure 4) gives a different patristic distance matrix; under the assumptions above, the double difference between these likelihoods follows a χ^2 distribution.

evolutionary process. In maximum parsimony, quite often several most parsimonious trees are found, and maximum likelihood methods sometimes give indistinguishably close likelihoods for many trees. This is because under a given model of sequence evolution (this occurs especially for complex models; Bakke and von Haeseler 1999), the topology may not be identifiable, i.e. there are several different trees that may have nearly equal probabilities of generating the sampled data, higher than the probability for other trees. The question is, then, what topologies could be (nearly) equally supported by the data. If the informational content (the phylogenetic signal) of the data is low, the number of such trees can be high. Presenting just one topology that is optimal – but only slightly better than the alternatives – may be highly misleading.

One way to address the issue of uncertainty in phylogenetic inference is to investigate what topological hypotheses, formulated *a priori*, can be supported by the data. Oftentimes though, many results of an analysis are not part of an explicit *a priori* hypothesis. It is of interest then to summarize the information conveyed by the data, by presenting support for the interior branches in the reconstructed topology.

Since DNA or RNA replication is a bifurcating process, gene trees are expected to be bifurcating. However, it must be noted that for the same group of species, different loci may have different gene trees: this may occur if the actual divergence of the genes occurred in the population of the common ancestor or may be an effect of recombination or hybridization. For some data, therefore, it is impossible to infer the tree because the underlying process is not tree-like, and so the relationships would be better represented by a graph that is not a tree (for a review on these so-called 'phylogenetic networks', see Huson and Bryant 2006).

Limitations of the data sampling and of the estimation procedure may lead to uncertainties about the branching order in gene trees and/or bias in the reconstruction. A particular situation, in which a bias in the reconstruction arises, is a combination

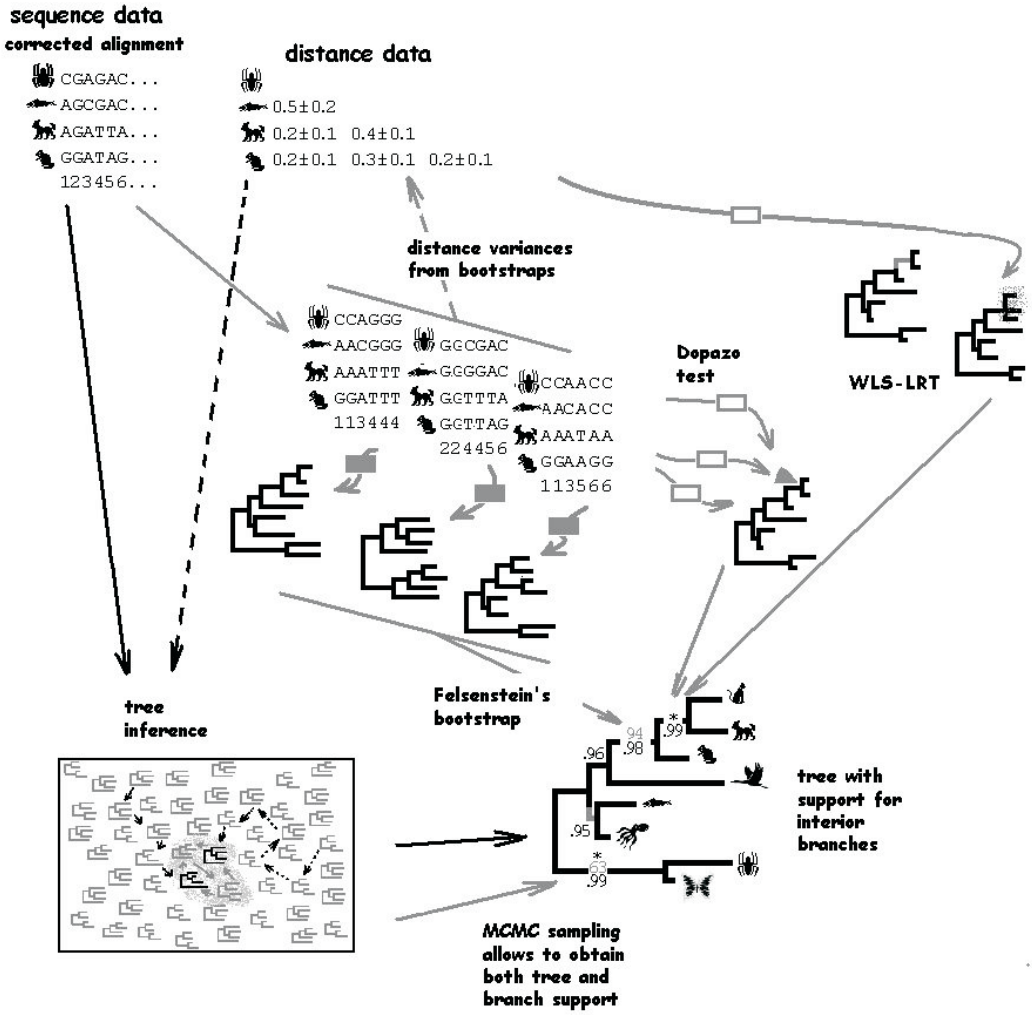

Figure 4. Measuring support for interior branches. If the Markov chain Monte Carlo method is used (see Figure 2), the fraction of the trees in which a given branch occurs in the posterior distribution of trees is the Bayesian support value (black numbers, without the leading zero). Another method to derive support for the bifurcations in the tree involves bootstrapping the character data (sampling with repetition of the columns from the alignment), tree inference for each such pseudosample of sequence data (shown as curved arrows with solid grey rectangles), and counting the number of times a given branch occurs in these trees (e.g. out of 100 bootstraps; grey numbers). Still another group of methods involves testing if the interior branches are significantly longer than zero (the occurrence of zero branch lengths is equivalent to multifurcations in the tree). In the Dopazo test, the distribution for the length of a given branch (grey histogram over a branch) is obtained by using the bootstraps (curved lines with hollow rectangles represent branch estimations). The weighted least squares likelihood ratio test (see the main body of the article and Figure 3 for a detailed description) constructs the test statistics by comparing a bifurcating tree with the tree in which one branch length is constrained to zero (the trifurcation is marked by a grey cloud). This method requires the estimation of distance variances. The variances may be derived from actual measurements or, for character data, by obtaining the distance matrices from the bootstraps. Various methods for measuring support sometimes do not agree: e.g. in the figure for grouping of the arthropods, the bootstrap support is low but the Bayesian support is high and the branch is significantly longer than zero (this is indicated by an asterisk). A particular situation, in which there is a high Bayesian support for very short wrong branches, is sometimes referred as 'a star tree paradox' (in the figure, the grey branch that groups cephalopods and fish).

of long and short branches (long-branch attraction; Felsenstein 1978; Hendy and Penny 1989; Bergsten 2005). Some other particular circumstances, such as homoplasy (similarity that is not due to common ancestry), may affect the reconstruction. In other words, there are conditions under which even a bifurcating process may lead to a non-bifurcating topology.

One of the simplest reasons for uncertainty in the branching order may be the short time scale corresponding to the bifurcation events (which results in short interior branches). If the time was not sufficient to accumulate the nucleotide or amino acid changes, then it may not be possible to resolve the multifurcation. The reasoning presented above is the basis of various so-called 'interior branch

tests', which are concerned if the branches in the reconstructed trees are significantly longer than zero.

Felsenstein's bootstrap and related methods

However, one of the first historically and to date perhaps the most commonly used method to assess the reliability of a branching pattern in phylogenetic trees is Felsenstein's (1985) bootstrap. The bootstrap (Efron 1979), and the related (and older) jackknife (Quenouille 1956; Tukey 1958) are methods to assess the variability of the estimate by repeated estimation using pseudosamples. In the case of jackknife, pseudosamples are formed by dropping observations from the actual sample, while in the case of bootstrap, by resampling from the obtained data with replacement (Figure 4).

For example, let us consider DNA hybridization data, which are usually a collection of repeated measurements of the affinity of the genomic DNA of various species. The appropriate measures of the affinity are, for instance, the inverse normalized percent hybridization values or differences in melting temperatures (Werman et al. 1996). The bootstrap procedure proposed by Krajewski and Dickerman (1990) for such data involves resampling the independent measurements preserving their number (the sample size), and then constructing a pseudoreplicate of the distance matrix by calculating the average. This method was criticised by Marshall (1991), who proposed instead estimating the parameters of a normal distribution of hybridization distance from the sample, and then constructing pseudoreplicates by sampling from this distribution (which makes this approach parametric).

As originally formulated for the molecular sequence data, Felsenstein's (1985) bootstrap considers each character (e.g. a nucleotide site) in a dataset of n characters (columns in the alignment of molecular sequences, i.e. the taxon-by-character data matrix) to be an independent sample. To apply the bootstrap, one samples n times with replacement from the set of n characters. In other words, some sites are sampled several times, and others are omitted (Figure 4). This is consistent with the view that each character is evolving independently from its neighbours. Again, this simplifying – although often false – assumption that the process of sequence evolution is Markovian, underlies most phylogenetic inference methods and models of sequence evolution. A rarely used improvement is the block bootstrap (Künsch 1989): the problem of close range (shorter than an arbitrary length b) correlations between nearby sites in nucleotide sequences is addressed by drawing n/b blocks of b sites, instead of simply n sites.

Character jackknife

The jackknife, another resampling procedure (Farris et al. 1996), has not gained the same popularity as the bootstrap. It can be argued that both are very similar (Felsenstein 2004, p. 339), especially the 'delete-half' jackknife, in which half the characters are drawn from the original data without replacement. One can consider the jackknife and the bootstrap as ways to assign weights to the original characters. In the bootstrap, the weights are the number of times a character is sampled; in the jackknife, the weights take the value 0 or 1, for presence or absence. If so, the delete-half scheme results in a similar coefficient of variation (the ratio of the standard deviation to the mean) for the weights as the bootstrap (1 and approximately 1, respectively). In practice, the delete-half jackknife support values tend to be lower and more variable than bootstrap values (Mort et al. 2000), which might have caused the method to be less popular.

Interpretation of bootstrap support values

Once the pseudoreplicates of the data are obtained, one can use the same method that was used to obtain the tree from the actual data to infer the 'bootstrap trees' (or 'jackknife trees'). The next step is to ask how many times each clade (in other words, a split, a bipartition, an interior branch, or the corresponding node) in the tree reconstructed from the actual data is present in those pseudoreplicate trees. The proportion of times a clade is represented in the bootstrap trees may be thought of as a 'confidence' assessment of the given clade (Figure 4).

The problem with the bootstrap is that it is not clear how the support values should be interpreted. This is related to a practical question of what bootstrap values should in practice be considered as 'high' and 'low' support for the clades, which leads to a widespread confusion as to a convenient 'cut-off value' to consider a clade 'significant'. One view is that no clear-cut 'threshold values' are necessary; the bootstrap support values for particular branches should be considered in the context of all the other branches. Low support indicates that the corresponding clades are the first candidates for suspecting error. This may actually be how most practitioners view the bootstrap values in real analysis (Soltis and Soltis 2003). Low sup-

port overall can be indicative of little 'phylogenetic signal' in the data. Some view the bootstrap as a measure of continuity (how a perturbation of the data affects the estimator; Sanderson 1989; Holmes 2003b).

Still, a high bootstrap value means that the clade has a high probability of recovery, should new data be generated by using the same underlying process (Felsenstein 1985; Hillis and Bull 1993; Berry and Gascuel 1996; Douady et al. 2003). The question is, therefore, if the bootstrap proportions can be interpreted directly as probabilities either in the frequentist or in the Bayesian framework.

Thus even though the bootstrap was developed over 20 years ago, its statistical properties are still controversial. Several different interpretations of the bootstrap have been presented (Berry and Gascuel 1996; Holmes 2003a,b). The original interpretation of Felsenstein (1985) is that the bootstrap proportions are a 'repeatability measure'. In the phylogenetic context, the repeatability is the probability that should another matrix of characters be obtained, a given clade would be reconstructed again. The second interpretation (Sanderson 1989; Zharkikh and Li 1992a,b; Hillis and Bull 1993) is that bootstrap proportions are estimates of the probability that a particular clade is present in the true tree. This is sometimes referred to as 'accuracy'. In the Bayesian framework, this corresponds to the view that the bootstrap values represent the posterior probabilities assuming a flat prior probability distribution (Efron et al. 1996; see below). It has been also postulated (a third interpretation) that they can be used for the purposes of the statistical hypothesis test (Felsenstein and Kishino 1993; Zharkihk and Li 1995; Efron et al. 1996; Anisimova and Gascuel 2006).

These last 2 more ambitious views hang on the consistency of the reconstruction method (the convergence on the true tree as more data are available). If the method is not consistent, the same wrong clades will be found for the pseudo-replicates of the data, which will lead to a high bootstrap proportion, even though the probability of their presence in the true tree is zero.

Corrected bootstrap values

The consensus is that depending on the actual conditions under which the data are generated, the nonparametric bootstrap may overestimate or underestimate phylogenetic accuracy. Several ways to correct the bootstrap proportions have been proposed (Rodrigo 1993; Efron et al. 1996; Zharkihk and Li 1995). Their purpose is to take into account the actual number of alternatives for a given clade and the curvature of the boundaries between the alternatives in the tree space: the number of neighbours of a tree influences the quality of the bootstrap estimate (Zharkikh and Li 1995; Efron et al. 1996; Holmes 2003b). However, these corrections have not gained popularity in the analysis of actual biological data. This could be because they have not been implemented in the popular tree-inference software, such as PHYLIP (Felsenstein 2000) or PAUP* (Swofford 2002). Perhaps more importantly, they are rather computationally costly. For example, the Rodrigo method relies on bootstrapping the bootstrap pseudoreplicates of the original data.

Sanderson and Wojciechowski (2000) demonstrated the usefulness of one of these methods (Efron et al. 1996) to correct for a common problem in bootstrap analysis: the tendency of the bootstrap support for a clade to decrease with increasing sampling of the clade. This behaviour of the bootstrap, observed long ago (Lecointre et al. 1993; Poe 1998), apparently stems mostly from the larger size of the tree space and its different geometry (increased number of the alternatives to the particular tree). Larger spaces make it more difficult to find an optimal tree for the bootstrap pseudoreplicates (a sub-optimal tree may not contain the clade of interest even if the optimal tree would). The corrected support declined much more slowly with the increased number of taxa (Sanderson and Wojciechowski 2000). It may not be possible to correct the problem completely because there is one more effect of increased clade sampling. As sequences are added to a clade, there is a possibility that some actually may have branched out before the previously considered lineages. This would lead to calculating support for a different branch (Sanderson and Wojciechowski 2000).

Bootstrap and the multiple test problem

The additional problem with interpreting the bootstrap support values as *P* values is the multiple test problem (Felsenstein 1985; Holmes 2003a,b). Are corrections, such as the Bonferroni correction, applicable if we are interested in more than one branch? One way to resolve this issue involves a multidimensional approach (Holmes 2003a,b), which – again – is hardly ever used in actual biological analysis.

If, on the other hand, the bootstrap proportions are to be interpreted as probabilities of a correct grouping (i.e. in the Bayesian sense), then the rela-

tionship between the values assigned to different groups is unclear. If they are assumed to be independent, then (for example) if both have 95% support, and the support is interpreted as posterior probabilities, the probability of both being true is about 0.90. As more groups are being considered, the value is quickly diminished. Of course, the branches on a tree correspond to lineages that share common evolutionary histories, so they are not independent. The question of the relationship between the probability values assigned to different clades remains difficult to address, and indeed is hardly ever addressed in bootstrap analyses over entire trees.

Taxon jackknife

In contrast to character pseudosampling, jackknifing across taxa (Lanyon 1985; Siddall 1995; Hovenkamp 2006) involves dropping one species (or more, exploring the 'delete space', Lapointe et al. 1994) for each pseudoreplicate. The critics of such jackknife note (as in the paragraph above) that taxa (genes, etc.) are not independent samples (Felsenstein 1988; Felsenstein 2004, pp. 338 and 357): they form clusters in the phylogeny to be reconstructed. The statistical properties of taxon jackknife are thus even more obscure than the properties of character jackknife/bootstrap. Some interpret, however, all these measures of support as nonprobabilistic (e.g. Helm-Bychowski and Cracraft 1993; Farris et al. 1996; Oxelman et al. 1999). According to this interpretation, the species jackknife can be viewed as an exploratory tool, especially if taxon sampling or long-branch attraction may be suspected to affect the results. In such a situation, however, it would be advisable to use a more directed approach, e.g. just to analyse the effects of removing a particular taxon that may be suspected to cause long-branch attraction (this is a 'long-branch extraction' procedure; Siddall and Whiting 1999).

Other methods to assign support: quartet-puzzling method

The beauty and the appeal of the bootstrap/jackknife method is that it does not depend on the inference procedure used to reconstruct the tree (in the sense that it can be used with any procedure; the actual numerical results will vary, of course). This is in contrast to some methods of assessing the uncertainty about the branches in topologies closely linked to particular reconstruction methods. One such method is based on the number of so-called 'supporting quartets'. The procedure is closely related to a short-course algorithm to obtain a tree (close to the optimal) based on the maximum likelihood (ML) criterion, the quartet-puzzling method (Strimmer and von Haesseler 1996). The first step in the procedure is to consider all possible 4-sequence trees (quartets). In the subsequent puzzling step, starting from one particular 4-leaf tree, the sequences are added one by one in a random order and in a position depending on their position in the quartets. The puzzling step is repeated several times, since in general no single n-taxon tree is consistent with all the quartets. The quartet-puzzling tree (possibly mutlifurcating) is obtained as a majority-rule consensus (Margush and McMorris 1981) of the trees obtained in the puzzling step. The fraction of times a particular clade occurs among these intermediate trees shows the reliability of a given branch. The statistical properties of these reliability values are much less investigated than the bootstrap; it has been noted that although in general they correlate with bootstrap support values (Strimmer and von Haesseler 1996), they are sometimes (perhaps misleadingly) higher (Cao et al. 1998).

The quartet-puzzling algorithm can also be used to derive an overall measure of the phylogenetic signal in the data. This measure is the number of 'unresolved quartets': the 4-sequence sets for which 2 or all 3 possible topologies have very similar likelihoods (from this originated the 'likelihood mapping' method; Strimmer and von Haesseler 1997; Nieselt-Struwe and von Haeseler 2001). Although in principle other criteria (e.g. maximum parsimony, minimum evolution, or least squares) could be used in a similar quartet-puzzling algorithm, this method has not, apparently, been used outside the ML framework.

Bayesian posterior probability support for clades

A technique for assessing the support for interior branches that is intimately linked to a very powerful method of phylogenetic inference is the calculation of the Bayesian posterior clade probabilities (Larget and Simon 1999; Huelsenbeck et al. 2001).

The Bayesian approach for obtaining trees is made possible by the Markov chain Monte Carlo (MCMC) procedure. The approach is related to the likelihood method in the sense that it is based on the model of the evolutionary process: in the case of molecular sequences, the substitution model for the evolution of the nucleotides or amino acids.

In the ML approach to tree inference (Figure 3), one constructs a function that gives the probability of observing the data (the molecular sequences) for given values of the parameters of the substitution model and for a given tree topology with branch lengths (or for a given distance matrix, if one is interested in calculating ML distances). The ML values of the parameters of the substitution model are those for which the likelihood function is at the maximum for a suboptimal tree (usually found by using a greedy approach, Figure 1). These values are kept fixed during the search in the treespace for a topology and branch lengths that maximize the likelihood (Figure 2).

In the Bayesian approach, instead, the idea is to move in the parametric space of trees (and the parameters of the molecular evolution model) by using the Metropolis algorithm (Metropolis et al. 1953). At each step in the sequence of steps, a new value of the parameters (including the topology) is considered in turn (Figure 2). The new value for a given parameter is drawn by using proposal mechanisms, e.g. by drawing from a uniform distribution of predefined width centred on the current value of a simple parameter, or by applying a small topological change in the case of proposal for trees. These topological changes are similar to the topological moves used when heuristically searching for the tree under optimality criteria (e.g. ML; Figure 2), such as nearest neighbour interchange or tree bisection-reconnection.

In the MCMC procedure, the new value is always accepted for the next iteration if it leads to an improvement of the tree likelihood. If the likelihood after introducing a proposed change (L_{new}) is lower than the likelihood calculated by using the previous values (L_{old}), it is accepted with a probability equal to the ratio between the 2 likelihoods ($P_{acceptance} = L_{new}/L_{old}$ if $L_{new} < L_{old}$, assuming that the prior probabilities for the 2 values of the parameter or the topologies are the same and that the possibility to move from the old state to the new state is equal to the possibility of the opposite move). The values of each parameter (and the tree topologies) are recorded at certain intervals (e.g. every 1000 steps). After discarding some steps from the beginning of the chain (this is called burn-in), it is assumed that the remaining values are a sample from the posterior distribution, and so one can calculate its mean, median or mode (in the case of trees: the maximum a posteriori probability tree, i.e. the tree that is sampled most often; Figure 2).

Provided that there is enough information in the data, and that the chain is run long enough, the distributions of the values of the parameters in the samples from the chain will be close to their posterior distributions. A way to summarise the posterior for the trees is to consider how often a particular clade is present in the trees sampled from the chain (Figure 4). This procedure is faster than heuristic searching for trees under the ML criterion for the bootstrap replicates (Figure 4).

Discrepancy between bootstrap support values and Bayesian posterior probabilities

At first it was assumed that the ML bootstrap frequencies and the Bayesian posterior probabilities should be similar (Efron et al. 1996; Durbin et al. 1998 p. 212; Huelsenbeck et al. 2001). The discrepancy between these support measures came as a surprise. The Bayesian probabilities for clades are on average substantially higher than bootstrap support values both in simulations (e.g. Suzuki et al. 2002; Wilcox et al. 2002; Alfaro et al. 2003; Cummings et al. 2003; Douady et al. 2003; Erixon et al. 2003; Mar et al. 2005; Anisimova and Gascuel 2006) and in analyses of actual biological data (e.g. Karol et al. 2001; Murphy et al. 2001; Buckley et al. 2002; Kauff and Lutzoni 2002; Leache and Reeder 2002; Reed, et al. 2002; Streelman et al. 2002; Miller et al. 2002; Whittingham et al. 2002; Collin 2003; DeBry 2003; Douady et al. 2003; Jordan, et al. 2003; Koepfli and Wayne 2003; Misawa and Nei 2003; Shoup and Lewis 2003; Taylor and Piel 2003; Steppan et al. 2004; Mar et al. 2005).

This discrepancy may result from the fact that these methods are based on different models (Alfaro et al. 2003) and so incorporate different sources of uncertainty. The different models could lead to different results. The same applies to differences in assumed prior distributions for the parameters in the same model. This, however, seems less of a problem in an analysis in which the results are dominated by the evidence from the data and the uninformative priors are used (but see Yang and Rannala 2005).

In a recent careful analysis, Svennblad et al. (2006) show that the bootstrap support values approximate the Bayesian posterior probabilities considering only one parameter: the vector of proportions of the possible patterns in the character data (Efron et al. 1996; Efron 2003; Alfaro and Holder 2006). The bootstrap method assumes a non-informative prior on this parameter, but the available approaches to the Bayesian and ML phylogenetic estimation do not use this parameter at all. The relevant parameters in the actual imple-

mentations are rather: the topology, the associated branch lengths, and parameters of the substitution model. Since apparently less patterns are informative for the search for the ML topology than for the search for the maximum posterior probability tree (Svennblad et al. 2006), more patterns in the dataset contribute to the maximum posterior probability tree. This results in lower support for the branches in the ML topology, as estimated by the bootstrap method.

Underestimation and overestimation of the support

Computational experiments show that when the same substitution model is used for the simulation and analysis of the data, the posterior probabilities tend to underestimate the support for clades, although admittedly nonparametric bootstrap underestimates them even more (Wilcox et al. 2002; Erixon et al. 2003). On the other hand, in the analysis of real data, it is rather the possible overconfidence in clades that is seen as a problem. The overconfidence in the Bayesian estimates in the phylogenetic setting is not limited only to clade probabilities (e.g. it affects also the dating of internal nodes, Wróbel et al. 2006). Incidentally, if the values obtained from the MCMC analysis were a perfect measure of clade probability (again, given the priors, the data, and the model), then they indeed should produce higher bounds on clade reliability. This is because they do not incorporate some important sources of uncertainty, e.g. the uncertainty about the alignment and the choice of the substitution model. Indeed, the Bayesian posterior probabilities for clades are more sensitive to model underparametrization than nonparametric bootstrap values (Erixon et al. 2003; Huelsenbeck and Rannala 2004).

The fact that the posterior probabilities are conditional on the model and the data makes them conceptually closer to the parametric bootstrap values (Huelsenbeck et al. 2002), which are calculated from the pseudoreplicates of the data matrix simulated by using the parameters estimated from the actual data rather than by sampling columns in the alignment. Indeed, some authors (e.g. Holmes 2003b) argue that it is more coherent to use parametric bootstrap in the ML framework. Conversely, the bootstrapped Bayesian posterior probabilities (in which the Bayesian analysis is repeated for pseudoreplicates of original data; this is a very computationally-intensive method) correlate much more strongly than the standard Bayesian posterior probabilities with the ML bootstrap values (Douady et al. 2003).

'Star tree paradox'

Some authors also note the increased tendency of the Bayesian method to assign occasionally high posterior probability to very short wrong branches (Figure 4; Alfaro et al. 2003; Douady et al. 2003; Cummings et al. 2003; Lewis, et al. 2005; Yang and Rannala 2005). Alfaro and Holder (2006) noted that this should not be worrying, because anyone using a statistical method is accustomed to accept a certain level of false positives. However, since the pattern frequencies generated by actual star trees ('hard polytomies') due to sampling errors will always be similar to frequencies generated from a tree with short interior branches, it may not be feasible to discriminate between these 2 cases.

In addition, as noted by Lewis et al. (2005), Bayesian inference programs (e.g. MrBayes, see Huelsenbeck and Ronquist 2001, 2003) assign zero prior probability to polytomies. This leads to an arbitrary resolution of the topology with the branch receiving high posterior probability, when in fact the data are most consistent with no mutational changes on a given branch (Lewis et al. 2005). This 'star tree paradox' does not disappear as sequences get longer (Steel and Matsen 2007; see also Kolaczkowski and Thornton 2006). A solution (Lewis et al. 2005), which leads to lowering the wrongly inflated support for such branches, is to give non-zero prior to polytomous tree topologies. The fact that some branches still retain high support after such an analysis (Lewis et al. 2005) suggests that for some branches other effects may be at hand (e.g. model violations; Waddell et al. 2002; Buckley 2002; Buckley et al. 2001, 2002; Huelsenbeck et al. 2002; Douady et al. 2003; Erixon et al. 2003; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004).

To sum up, especially until the Lewis et al.'s (2005) improvement is incorporated into commonly used Bayesian inference software, users should be particularly wary of short branches receiving very high support. It is also advisable to use other methods to measure branch support to detect problematic splits in the tree.

Interior branch tests

There are several (frequentist) methods that allow asking directly the question if particular branches are significantly longer than zero. First of all, it can be done by estimating the variances of interior branch lengths, using a parametric approach. One way to do this requires the computation of the covariance matrix for the distances (Nei et al.

1985; Li 1989; Rzhetsky and Nei 1992a; Tajima 1992), which is computationally difficult (Bulmer 1991; Susko 2003; Sanjuan and Wróbel 2005; Czarna et al. 2006), although the task may be simplified (Sanjuan and Wróbel 2005). The alternative is to estimate the distribution of the branch lengths in a nonparametric manner, by using the bootstrap (Figure 4; Dopazo 1994; Sitnikova 1996). Both methods give similar results when this distribution is close to normal. The advantage of the Dopazo method is that it also has good properties when this is not the case, e.g. when the substitution rate varies among sites (Sitnikova 1996). The null hypothesis in either case is that the length of a particular branch being tested is equal to zero.

Dopazo method

In the Dopazo method, branch length in the topology is being reestimated for each bootstrap (Figure 4). The test has been originally formulated for the distance methods, but could easily be extended to ML, for example. This would require a ML estimate of a branch length for each bootstrap. The fraction of the bootstrap replicates for which the length is positive gives the support for the clade. Sitnikova (1996) offered the appropriate manner in which the *P* values need to be calculated from this fraction for the case when the topology is obtained from the same data that are being used to test the branch. The reasoning is based on the worst-case scenario observed when one interior branch is of zero length and the other branches are long (Sitnikova et al. 1995). The correction is necessary because the distribution conditional on a particular reconstruction constitutes only 1/3 of the unconditional distribution (there are 3 alternative arrangements of subtrees around a branch of interest; Sitnikova 1996). Thus if the *P* value is higher than 2/3, the corrected value is $P' = 3P - 2$; otherwise it is zero (this leads to a deviation from the uniform distribution for *P*' at small values, which is of little importance since the values of interest are close to 0.9).

As was mentioned above, the variances of the branch lengths can be calculated analytically or estimated by bootstrap. They could also be approximated from the curvature of the log-likelihood of the tree as the branch length under test changes (while fitting the lengths of the other branches).

Likelihood ratio test for branches

The ML framework enables another way to test the hypothesis of zero branch lengths: the likelihood ratio test (LRT). A fully resolved unrooted tree has $2n-3$ branches. If we constrain one branch to be of zero length, then the number of branch lengths to be estimated is one less. Conditional on a particular tree, this constructs a nested hypothesis of a trifurcation that can be tested (Figure 3 and 4). The test statistic is thus double difference in maximum log-likelihoods for the bifurcating tree ($\ln L_{bifurcating}$) and the constrained tree ($\ln L_{constrained}$): the ratio of likelihoods corresponds to the difference in log-likelihoods. What is the distribution of the test statistic? Usually, it is assumed that the distribution of the branch lengths in the tree is multivariate normal, and that the likelihood ratio statistic is χ^2-distributed asymptotically: $2(\ln L_{bifurcating} - \ln L_{constrained}) \sim \chi^2$. The number of degrees of freedom for this χ^2 distribution is equal to the difference in the number of parameters between the nested hypothesis and the unrestricted one (here this difference is one). Felsenstein (1988), followed by Li and Gouy (1990), proposed therefore to use this distribution for the purpose of constructing a LRT for branches.

An approximation to such a test is the weighted least squares LRT (Sanjuan and Wróbel 2005; Figures 3 and 4). The procedure requires that the distance matrix and at least some of the variances associated with these distances are known. When the distances are derived from characters, the variances can be estimated by bootstrapping. The test can be applied not only to sequence data, but also to DNA-DNA hybridization data, for example (Wysocka et al. 2006). The procedure ignores the covariances between the distances for computational reasons (the calculation of covariances, as was noted above, is difficult). An additional simplification was proposed by Sanjuan and Wróbel (2005): nonlinear regression of the variances on the corresponding distances allows using just 2 regression parameters in the calculations, instead of the matrix of variances. This simplification is appropriate for sequence data, for example. If one is interested in more than one branch in the topology, corrections for multiple tests can be applied (Sanjuan and Wróbel 2005).

However, it was noted previously (Gaut and Lewis 1995) that the application of the χ^2 distribution with one degree of freedom is not obvious. In the case of LRT for branches, the parameter under consideration takes value zero, which is at the boundary of the parameter space. Gaut and Lewis (1995) note that a mixture of the χ^2 distributions ($^1/_2\chi_0^2 + {}^1/_2\chi_1^2$; Goldman and Whelan 2000; Ota et al. 2000; Anisimova and Gascuel 2006) would

be more appropriate. The χ^2 distribution with 0 degrees of freedom takes zero with probability one. In the mixed distribution, a value is drawn with the probability 0.5 from the χ_0^2 distribution, and with probability 0.5 from the χ^2 distribution with 1 degree of freedom. Thus the mixed distribution results in a less conservative test than when the χ_1^2 distribution is used. That is, it leads to accepting more branches as significantly longer than zero. Simulations show that such a LRT branch test behaves very well in terms of both accuracy and power, unless the model is seriously misspecified (Anisimova and Gascuel 2006).

Nonetheless, the question that remains is whether the tests on the significance of positive branch lengths are appropriate at all. The most obvious criticism is that the expectation of the length of the wrong branch in an incorrect topology can be larger than zero (Sitnikova et al. 1995). For actual datasets, such a situation may be rare (Felsenstein 2004, p. 321). Anisimova and Gascuel (2006) provide an argument that since the interesting question is not whether a branch length is zero but whether it is incorrect, the test statistic should not consider the difference in likelihoods between the unrestricted and restricted topology but instead the difference between the likelihood for a given arrangement around a branch and the second best alternative (out of 3 possible configurations of subtrees around a given branch, corresponding to 3 possible unrooted trees for 4 taxa).

As noted above, for real datasets, the second best topology should not provide a much better fit than the star topology for this branch. For a 4-leaf tree, the least squares expectation for an incorrect branch is actually negative (Sitnikova et al. 1995). For larger incorrect topologies, some branches may have positive expectations, but it has been conjectured (Sitnikova et al. 1995) that at least one expectation will be negative. Assuming that this conjecture is true, a test statistic (such as the weighted least squares statistic) would be higher for a wrong topology than for the unrestricted estimate if negative branch lengths were not allowed. At the same time, the difference for the true topology should be negligible (Susko 2003; Sanjuan and Wróbel 2005).

Interior branch tests and topology testing

If the conjecture discussed above is correct, there exists a relation between the interior branch tests and topology testing. At first, the 2 problems seem quite different. Support values for the clades in a reconstructed topology are a way to summarize the uncertainty of the reconstruction. But it is often the case that there is a certain *a priori* knowledge or hypotheses of the relationships between the analysed objects (taxa, members of different populations, etc.). In molecular systematics, there can be consensus on the relationships between the analysed taxa. In molecular ecology, hypotheses stemming from the knowledge of the physical particularities of the environment may be relevant to the phylogenetic problem at hand. Molecular parasitologists are often interested if the phylogeny of the symbionts corresponds to the phylogeny of the hosts. In molecular epidemiology, a relevant question is whether the inferred relationships between pathogen sequences obtained from patients are consistent with other epidemiological data concerning the exposure of the patients to the pathogens. All these and many other questions can be framed as problems in topology testing.

Kishino-Hasegawa, Shimodaira-Hasegawa and other topology tests

In the Bayesian framework, topology testing is really not very different from branch testing: one only needs to see how often trees conforming to a particular topological constraint occur in the MCMC sample. On the other hand, testing the hypothesis of zero branch length in a particular topology is quite different from testing the question whether 2 different topologies can be supported by the data. If the 2 topologies are specified *a priori*, one can use molecular sequence data and the method popularized by Kishino and Hasegawa (1989) to estimate the confidence interval of the difference in log-likelihoods between 2 distinct topologies. The test assumes that the sites evolve independently and uses bootstrap sampling (for a detailed explanation, see Goldman et al. 2000). Briefly, if the expected likelihood for the 2 topologies is equal, the expectation of the distribution of the difference in log-likelihood for 2 trees calculated at each site will be zero.

If there are at least 2 competing topologies and one of them is the ML tree, then the Shimodaira -Hasegawa (1999) test is more appropriate (Goldman et al. 2000): it adjusts for the fact that the expectation of the difference in likelihood is larger than zero (this is called 'centring'). Other topology tests include the Swofford-Olsen-Wadden-Hillis test (Swofford et al. 1996; Goldman et al. 2000), the expected likelihood weights test (Strimmer and Rambaut 2002), and

the generalized least squares test (Susko 2003; Czarna et al. 2006).

In practice, the topology tests are often employed in the following fashion: first the *a priori* hypotheses are stated and they are formulated in terms of monophyly of certain clades or other topological restrictions (these might be temporal restrictions, see e.g. Wróbel et al. 2006). Then, trees optimal under the commonly used criteria (e.g. ML or least squares) and given these topological constraints are obtained. Finally, whether these trees belong to the confidence set is determined by employing the tests of topologies. This procedure can be employed in assessing clade support by considering each split in turn. First, the best alternative tree (optimal using the criteria at hand, e.g. ML) that is inconsistent with the clade is found. Then, the topology testing methods are employed to determine if the alternative trees belong to the confidence set (Huelsenbeck et al. 1996; Lee 2000). This is similar to the interior branch test formulated by Anisimova and Gascuel (2006), which was discussed above.

Unfortunately, in some situations various tests of topologies give contradictory results (e.g. Goldman et al. 2000; Strimmer and Rambaut 2002; Shi et al. 2005; Czarna et al. 2006). In particular, the Shimodaira-Hasegawa test seems to be too conservative (i.e. its confidence sets for trees are too large), especially in comparison to the Swofford-Olsen-Wadden-Hillis test, which often rejects all but the ML topology.

Such a situation is highly uncomfortable because it makes it possible to reject or accept a hypothesis, depending on one's interests, simply by choosing either a 'more conservative' or a 'less conservative' test. A more careful approach is to report results given by several methods, both when investigating the support of branches in a given topology and when analysing the possible phylogenetic alternatives.

Conclusions

In most papers resolving actual phylogenetic problems it is a common practice to present several phylogenetic reconstructions by using different methods. In fact, it is an informal way to address the question of the uncertainty of the reconstruction. With maximum parsimony methods slowly losing popularity (mainly due to the demonstrable inconsistency of the method), it is common to present the results of an analysis based on ML distances and a greedy approach (such as neighbour joining). Then an attempt is made to improve such a topology by a heuristic search using the ML criterion, and also perhaps a Bayesian approach. Both the heuristic search and the MCMC analysis can start from a random topology, but this may present a problem when the number of leaves is large. Although starting at a random tree should always be attempted if possible, when the number of sequences is large the only practicable approach may be to start in a region of the tree space likely to be close to the global optimum, although with a risk of not exploring the space properly, which is asociated with obvious consequences when calculating Bayesian measures of support for the clades. The same applies to the nonparametric bootstrap in the ML analysis, where it may be necessary for computational reasons to start from a neighbour-joining topology for each pseudoreplicate.

Mostly by virtue of being the first to be proposed and quite easy to implement, the bootstrap remains not only the most commonly used but also the most thoroughly investigated method of assessing the uncertainty in the phylogenetic analysis. One of the important issues that have been investigated is the effect of model under- and overparametrization. Opposite effects of increasing the number of parameters in the substitution model have been reported: an increase in support because of the greater accuracy of tree selection, and a decrease in support due to the higher variance in estimates as the number of parameters increases (Sullivan et al. 1997; Waddell and Steel 1997; Buckley et al. 2001; Buckley and Cunningham 2002). Another, already mentioned issue is the inclusion of additional taxa in the analysis, which – as discussed briefly above – results in decreasing the bootstrap support (Sanderson and Wojciechowski 2000).

A number of studies have been devoted to the effects of the model on the Bayesian posterior probabilities (Buckley 2002; Suzuki et al. 2002; Waddell et al. 2002; Douady et al. 2003; Erixon et al. 2003; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004). The results obtained by using simulations or empirical data, for which a true tree was known, indicate that whereas slight overparametrization seems harmless, the underparametrization is reported to result in greatly inflated support. Thus it is advisable to use complex models, especially when the number of analysed sequences is high. A high number of leaves means that the number of additional parameters com-

pared with a simpler substitution model is small compared with the number of all the estimated parameters, most of which are branch lengths.

When discussing the comparison of the other methods with the bootstrap, it is especially important to note that conclusions drawn in simple cases of 4-taxon trees may not extend to more complex topologies. It is also important to discuss clearly the restrictions of the various approaches, especially whether the theoretical basis of a particular method allows it to be used to assess the support for branches by using the same data as was used to obtain the topology, or whether is it only appropriate to test *a priori* hypotheses.

Another essential point is the proper understanding of the sources of uncertainty that are actually addressed by each method. Obviously, the more uncertainty is allowed to enter in the analysis, the more conservative is the approach. For example, it is expected that nonparametric bootstrap will be more conservative than parametric bootstrap. It has been mentioned above that some authors (e.g. Holmes 2003b) propose that switching to a nonparametric paradigm after using a parametric approach (e.g. ML) is problematic. This would apply also to some of the interior branch tests, e.g. the Dopazo (1994) method or the weighted least squares approach in which the variance of branch lengths is estimated by using bootstrap (Sanjuan and Wróbel 2005).

In the parametric bootstrap, the pseudodata are simulated by using the model estimated from the actual data (i.e. the ML estimates of the parameters of the substitution model and the topology). A different approach, which would incorporate the uncertainty about the model, would allow sampling from the estimated distribution of the parameters and possibly also topologies (e.g. from the confidence set of topologies if not from the set of topologies obtained by using the MCMC algorithm). Still another source of uncertainty, which regains the interest of investigators (e.g. Redelings and Suchard 2005; Lunter et al. 2005), is the uncertainty about the alignment.

When using the phylogenetic analysis in systematic studies, it is sometimes assumed that a highly supported bifurcation in a tree is a suitable basis for (re)classification. However, the measures of support for interior branches can only be viewed as guides. This is because the statistical measures of support may be contradicted by different taxon sampling or different sampling of sequences for the same taxa. Also, these measures depend on the assumed model – and the evolutionary model is always wrong to some degree. In other words, differences between the answers given by different methods are expected, since the assumptions are different. Thus the interior branch tests assume that a given topology is true and only consider a particular branch length. If a tree is incorrect, a wrong branch (a low bootstrap or Bayesian support may be an indication) may have a non-zero length (see above). Similarly, if the substitution model is erroneous (underparametrized), the length of a branch may be overestimated, and the Bayesian support for the branch may be inflated because the acceptance of the topological change proposals (Figure 2) depends directly upon the likelihood under the model. The standard nonparametric bootstrap, on the other hand, approximates the variance of the data under the real model of sequence evolution, because it involves direct resampling from this data. Thus the discrepancy between the Bayesian support and the bootstrap support may signal model inaccuracy.

It is sometimes argued that overestimation of support is more dangerous, since it leads to complacency, while underestimation provides an incentive to collect more data. Provided that the model is correct, the methods are consistent, and the data are not biased, all methods would give absolute support with the increasing amount of data. This is indeed observed in the analysis of large sequence datasets (e.g. Rokas et al. 2003; Dopazo et al. 2004; Dopazo and Dopazo 2005; Edwards et al. 2007). For such data it may be interesting to include additional sources of uncertainty about the reconstruction, mentioned above. However, not all phylogenetic questions can be resolved by collecting longer sequences. In many cases, the interest is in the phylogenies of large assemblies of genes. But many analyses concern, instead, the relationships between single genes or small genomes, such as viral genomes or modules in the genomes (e.g. Wróbel and Węgrzyn 2002).

Thus the investigation of the methods of measuring support remains central for phylogenetic analysis, and so is the investigation of the discrepancies between the results of different methods. Even though the Bayesian approach has the strong appeal of being well funded in the statistical theory, one should not dismiss the frequentist methods. Under certain conditions (such as uninformative priors), ML solutions often coincide with the maximum posterior probability solutions.

Much insight has been gained from the recent flurry of papers discussing the issue of the discrep-

ancy between the bootstrap and the Bayesian support. It has resulted in a deeper understanding of the statistical underpinnings of both methods and improvements in algorithms. This knowledge is useful when analysing the actual data, and it is often suggested to use several methods to explore the range of node support estimates (Douady et al. 2003). This is especially important if the model inadequacy may be an issue, since different methods are sensitive to a different degree to model misspecifications. The observed discrepancies may lead to the use or indeed development of more realistic models of sequence evolution, but the conclusion now is that no perfect method to address the uncertainty in the reconstruction is yet available. In practical applications, therefore, it is recommended that all 3 approaches are used, followed by a detailed discussion of the possible reasons for the discrepancies in the results, if they arise.

Acknowledgments. The author acknowledges the support of the EU Marie Curie Training and Mobility Program (HPMD-CT-2000-00056, MERG-CT-2004-006328), the Polish Ministry of Science and Education (72-6PRUE-2005-7), and the Foundation for Polish Science. The first draft of the paper was written during a research cruise on r/v Oceania. Special thanks go to Aleksandra Czarna, Anna Gambin, Fernando Gonzalez Candelas, and Rafael Sanjuan Verdeguer, for critical reading of the manuscript.

REFERENCES

Alfaro ME, Zoller S, Lutzoni F, 2003. 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: 255–266.

Alfaro ME, Holder MT, 2006. The posterior and the prior in Bayesian phylogenetics Annu Rev Ecol Evol Syst 37:19–42.

Anisimova M, Gascuel O, 2006. Approximate likelihood-ratio test for branches: a fast, accurate and powerful alternative. Syst Biol 55: 539–552.

Bakke E, von Haeseler A, 1999. Distance measures in terms of substitution process. Theor Popul Biol 55: 166–175.

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

Brandley MC, Leache AD, Warren DL, McGuire JA, 2006. Are unequal clade priors problematic for Bayesian phylogenetics? Syst Biol 55: 138–146.

Berry V, Gascuel O, 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol Biol Evol 13: 999–1011.

Buckley TR, 2002. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. Syst Biol 51: 509–523.

Buckley TR, Simon C, Chambers GK, 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol 50: 67–86.

Buckley TR, Cunningham CW, 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol Biol Evol 19: 394–405.

Buckley TR, Arensburger P, Simon C, Chambers GK, 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst Biol 51: 4–18.

Bulmer M, 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol Biol Evol 8: 868–883.

Cao Y, Adachi J, Hasegawa M, 1998. Comment on the quartet puzzling method for finding maximum-likelihood tree topologies. Mol Biol Evol 15: 87–89.

Collin R, 2003. Phylogenetic relationships among calyptraeid gastropods and their implications for the biogeography of marine speciation. Syst Biol 52: 618–640.

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K, 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst Biol 52: 477–487.

Czarna A, Sanjuan R, Gonzalez-Candelas F, Wróbel B, 2006. Topology testing of phylogenies using least squares methods. BMC Evol Biol 6: 105.

DeBry RW, 2003. Identifying conflicting signal in a multigene analysis reveals a highly resolved tree: The phylogeny of Rodentia. Syst Biol 52: 604–617.

Desper R, Gascuel O, 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol Biol Evol 21: 587–598.

Dopazo J, 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J Mol Evol 38: 300–304.

Dopazo H, Dopazo J, 2005. Genome scale evidence for the nematode-arthropod clade. Genome Biol 6: R41.

Dopazo H, Santoyo J, Dopazo J, 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics 20: I116–I121.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJP, 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol 20: 248–254.

Durbin R, Eddy S, Krogh A, Mitchison G, 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge Univ Press: 212.

Edwards AWF, Cavalli-Sforza LL, 1963. The reconstruction of evolution. Ann Hum Gen 27: 105-106.

Edwards SV, Liu L, Pearl DK, 2007. High resolution species trees without concatenation. Proc Natl Acad Sci USA 104: 5936–5941.

Efron B, 1979. Bootstrap methods: another look at the jackknife. Ann Statist 7: 1–26.

Efron B, 2003. Second thoughts on the bootstrap. Stat Sci 18: 135–140.

Efron B, Halloran E, Holmes S, 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 93: 13429–13434.

Erixon P, Svennblad B, Britton T, Oxelman B, 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol 52: 665–673.

Farris JS, Albert VA, Kallersjo M, Lipscomb D, Kluge AG, 1996. Parsimony jackknifing outperforms bootstrapping. Cladistics 12: 99–124.

Felsenstein J, 1978. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol J Linn Soc 16: 183–196.

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. Ann Rev Genet 22: 521–565.

Felsenstein J, 2000. PHYLIP (Phylogeny Inference Package). Distributed by the author, University of Washington, Seattle.

Felsenstein J, 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.

Felsenstein J, Kishino H, 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst Biol 42: 193–200.

Gascuel O, 1997a. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685–695.

Gascuel O, 1997b. Concerning the NJ algorithm and its unweighted version, UNJ. In: Mirkin B, McMorris F, Roberts F, Rhetsky A, eds. Mathematical hierarchies and biology. Providence, RI: American Mathematical Society: 149–170.

Gascuel O, Steel M, 2006. Neighbor joining revealed. Mol Biol Evol 23: 1997–2000.

Gaut BS, Lewis PO, 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12: 152–162.

Goldman N, Anderson JP, Rodrigo AG, 2000. Likelihood-based tests of topologies in phylogenetics. Syst Biol 49: 652–670.

Goldman N, Whelan S, 2000. Statistical tests of gamma distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol 17: 975–978.

Helm-Bychowski K, Crafcraft J, 1993. Recovering phylogenetic signal from DNA sequences: relationships within the corvine assemblage (class Aves) as inferred from complete sequences of the mitochondrial DNA cytochrome-*b* gene. Mol Biol Evol 10: 1196–1214.

Hendy MD, Penny D, 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38: 297–309.

Hillis DM, Bull JJ, 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42: 182–192.

Holder M, Lewis PO, 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Rev Genet 4: 275–284.

Holmes S, 2003a. Bootstrapping phylogenetic trees: theory and methods. Stat Sci 18: 241–255.

Holmes S, 2003b. Statistics for phylogenetic trees. Theor Popul Biol 63: 17-32.

Hovenkamp P, 2006. Can taxon-sampling effects be minimized by using branch supports? Cladistics 22: 264–275.

Huelsenbeck JP, Hillis DM, Nielsen R, 1996. A likelihood-ratio test of monophyly. Syst Biol 45: 546–558.

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

Huelsenbeck JP, Larget B, Miller RE, Ronquist F, 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51: 673–688.

Huelsenbeck JP, Rannala B, 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol 53: 904–913.

Huelsenbeck JP, Ronquist, FR, 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17: 754–755.

Huelsenbeck JP, Ronquist, FR, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

Huson DH, Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.

Jordan S, Simon C, Polhemus D, 2003. Molecular systematics and adaptive radiation of Hawaii's endemic damselfly genus *Megalagrion*. Syst Biol 52: 89–109.

Karol KG, McCourt RM, Cimino MT, Delwiche CF, 2001. The closest living relative of land plants. Science 294: 2351–2353.

Kauff F, Lutzoni F, 2002. Phylogeny of the Gyalectales and Ostropales (Ascomycota, Fungi): among and within order relationships based on nuclear ribosomal RNA small and large subunits. Mol Phylogenet Evol 25: 138–156.

Kishino H, Hasegawa M, 1989. 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: 170–179.

Kolaczkowski B, Thornton JW, 2006. Is there a star tree paradox? Mol Biol Evol 23: 1819–1823.

Koepfli K-P, Wayne RK, 2003. Type I STS markers of more informative than cytochrome *b* in phylogenetic reconstruction of the Mustelidae (Mammalia: Carnivora). Syst Biol 52: 571–593.

Krajewski C, Dickerman AW, 1990. Bootstrap analysis of phylogenetic trees derived from DNA hybridization distances. Syst Zool 39: 383–390.

Künsch H, 1989. The jackknife and the bootstrap for general stationary observations. Ann Statist 17: 1217–1241.

Lanyon SM, 1985. Detecting internal inconsistencies in distance data. Syst Zool 34, 397–403.

Lapointe F-J, Kirsch JAW, Bleiweiss R, 1994. Jackknifing of weighted trees: validation of phylogenies reconstructed from distance matrices. Mol Phylogenet Evol 3: 256–267.

Larget B, Simon D, 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16: 750–759.

Lecointre G, Philippe H, Le HLV, Le Guyader H, 1993. Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol 2: 205–224.

Leache ADT, Reeder W, 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood, and Bayesian approaches. Syst Biol 51: 44–68.

Lee MSY, 2000. Tree robustness and clade significance. Syst Biol 49: 829–836.

Lemmon AR, Moriarty EC, 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst Biol 53: 265–277.

Lewis PO, Holder MT, Holsinger KE, 2005. Polytomies and Bayesian phylogenetic inference. Syst Biol 54: 241–53.

Li W-H, 1989. A statistical test of phylogenies estimated from sequence data. Mol Biol Evol 6: 424–435.

Li W-H, Gouy M, 1990. Statistical tests of molecular phylogenies. Methods Enzymol 183: 645–659.

Lunter GA, Miklós I, Drummond AJ, Jensen JL, Hein J, 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinform 6: 83.

Mar JC, Harlow TJ, Ragan MA, 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol Biol 5: 8.

Margush T, McMorris FR, 1981. Consensus *n*-trees. Bull Math Biol 43: 239–244.

Marshall CR, 1991. Statistical tests and bootstrapping: assessing the reliability of phylogenies based on distance data. Mol Biol Evol 8: 386–391.

Metropolis N, Rosenbluth AE, Rosenbluth MN, Teller AH, Teller E, 1953. Equation of state calculations by fast computing machines. J Chem Phys 21: 1087–1092.

Miller RE, Buckley TR, Manos P, 2002. An examination of the monophyly of morning glory taxa using Bayesian phylogenetic inference. Syst Biol 51: 740–753.

Misawa K, Nei M, 2003. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. J Mol Evol 57: S290–S296.

Mort ME, Soltis PS, Soltis DE, Mabry M, 2000. Comparison of three methods for estimating internal support on phylogenetic trees. Syst Biol 49: 160–171.

Murphy WJ, Eizirik E, O'brien SJ, Madsen O, Scally M, Douady CJ, et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348–2351.

Nei M, Stephens JC, Saitou N, 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol Biol Evol 2: 66–85.

Nieselt-Struwe K, von Haeseler A 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol Biol Evol 18: 1204–1219.

Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H, 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol Biol Evol 17: 798–803.

Oxelman B, Backlund M, Bremer B, 1999. Relationships of the Buddlejaceae *s. l.* investigated using parsimony jackknife and branch support analysis of chloroplast *ndhF* and *rbcL* sequence data. Syst Bot 24: 164–182.

Poe S, 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst Biol 47: 18–31.

Quenouille MH, 1956. Note on bias and estimation. Biometrika 43: 353–360.

Redelings BD, Suchard MA, 2005. Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54: 401–418.

Reed DL, Carpenter KE, deGravelle MJ, 2002. Molecular systematics of the jacks (Perciformes: Carangidae) based on mitochondrial cytochrome *b* sequences using parsimony, likelihood, and Bayesian approaches. Mol Phylogenet Evol 23: 513–524.

Rodrigo AG, 1993. Calibrating the bootstrap test of monophyly. Int J Parasitol 23: 507–514.

Rokas A, Williams BL, King N, Carrol SB, 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.

Rzhetsky A, Nei M, 1992a. A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9: 945–967.

Rzhetsky A, Nei M, 1992b. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J Mol Evol 35: 367–375.

Rzhetsky A, Nei M, 1993. Theoretical foundation of the minimum evolution method of phylogenetic inference. Mol Biol Evol 10: 1073–1095.

Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.

Sanderson MJ, 1989. Confidence limits on phylogenies: the bootstrap revisited. Cladistics 5: 113–129.

Sanderson MJ, Wojciechowski MF, 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-*Astragalus* (Leguminosae). Syst Biol 49: 671–685.

Sanjuan R, Wróbel B, 2005. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. Syst Biol 54: 218–229.

Shi X, Gu H, Susko E, Field C, 2005. The comparison of the confidence regions in phylogeny. Mol Biol Evol 22: 2285–2296.

Shimodaira H, Hasegawa M, 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.

Shoup S, Lewis L, 2003. Polyphyletic origin of parallel basal bodies in swimming cells of chlorophycean green algae (Chlorophyta). J Phycol 39: 789–796.

Siddall ME, 1995. Another monophyly index: revisiting the jackknife. Cladistics 11: 33–56.

Siddall ME, Whiting MF, 1999. Long-branch abstractions. Cladistics 15: 9–24.

Sitnikova T, Rzhetsky A, Nei M, 1995. Interior-branch and bootstrap tests of phylogenetic trees. Mol Biol Evol 12: 319–333.

Sitnikova T, 1996. Bootstrap method of interior-branch test for phylogenetic tress. Mol Biol Evol 13: 605–611.

Soltis PS, Soltis DE, 2003. Applying the bootstrap in phylogeny reconstruction. Stat Sci 18: 256–267.

Steel M, Matsen FA, 2007. The Bayesian "star paradox" persists for long fine sequences. Mol Biol Evol 24:1075–1079.

Steppan SJ, Adkins RM, Anderson J, 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst Biol 53: 533–553.

Streelman JT, Alfaro ME, Westneat MW, Bellwood DR, Karl SA, 2002. Evolutionary history of the parrotfishes: biogeography, ecomorphology, and comparative diversity. Evolution 56: 961–971.

Strimmer K, von Haeseler A, 1996. Quartet Puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13: 964–969.

Strimmer K, von Haeseler A, 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA 94: 6815–6819.

Strimmer K, Rambaut A, 2002. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond Ser B 269: 137–142.

Sullivan J, Markert JA, Kilpatrick CW, 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst Biol 46: 426–440.

Susko E, 2003. Confidence regions and hypothesis tests for topologies using generalized least squares. Mol Biol Evol 20: 862–868.

Suzuki Y, Glazko GV, Nei M, 2002. Overcredibility of molecular phylogenetics obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA 99: 16138–16143.

Svennblad B, Erixon P, Oxelman B, Britton T, 2006. Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics. Syst Biol 55L: 116–121.

Swofford DL, 2002. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Sunderland, MA: Sinauer Associates.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM, 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, eds. Molecular systematics, Sunderland, MA: Sinauer Associates: 407–514.

Tajima F, 1992. Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. Mol Biol Evol 9: 168–181.

Taylor DJ, Piel WH, 2003. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol 21: 1534–1537.

Tuffley C, Steel M, 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 59: 581–607.

Tukey JW, 1958. Bias and confidence in no quite large samples. Ann Math Stat 29: 614.

Waddell PJ, Kishino H, Ota R, 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform 13: 82–92.

Waddell PJ, Steel MA, 1997. General timereversible distances with unequal rates across sites: Mixing G and inverse Gaussian distributions with invariant sites. Mol Phylogenet Evol 8: 398–414.

Werman SD, Springer MS, Britten RJ, 1996. Nucleic acids I: DNA-DNA hybridization. In: Hillis DM, Moritz C, Mable BK, eds. Molecular systematics, Sunderland, MA: Sinauer Associates: 169–203.

Whittingham LA, Slikas B, Winkler DW, Sheldon FH, 2002. Phylogeny of the tree swallow genus, *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol Phylogenet Evol 22: 430–441.

Wilcox TP, Zwickl DJ, Heath TA, Hillis DM, 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol Phylogenet Evol 25: 361–371.

Wróbel B, Torres-Puente M, Jimenez N, Bracho M, Garcia-Robles I, Moya A, Gonzalez-Candelas F, 2006. Analysis of the overdispersed clock in the short-term evolution of Hepatitis C Virus: Using the E1/E2 gene sequences to infer infection dates in a single source outbreak. Mol Biol Evol 23: 1242–1251.

Wróbel B, Wegrzyn G, 2002. Evolution of lambdoid replication modules. Virus Genes 24: 163–171.

Wysocka A, Konopa G, Wegrzyn G, Wróbel B, 2006. Genomic DNA hybridization as an attempt to evalu-

ate phylogenetic relationships of Ostracoda. Crustaceana 79: 1309–1322.

Yang Z, Rannala B, 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol 54: 455–470.

Zharkikh A, Li W-H, 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol Biol Evol 9: 1119–1147.

Zharkikh A, Li W-H, 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J Mol Evol 35: 356–366.

Zharkikh A, Li W-H, 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol Phyl Evol 4: 44–63.