Syst. Biol. 53(6):904–913, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490522629

Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models

JOHN P. HUELSENBECK[1] AND BRUCE RANNALA[2]

[1]*Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093-0116, USA*
E-mail: johnh@biomail.ucsd.edu
[2]*Genome Center & Section of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, California 95616, USA*

Abstract.— What does the posterior probability of a phylogenetic tree mean? This simulation study shows that Bayesian posterior probabilities have the meaning that is typically ascribed to them; the posterior probability of a tree is the probability that the tree is correct, assuming that the model is correct. At the same time, the Bayesian method can be sensitive to model misspecification, and the sensitivity of the Bayesian method appears to be greater than the sensitivity of the nonparametric bootstrap method (using maximum likelihood to estimate trees). Although the estimates of phylogeny obtained by use of the method of maximum likelihood or the Bayesian method are likely to be similar, the assessment of the uncertainty of inferred trees via either bootstrapping (for maximum likelihood estimates) or posterior probabilities (for Bayesian estimates) is not likely to be the same. We suggest that the Bayesian method be implemented with the most complex models of those currently available, as this should reduce the chance that the method will concentrate too much probability on too few trees. [Bayesian estimation; Markov chain Monte Carlo; posterior probability; prior probability.]

Quantifying the uncertainty of a phylogenetic estimate is at least as important a goal as obtaining the phylogenetic estimate itself. Measures of phylogenetic reliability not only point out what parts of a tree can be trusted when interpreting the evolution of a group, but can guide future efforts in data collection that can help resolve remaining uncertainties. The reliability of a tree can be assessed using frequentist and Bayesian approaches. In the frequentist approach, the reliability of a phylogenetic tree could be described by a confidence set of trees, in which the confidence level or coverage probability would be $100(1-\alpha)\%$ (here α is a number between 0 and 1 that controls the size of the confidence set of trees). The interpretation of a 95% ($\alpha = 0.05$) confidence set of trees (simply a list of trees) requires a thought experiment: if we could take random samples of the same size as the original data matrix and create confidence sets of trees in the same way for each sample, then 95% of those confidence sets would contain the true tree. An alternative interpretation is that if an investigator calculated a $100(1-\alpha)\%$ confidence set of trees every time he or she performed a phylogenetic analysis, then at most a proportion α of them would not contain the true tree over the investigator's lifetime. In practice, phylogeneticists do not calculate confidence sets of trees, but assign measures of phylogenetic reliability to individual branches of a phylogenetic tree. The general idea, however, is the same; if the investigator's confidence in a particular group is 95%, then ideally the grouping of taxa would be incorrect 5% of the time (on repeated sampling). Assessing the uncertainty in individual clades on a tree has the benefit that it allows the systematist to evaluate specific hypotheses of monophyly.

The task of developing good frequentist measures of phylogenetic reliability, however, has proven particularly difficult, and has resulted in many different approaches being taken. One group of methods calculates a test statistic for a branch on the phylogeny, such as the Bremer decay index (Bremer, 1988, 1994), and then uses permutation procedures to generate a null distribution for the test statistic. The T-PTP test (Faith, 1991), for example, follows this strategy and aims at testing the hypothesis of monophyly for a group. The problem with this approach, however, is that the null distribution for the test statistic has nothing to do with the null hypothesis of monophyly (Swofford et al., 1996; DeBry, 2001); if a group truly were monophyletic, then one would not expect the observations supporting that monophyly to be random. An alternative method for assessing the reliability of a phylogenetic group involves testing the null hypothesis that the branch supporting the monophyly of the group is zero in length. Although this type of test is clearly related to the support of a group (a long branch subtending a group on a tree means that many characters evolved on that branch), it does not directly assess the reliability of a group.

The most widely used, and generally accepted, method for assessing phylogenetic reliability is the nonparametric bootstrap method (Efron, 1979; Efron and Tibshirani, 1993) introduced to the phylogenetics literature by Felsenstein (1985). The nonparametric bootstrap method constructs new bootstrapped data matrices of the same size as the original data matrix by randomly sampling characters (e.g., columns in an alignment of DNA sequences) with replacement. The same method of estimating phylogeny that was applied in the analysis of the original data matrix is then applied to each of the bootstrapped data matrices. The fraction of the time a particular clade appears in analysis of the bootstrapped data sets represents the confidence value for that grouping of taxa; if a group appeared in 88% of the analyses of the bootstrapped data matrices, then its confidence level is approximately 88%. The jackknife method is similar in spirit to the bootstrap method (Mueller and Ayala, 1982; Farris et al., 1996). Instead of resampling the original data matrix with replacement, however, a fixed number of characters (columns) are randomly

deleted from the original analysis in construction of the jackknifed data matrices. Each of these jackknifed data matrices is then analyzed using the same phylogenetic method as was applied to the original data. Although the end goal of the jackknife method is the same as the bootstrap method, it is not clear what fraction of the sites should be dropped in any particular analysis so that the jackknife fraction for a group matches the desired confidence level (Farris et al., 1996; Felsenstein, 2003).

There are three problems with using the nonparametric bootstrap method to assess phylogenetic reliability. None of these problems is necessarily fatal, but taken together they greatly complicate the application of the bootstrap method in phylogenetics and the interpretation of bootstrap proportions on trees. The first problem concerns the computational complexity of the method, especially when applied with the method of maximum likelihood to estimate phylogeny. Simply put, bootstrap analyses can be computationally prohibitive when the models of sequence analysis are complicated or the data sets are large. This problem can be alleviated by clever programming and fast computers, but bootstrap analyses will always take many times longer than analysis of the original data. Another problem is that it is difficult to apply the bootstrap method under some sorts of models. For example, the standard bootstrap method in which individual sites are sampled is inappropriate for models that incorporate correlation in rates (Yang, 1995; Felsenstein and Churchill, 1996) because the sampling procedure destroys the autocorrelation structure of rates in the bootstrap matrices. Potentially, this problem can be dealt with by sampling blocks of characters (Künsch, 1989), but this solution remains unexplored. Finally, the interpretation of the bootstrap proportion as applied in phylogenetics is problematic.

In most statistical applications, the bootstrap procedure is used to approximate the sampling distribution of a statistic. We imagine that we have some observations from which we estimate the value of some parameter that has a true value of θ_T. The estimated value of the parameter is denoted $\hat{\theta}$. The observations are treated as random variables (variables that take their values by chance), so that the parameter, which is a function of the observations, is also a random variable.The probability distribution of the parameter estimate is called the sampling distribution, and several very useful quantities can be calculated from it. Importantly, the sampling distribution can be used to assess the variability in an estimate. For example, one can construct a confidence interval, which is a range of parameter values that contains the true value of the parameter with some prespecified coverage probability (usually 95%). The bootstrap provides a computer intensive, but straightforward way to approximate the sampling distribution of a parameter even when the modeling assumptions are quite complex. Unfortunately, the application of the bootstrap in phylogenetics is not as simple as it is for most statistical problems. One of the main problems is that there is not a natural way to measure the variability of trees (Holmes, 2003). Another problem is that the parameter estimates change discontinously; small changes in the data can result in different tree topologies being chosen as best. The bootstrap proportions have been variously interpreted as the probability that a clade is correct (a notion explored by Hillis and Bull, 1993), the robustness of the results of a phylogenetic analysis to perturbation (Holmes, 2003), and the probability of incorrectly rejecting a hypothesis of monophyly (Felsenstein and Kishino, 1993). Hillis and Bull (1993) performed a simulation study that investigated the coverage probability of the bootstrap proportion. In their simulation analyses, the bootstrap proportion for a clade usually underestimated the true probability that the clade is correct. The general results of the Hillis and Bull study have been replicated in other studies (e.g., Alfaro et al., 2002; Zharkikh and Li, 1992a, 1992b). The bootstrap proportions can be interpreted in a hypothesis testing framework. In this case, the hypothesis to be tested is one of monophyly for a group. One minus the bootstrap proportion for a clade can be interpreted as the probability of incorrectly rejecting the hypothesis of monophyly. The iterated bootstrap (Rodrigo, 1993), full-and-partial bootstrap (Zharkikh and Li, 1995), and corrected bootstrap of Efron et al. (1996) all attempt to improve the accuracy of the bootstrap as an estimate of one minus the probability of incorrectly rejecting a hypothesis of monophyly (Sanderson and Wojciechowski, 2000). These correction procedures, however, add considerable complexity to the bootstrap procedure. To date, only Sanderson and Wojciechowski (2000) have applied the corrected bootstrap, and only to a single clade. Although the best interpretation of the bootstrap proportion on a phylogenetic tree appears to be in a hypothesis testing framework, where the bootstrap proportion is related to the type I error, most practicing systematists appear to interpret bootstrap support as the probability that the clade is correct. Although this interpretation and the hypothesis testing one are clearly related (a high bootstrap proportion should be associated with correct clades more often than low bootstrap proportions), the questions are different. To our knowledge, only the Bayesian approach directly addresses the probability that a clade is correct, conditional on the observations.

The methods discussed above all use the frequentist interpretation of probability. The complication in frequentist statistics is that the parameter (e.g., phylogenetic tree) is considered to take a fixed but unknown value; the parameter is not treated as a random variable, and hence cannot be directly assigned a probability. To obtain an idea of the variability of an estimated phylogeny, then, one must resort to the thought experiment of sampling data sets and reconstructing phylogeny on each. The distribution of phylogenetic trees obtained in this manner is an approximation of the sampling distribution of phylogeny. In a Bayesian analysis, on the other hand, the parameters are treated as random variables and can be directly assigned probabilities. Bayesian inference of phylogeny suggests a natural way to assess the uncertainty in a phylogeny: the probability that a tree is correct is simply the posterior probability of the tree. The posterior probability of a tree is conditioned on the data

and the model used in the analysis both being correct. The use of posterior probabilities has some intuitive appeal. For one, the interpretation of posterior probabilities is direct and simple; one does not need to invoke a thought experiment of repeated sampling to interpret the results of a Bayesian analysis. More practically, posterior probabilities can be approximated using Markov chain Monte Carlo even under complex models of substitution in a fraction of the time that would be required for maximum likelihood bootstrapping (the closest analog to a Bayesian analysis of phylogeny; Larget and Simon, 1999).

The proper interpretation of posterior probabilities on trees has recently attracted a lot of attention from biologists. Many systematists have noted that posterior probabilities on clades tend to be higher than the bootstrap proportions on the same clades (Douady et al., 2003; Erixon et al., 2003; but see Cummings et al., 2003), and Murphy et al. (2001) speculated that posterior probabilities do not suffer from being too conservative like the nonparametric bootstrap (Hillis and Bull, 1993). The results of simulation studies comparing posterior probabilities and bootstrap proportions have been mixed. One important problem that can be addressed in simulation is the robustness of posterior probabilities to violation of model assumptions. This can be done by simulating data under a complicated model of evolution, and then using an incorrect (oversimplified) model of evolution in the Bayesian analysis. Suzuki et al. (2002) took this approach and argued that posterior probabilities are more sensitive to model misspecification than the bootstrap method. Their method for violating the assumptions of the method, however, was extreme; they simulated data sets on the three possible trees for four species, and then concatenated the data matrices. This means that each third of the concatenated data set had a different underlying phylogenetic tree. The Bayesian analysis, however, assumed a common tree for the data matrix. Although there are many cases in which evolutionary biologists think that different phylogenies may underlie different parts of the genome (e.g., for coalescence trees), this type of model violation is extreme and does not mimic the more universal concern that the substitution model assumed in the analysis is incorrect. Indeed, analysis of model adequacy suggests that models currently used in phylogenetic analysis fail to capture important evolutionary processes (Goldman, 1993). It is also important to analyze the behavior of Bayesian posterior probabilities (and other methods for assessing phylogenetic reliability) when all of the assumptions of the analysis are satisfied. This gives a "best case" picture of the statistical behavior of a method. It is already known that it is difficult to interpret the uncorrected bootstrap support for a clade as the probability that the clade is correct (Hillis and Bull, 1993; Holmes, 2003), even when all of the assumptions of the method are satisfied. A few studies have examined the statistical properties of posterior probabilities when the assumptions of the Bayesian analysis are satisfied. Lemmon and Moriarty (2004) examined the behavior of posterior probabilities in simulation. They examined cases in which the Bayesian model was over- and underspecified. In general, underspecification of the phylogenetic modelled to biased estimates of parameters. Alfaro et al. (2003) and Wilcox et al. (2002) compared posterior probabilities and bootstrap proportions in simulation and found that posterior probabilities gave a more accurate representation of phylogenetic confidence than the bootstrap method (when the assumptions of the method were satisfied). In general, a posterior probability of, say, 0.91 was more likely to correspond to a probability that the tree was correct of 0.91 than the bootstrap method. Importantly, the posterior probabilities did not perfectly match the probability that a clade is correct in these simulations. For example, Wilcox et al. (2002: 369) point out that "...Bayesian support values provide much closer estimates of phylogenetic accuracy (even though they are still somewhat conservative) than the estimates provided by corresponding bootstrap proportions." A similar result is seen in Figure 4 of Alfaro et al. (2003). This result is worrisome, because it suggests that posterior probabilities may not have the meaning that is ascribed to them (i.e., that the posterior probability of a clade is the probability that the clade is correct).

In this study, we examine the statistical properties of Bayesian posterior probabilities on small (six taxon) trees. We examine the behavior of the method when all of the assumptions are satisfied, and also when the assumptions of the method are violated. We show that posterior probabilities do have a well-defined and easily interpreted meaning when the assumptions of the method are satisfied. We also show that posterior probabilities can be sensitive to model misspecification, suggesting that care should be taken to use models of sequence evolution that are as realistic as possible in Bayesian analysis.

METHODS

We evaluate the statistical properties of posterior probabilities using computer simulation. In a traditional computer simulation, one fixes parameters of the phylogenetic model or varies them systematically over a parameter space (e.g., Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995). For example, one might decide to evaluate the behavior of some phylogenetic method on many data matrices that were simulated on a common tree and set of branch lengths. This traditional simulation procedure is how earlier studies examined the properties of Bayesian posterior probabilities (Alfaro et al., 2003; Suzuki et al., 2002; Wilcox et al., 2002). However, if the goal is to study posterior probabilities when all of the assumptions of the Bayesian analysis are satisfied, then the traditional approach cannot be used. The model in a Bayesian analysis of phylogeny has two parts: One part involves assumptions about how the substitutions occur on the tree; the other part of the model describes the prior probability distribution of the parameters. The traditional simulation approach does not treat the prior of the Bayesian analysis seriously. In effect, a traditional simulation treats the parameters as fixed, whereas a Bayesian analysis treats the parameters as random variables. We use a modification

Traditional Simulation:

Step 1. Pick a phylogenetic model, specifying a fixed tree, set of branch lengths, and substitution model parameters.

Step 2. Simulate a data matrix under the model specified in step 1.

Step 3. Perform a Bayesian analysis of the data simulated in step 2, approximating posterior probabilities using MCMC.

Bayesian Simulation:

Step 1. Pick a phylogenetic model by randomly drawing a tree,set of branch lengths, and substitution model parameters from the prior probability distribution.

Step 2. Simulate a data matrix under the model specified in step 1.

Step 3. Perform a Bayesian analysis of the data simulated in step 2, approximating posterior probabilities using MCMC.

FIGURE 1. The traditional simulation procedure fixes parameters of the evolutionary model (such as the tree, branch lengths, and substitution model parameters), simulating many data matrices on the fixed tree. The procedure used in this study first draws the tree, branch lengths, and substitution model parameters from the prior probability distribution before simulating data. This is done for each replicate in the simulation.

of the traditional approach, drawing the phylogenetic parameters such as the tree and branch lengths from a probability distribution and then simulating a single data matrix on every such draw from the prior (Fig. 1). In other words, instead of fixing the parameters of the simulation, we fix the prior probability distribution from which the phylogenetic parameters are drawn.

We simulated data under the JC69 (Jukes and Cantor, 1969) model and under the general time reversible model with gamma distributed rate variation (GTR+Γ; Tavaré, 1986; Yang, 1993, 1994). The JC69 model is very simple, assuming that the rate of substitution is equal across sites, that the four nucleotide frequencies are the same, and that the rates of change among nucleotides are equal. The GTR+Γ model is considerably more complex. The GTR+Γ model of DNA substitution allows the rates of substitution to differ among sites in the sequence (they are random variables drawn from a gamma distribution with shape parameter a), the nucleotide frequencies to be different, and the rate of substitution among nucleotides to be different. We simulated data under the JC69 and GTR+Γ models, but analyzed the data under these models as well as many of the models in between the JC69 and GTR+Γ models. Table 1 shows the models used in this study and the prior probability distribution of the parameters for each.

We simulated data matrices of $c = 100$, $c = 500$, and $c = 1000$ sites on trees of six species. We chose to examine the six species case for a number of reasons. For one, six species is the smallest number of species for which there are several tree shapes for unrooted trees. More importantly, there are only 105 possible unrooted trees for six species, and the posterior probability of each can be reasonably approximated using Markov chain Monte Carlo. If we had decided to examine more species, then the posterior probabilities of individual trees would have been more difficult to accurately estimate. We used MrBayes v3.0 (Huelsenbeck and Ronquist, 2001) to approximate posterior probabilities using Markov chain Monte Carlo. We ran a single Markov chain for 200,000 cycles for each simulated data matrix, discarding samples taken during the first 50,000 cycles. For each set of model conditions, we simulated a total of 10,000 matrices.

Consider just one of the simulations that was performed in this study. In one set of simulations, we simulated data matrices of $c = 100$ sites under the GTR+Γ model of DNA substitution and analyzed the simulated data matrices under the JC69 model. Each of the 10,000 simulated data matrices was generated as follows: First, we picked an unrooted tree from the prior probability distribution of trees. We considered all trees to be equally probable. Hence, each of the 105 trees had a probability of 1/105 of being chosen. Second, once the tree was chosen, we assigned branch lengths to the tree. The branch lengths were randomly assigned by drawing each from an exponential distribution with parameter $\lambda = 5$. The exponential distribution has parameter λ. The mean of the exponential is $1/\lambda$ and the variance is $1/\lambda^2$. Hence, the average branch length on the trees in this study was $1/5 = 0.2$ expected substitutions per site. Third, we chose the nucleotide frequencies from a Dirichlet (5, 5, 5, 5) distribution. Fourth, we chose the gamma-shape parameter for among-site rate variation from an exponential distribution with parameter 2 (resulting in an average shape parameter of $a = 0.5$ over simulations). Fifth, we chose the rates of substitution by generating six independent exponential random variables with parameter 1. These six exponential random variables correspond to the rates r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, and r_{GT}. Sixth, after the parameters of the model had been chosen, 100 sites were simulated on the tree. Finally, a Bayesian analysis was performed on the simulated data matrix by running a Markov chain for 200,000 cycles. The output from the Markov chain Monte Carlo allowed the posterior probabilities of individual trees to be approximated. Of course, because the tree is known, we can ask whether a high posterior probability also corresponded to a high probability that the tree was correct. The entire procedure described here was repeated 10,000 times. Note that no two of the 10,000 simulated data matrices had precisely the same parameter values (i.e., they may have differed in topology but certainly differed in the branch lengths and the substitution model parameters). Each of the figures that summarize the results of the simulations is based on $105 \times 10\,000 = 1\,050\,000$ data points, and these million plus data points were assigned (according to posterior probability) to 20 bins.

TABLE 1. Parameter settings for the models examined in this study. The prior probability distributions for the parameters specify either the probability distribution from which parameters for simulated data sets were drawn, or the settings for the Bayesian analysis of the data. (JC69: Jukes and Cantor, 1969; F81: Felsenstein, 1981; SYM: symmetric model; GTR: general time reversible model, Tavaré, 1986.)

Model	Nucleotide frequencies	Substitution rates	Gamma rate variation	Topology	Branch lengths
JC69	Fixed (1/4, 1/4, 1/4, 1/4)	Fixed (1, 1, 1, 1, 1, 1)	Fixed (∞)	Uniform	Exponential (5)
F81	Dirichlet (5, 5, 5, 5)	Fixed (1, 1, 1, 1, 1, 1)	Fixed (∞)	Uniform	Exponential (5)
SYM	Fixed (1/4, 1/4, 1/4, 1/4)	Dirichlet (1, 1, 1, 1, 1, 1)	Fixed (∞)	Uniform	Exponential (5)
JC69+Γ	Fixed (1/4, 1/4, 1/4, 1/4)	Fixed (1, 1, 1, 1, 1, 1)	Exponential (2)	Uniform	Exponential (5)
SYM+Γ	Fixed (1/4, 1/4, 1/4, 1/4)	Dirichlet (1, 1, 1, 1, 1, 1)	Exponential (2)	Uniform	Exponential (5)
F81+Γ	Dirichlet (5, 5, 5, 5)	Fixed (1, 1, 1, 1, 1, 1)	Exponential (2)	Uniform	Exponential (5)
GTR	Dirichlet (5, 5, 5, 5)	Dirichlet (1, 1, 1, 1, 1, 1)	Fixed (∞)	Uniform	Exponential (5)
GTR+Γ	Dirichlet (5, 5, 5, 5)	Dirichlet (1, 1, 1, 1, 1, 1)	Exponential (2)	Uniform	Exponential (5)

RESULTS AND DISCUSSION

The Meaning of Posterior Probabilities

In the best-case scenario in which all of the assumptions of the Bayesian analysis are satisfied, the posterior probability of a tree is equal to the probability that the tree is correct. Figure 2 shows the relationship between the posterior probability for a phylogenetic tree and the frequency at which trees with that posterior probability are correct (identical to the true tree used in the simulation) for the four simulations that were performed in which all of the assumptions of the Bayesian analysis are satisfied. For large numbers of replicate simulations, the frequency of correct trees will approximate the "frequentist" probability that a tree is correct. Thus, we subsequently refer to the frequency of correct trees as the probability the tree is correct in all figures. The relationship is linear, indicating that posterior probabilities have the meaning ascribed to them: *the posterior probability of a tree is the probability that the tree is correct (assuming that the model is correct)*. Importantly, the Bayesian method is currently the only phylogenetic method that has this property.

Robustness of Posterior Probabilities

Two sets of simulations were performed in which the assumptions of the Bayesian analysis were violated. In the first set, the assumptions of the Bayesian analysis

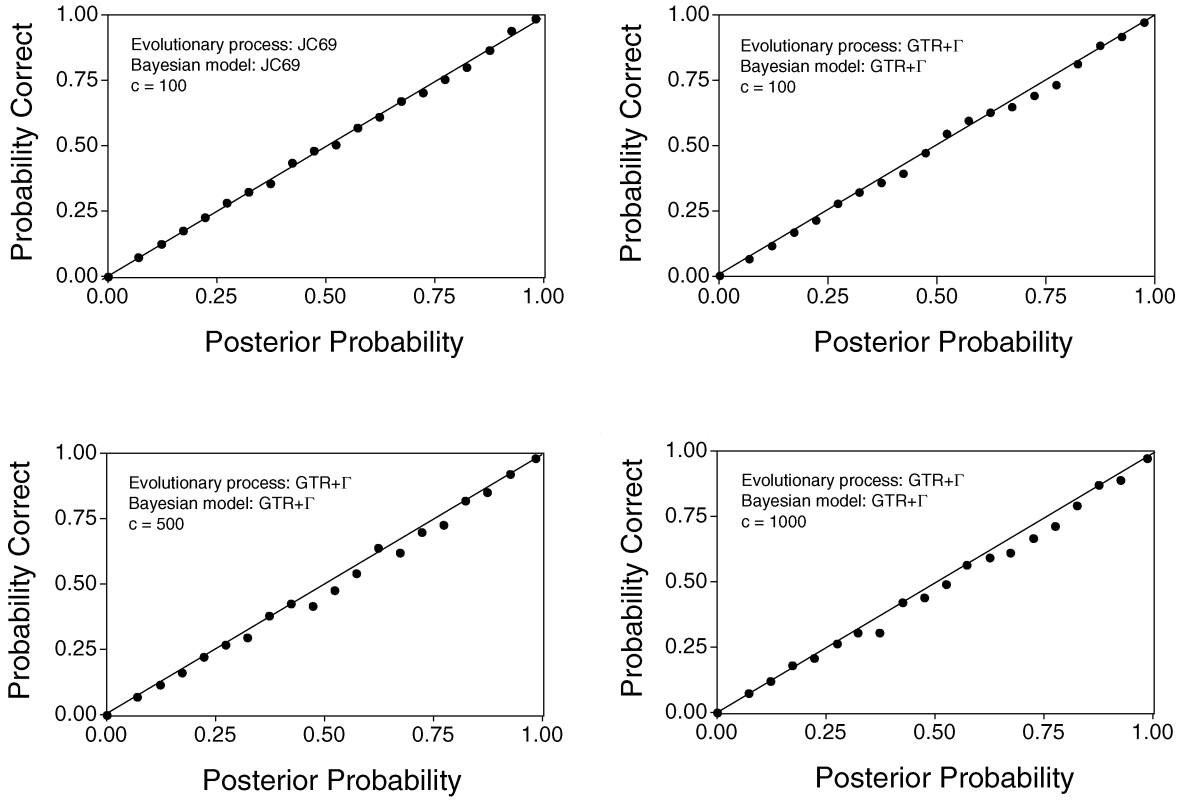

FIGURE 2. The relationship between posterior probabilities on trees and the probability that the tree is correct when the assumptions of the Bayesian analysis are satisfied.

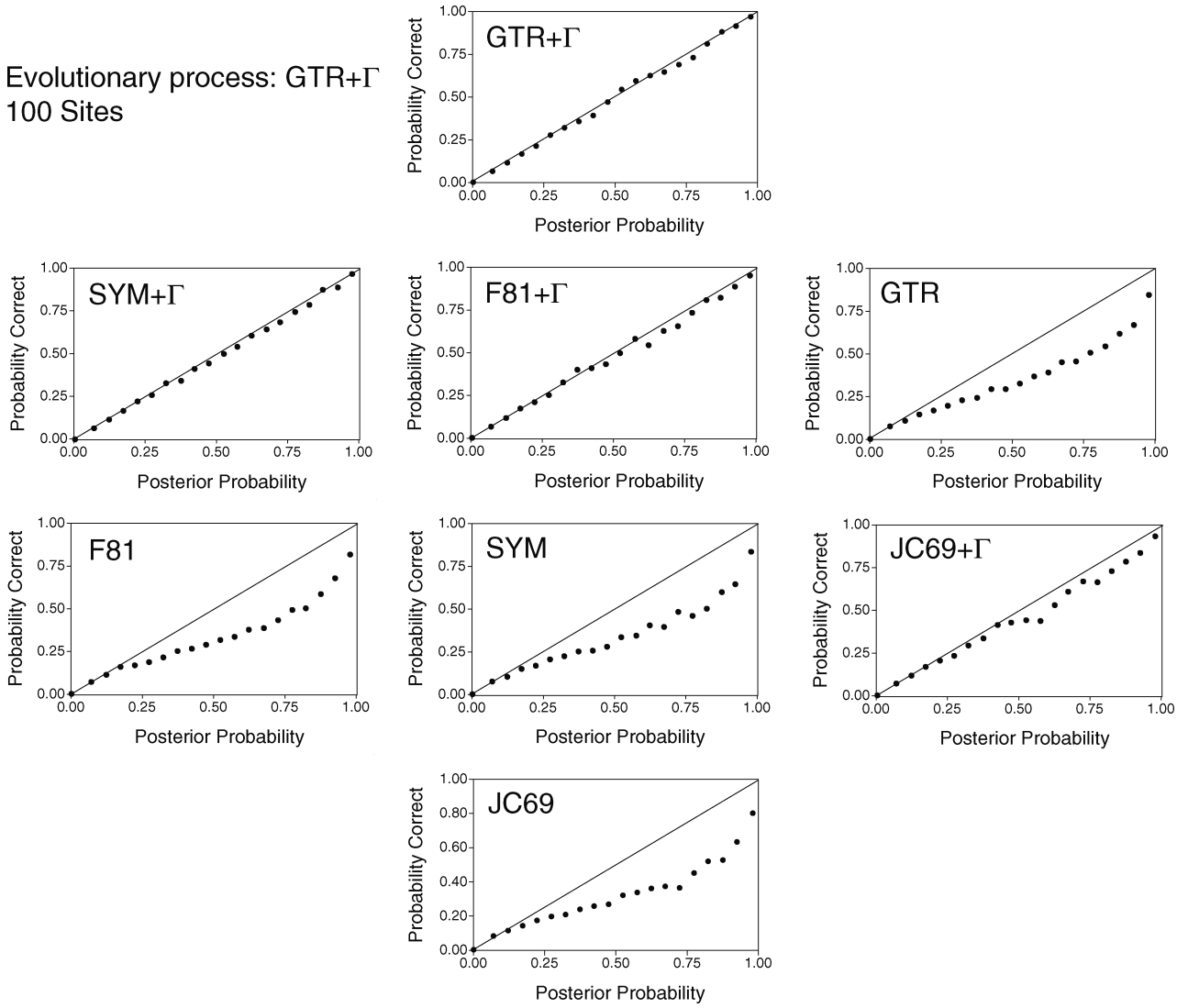

FIGURE 3. The relationship between posterior probabilities on trees and the probability that the tree is correct when the assumptions of the Bayesian analysis are satisfied (top graph) or when the model used in the Bayesian analysis is underspecified (all other graphs).

were under-specified. The evolutionary process generating the DNA sequences followed the GTR+Γ model of DNA substitution. However, the assumptions of the Bayesian analysis did not capture all aspects of the process that generated the DNA sequences. In the simplest model examined—the JC69 model—the Bayesian analysis incorrectly assumed equal nucleotide frequencies, equal rates across sites, and equal rates of substitution among nucleotides. In this case, high posterior probabilities corresponded to smaller probabilities that the tree was correct (Fig. 3). The other models examined had either two components of the true model missing (F81, SYM, and JC69+Γ) or one component of the true model missing (SYM+Γ, F81+Γ, and GTR). For example, the F81 model captures the fact that base frequencies are potentially different, but fails to account for the fact that the actual evolutionary process generating the DNA sequences had rate variation across sites and different rates among the nucleotides. Inspection of the graphs having one or two components of the true model missing suggests that failure to correctly model among-site rate variation had a more serious effect on posterior probabilities than failure to correctly model other aspects of the evolutionary process (Fig. 3); when gamma rate variation was assumed, the posterior probabilities more nearly have their intended interpretation.

In the second set of simulations, summarized in Figure 4, the assumptions of the Bayesian analysis were violated in a different way. In this set of simulations, the evolutionary process followed the JC69 model. However, the Bayesian analysis was overcomplicated, assuming model parameters that were unnecessary. The effect of using an overspecified model was negligible and resulted in a very slight overestimation of the probability that a tree was correct. The bias appears to be in the same direction as the nonparametric bootstrap.

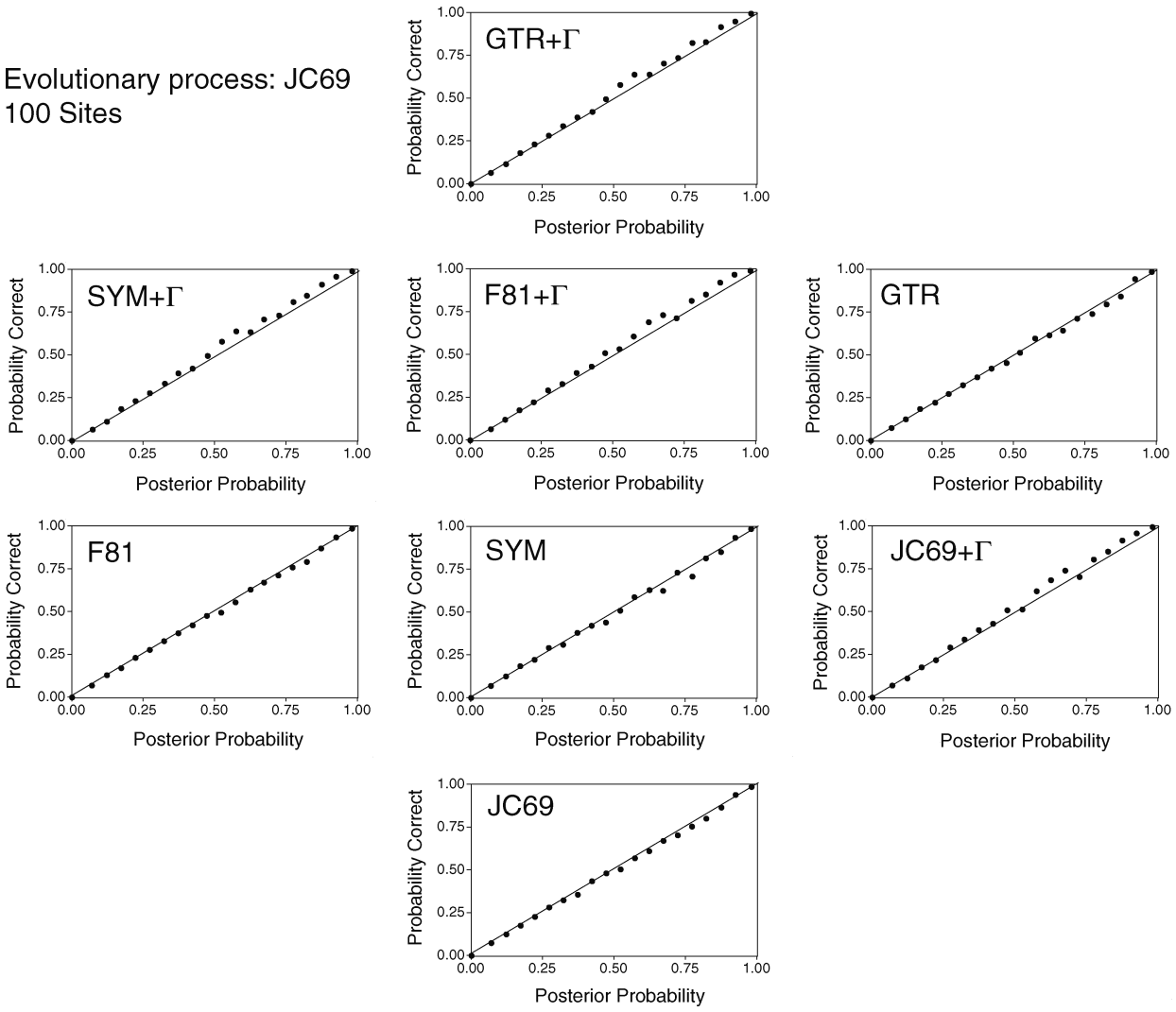

FIGURE 4. The relationship between posterior probabilities on trees and the probability that the tree is correct when the assumptions of the Bayesian analysis are satisfied (bottom graph) or when the model used in the Bayesian analysis is overspecified (all other graphs).

We also collected information on the coverage probability of the 95% credible set of trees and the number of trees contained in the 95% credible set. These results are summarized in Table 2 and reinforce the conclusions drawn from the figures. A 95% credible set of trees is constructed by first ordering all trees by their posterior probability. Trees are included in the credible set starting with the tree of highest posterior probability until the cumulative probability of trees contained in the set is 0.95. It is unlikely that one can obtain a cumulative probability of precisely 0.95. We get around this problem by randomly including or excluding the first tree that exceeds a cumulative probability of 0.95. For example, imagine that the posterior probability of one of the 105 trees is 0.90 and the posterior probability of the tree with the next highest posterior probability is 0.07. Clearly, the first tree should be included in a 95% credible set of trees. However, if the second tree is included in the set, then we do not have a 95% credible set of trees but rather a 97% credible set. If the tree is excluded, then we have a 90% credible set. We resolve this problem by including the second tree in the set with a probability of $0.05/0.07 = 0.714$. Table 2 shows that when the assumptions of the Bayesian analysis are satisfied, that the 95% credible set of trees contains the true tree with a probability that is approximately 0.95. When the assumptions of the analysis are violated, then the 95% credible set of trees contains the true tree with a smaller probability. In the worst case we examined, the 95% credible set of trees contained the true tree with a probability of just 0.71. The number of trees contained in the credible set gives an idea of the variability of the estimate. When the model is underspecified (the true model is more complicated than the model assumed in the Bayesian analysis), the number of trees contained in the credible set is smaller than it should be; the posterior probability is concentrated on too few trees. On the other hand, overspecifying the model has a much smaller effect on the variability of the estimate.

TABLE 2. The coverage probability and average size of the 95% credible set of trees.

Evolutionary process	Bayesian model assumptions	Number of sites	95% credible set coverage probability	95% credible set size
GTR+Γ	JC69	100	0.72	6.05
GTR+Γ	JC69	500	0.71	1.74
GTR+Γ	JC69	1000	0.72	1.28
GTR+Γ	F81	100	0.74	6.47
GTR+Γ	JC69+Γ	100	0.91	16.09
GTR+Γ	SYM	100	0.75	6.37
GTR+Γ	SYM+Γ	100	0.93	16.25
GTR+Γ	GTR	100	0.76	6.64
GTR+Γ	F81+Γ	100	0.94	17.77
GTR+Γ	GTR+Γ	100	0.94	16.69
GTR+Γ	GTR+Γ	500	0.93	4.59
GTR+Γ	GTR+Γ	1000	0.93	2.80
JC69	JC69	100	0.95	4.62
JC69	F81	100	0.95	4.63
JC69	JC69+Γ	100	0.96	5.50
JC69	SYM	100	0.95	4.70
JC69	SYM+Γ	100	0.96	5.64
JC69	GTR	100	0.95	4.72
JC69	F81+Γ	100	0.96	5.53
JC69	GTR+Γ	100	0.96	5.67

For the simulations where the evolutionary process generating the DNA sequences followed the JC69 model, the credible set has about one additional tree in it when the model is very complex (GTR+Γ) as compared to when the model is correct (JC69).

A Comparison of Posterior Probabilities and Nonparametric Bootstrapping

We also attempted a direct comparison of Bayesian posterior probabilities with the nonparametric bootstrap method. Maximum likelihood is the obvious method to compare to Bayesian posterior probabilities because the likelihood function can be calculated under the same assumptions for both. However, a direct comparison was difficult to accomplish. For one, typically implemented maximum likelihood and Bayesian analysis treat nuisance parameters differently. In a maximum likelihood analysis, the likelihood is maximized with respect to the nuisance parameters (such as the branch lengths and substitution model parameters), whereas in a Bayesian analysis they are integrated over a prior probability distribution. Also, we could only examine bootstrap proportions for maximum likelihood under the simplest models. Bootstrap analysis under the more parameter rich models takes too long to complete. Hence, we only examined the bootstrap method implemented with maximum likelihood under the Jukes-Cantor model (JC69; Jukes and Cantor, 1969).

Figure 5 shows a comparison of Bayesian posterior probabilities and maximum likelihood bootstrapping when the assumptions of the analysis are violated (the two graphs on the left) and when the assumptions of the analysis are satisfied (the two graphs on the right). As expected, the bootstrap method is too conservative when its assumptions are satisfied. Bootstrap proportions generally corresponded to higher probabilities of the tree being correct when the evolutionary process followed the Jukes-Cantor model and the maximum likelihood analysis also assumed the Jukes-Cantor model. This finding is consistent with previous simulation studies (Hillis and Bull, 1993; Alfaro et al., 2003). As noted earlier, there is a linear relationship between the posterior probability of a tree and the probability that the tree is correct.

The nonparametric bootstrap is not as biased as the Bayesian posterior probabilities when the assumptions of the analysis are violated. When the evolutionary process followed the GTR+Γ model of DNA substitution, but the analysis assumed the Jukes-Cantor model, the bootstrap values more nearly had a linear relationship with the probability that the tree is correct.

RECOMMENDATIONS

Bayesian inference is the only method available (with the possible exception of the corrected bootstrap method), that provides estimates of phylogeny and an indication of phylogenetic uncertainty that are both correct and easily interpretable. The posterior probability of a phylogenetic tree is the probability that the tree is correct, assuming that the model is correct. On the other hand, all bets are off when the assumptions of a Bayesian analysis are not satisfied, the same conclusion reached by Waddell et al. (2002). In the simulations performed in this study, the Bayesian method was more sensitive to underspecification of the evolutionary model than to overspecification. More specifically, in the simulations performed here, failure to account for among-site rate variation more severely affected posterior probabilities than failure to properly model other aspects of the evolutionary process. Sullivan and Swofford (2001) showed a similar sensitivity of maximum likelihood when among site rate variation is not accounted for..

It may be too much to hope that there exists a method that (1) provides a direct measure of phylogenetic reliability while (2) also being robust to violation of model assumptions. Clearly, the nonparametric bootstrap is not such a method. It is already known that it is invalid to use the bootstrap to measure the probability that a clade is correct, and one of the simulations performed in this study further confirms this. The corrected bootstrap method may more nearly have the correct coverage probability, but it is currently unknown how sensitive the corrected bootstrap is to model misspecification. A simulation study that examined the corrected bootstrap could address the sensitivity of the method to model misspecification (and would represent an impressive amount of computation because the method is so difficult to implement). The Bayesian method, too, does not fulfill the properties of providing a correct measure of phylogenetic reliability while at the same time being robust to model misspecification. When its assumptions are satisfied, the Bayesian method correctly measures phylogenetic reliability. Unfortunately, it appears to be more sensitive than the bootstrap method to model violation.

This simulation study was the first that correctly evaluated the coverage probability of Bayesian posterior

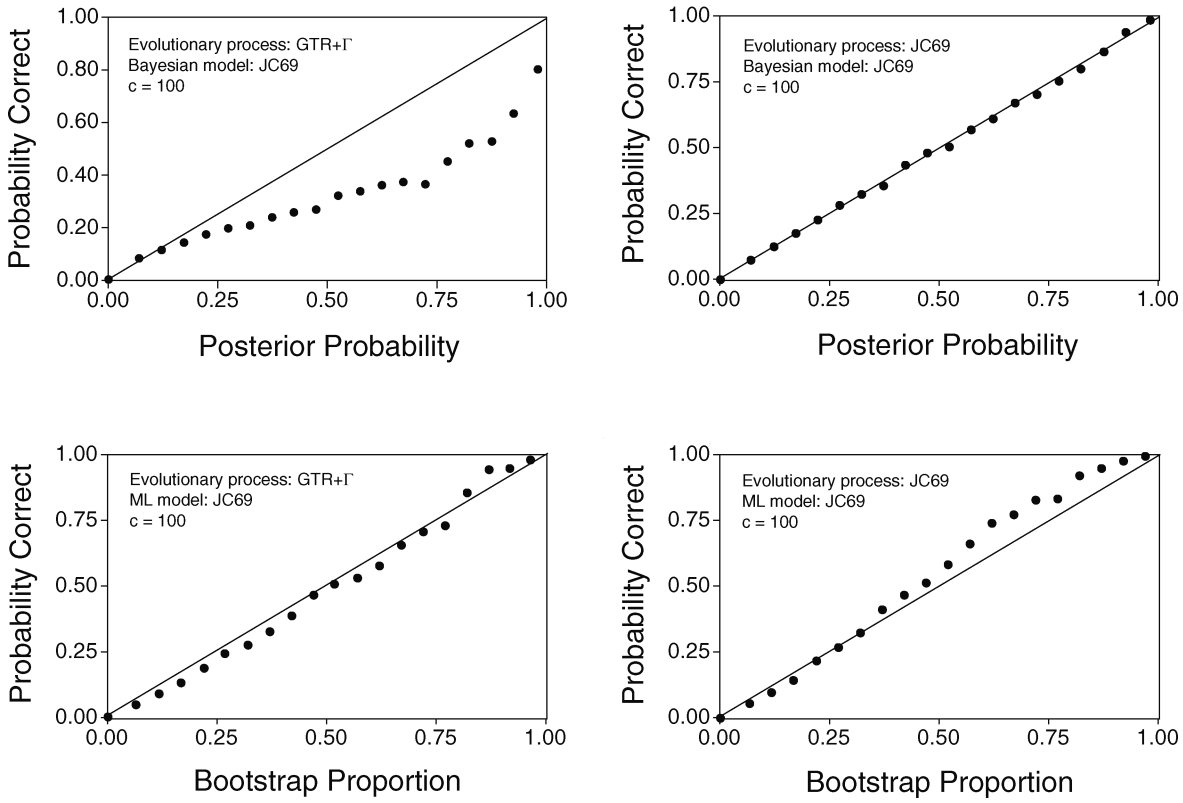

FIGURE 5. A comparison of posterior probabilities and nonparametric bootstrap proportions. The top two graphs show the relationship between posterior probabilities and the probability that the tree is correct. The bottom two graphs show the relationship between bootstrap values and the probability that the tree is correct.

probabilities. The basic problem in earlier studies is that they did not treat the model parameters as random variables. The Bayesian method treats all parameters of a model as random variables, with a prior probability distribution on each. Earlier studies had fixed the tree (and other aspects of the phylogenetic model), and therefore failed to satisfy the assumptions of the Bayesian method. Here, we simulated data sets on trees that were first drawn from a prior probability distribution. In this respect, our simulation satisfied the assumptions of the Bayesian method.

The results of this study suggest that careful attention must be paid to the model used in a Bayesian analysis. Specifically, the model should be as complex as possible while still allowing parameters to be identified. There are a number of strategies that can be currently used, such as partitioning data and modelling the evolutionary process separately in each. This can be done using MrBayes v3.0 (Huelsenbeck and Ronquist, 2001). The idea is to increase the number of trees visited by the Markov chain Monte Carlo method, and keep the Bayesian method from placing too much probability on too few trees (Castoe et al., 2004; Nylander et al., 2004; Lin et al., 2004). Unfortunately, there are limits to this approach. The universe of phylogenetic models is currently quite small and the types of evolutionary processes that are accommodated is limited. One can apply the current models to small parts of a data matrix, but if the model fails to capture important evolutionary processes, then it is not clear how much improvement there will be in the estimate of phylogeny or the assessment of variability in the phylogeny in the Bayesian method. What is also needed is an expansion of the universe of possible models. Specifically, virtually all of the models currently used assume that the evolutionary process is homogenous over the entire phylogenetic history of a group. This assumption can be relaxed. For example, the covarion-like model (Tuffley and Steel, 1997) relaxes the assumption that the rate of substitution at a site is constant over time. It might also be possible to relax the assumption that nucleotide composition is constant over time. For example, in a Bayesian framework, one might assume that nucleotide composition changes discretely, and use Markov chain Monte Carlo to integrate over different histories of nucleotide composition change (e.g., in a manner similar to Huelsenbeck et al., 2000). Adoption of this strategy—using more complex models chosen from a more extensive pool of candidate phylogenetic models—does not necessarily mean abandoning formal model choice. It may still be possible to choose a model that best explains the alignment without introducing superfluous parameters using Bayesian model choice (e.g., Huelsenbeck et al., 2004) or information criteria (see Burnham and Anderson, 1998).

ACKNOWLEDGMENTS

J.P.H. was supported by NSF grants DEB-0075406 and MCB-0075404, NIH grant GM-069801, and by a Guggenheim Fellowship. B.R. was supported by NIH grant HG01988 and CIHR grant MOP44064.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 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.

Bremer, K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803.

Bremer, K. 1994. Branch support and tree stability. Cladistics 10:295–304.

Burnham, K. P., and D. R. Anderson. 1998. *Model selection and inference*. Springer-Verlag, New York.

Castoe, T. A., T. M. Doan, and C. L. Parkinson. 2004.Data partitions and complex models in Bayesian analysis: The phylogeny of Gymnophthalmid lizards. Syst. Biol. 53:448–470.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

DeBry, R. 2001. Improving interpretation of the decay index for DNA sequences. Syst. Biol. 50:742–752.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

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

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

Efron, B., and R. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, London.

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

Faith, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40:366–375.

Farris, J. S., V. A. Albert, M. Kallersjo, D. Lipscomb, and A. G. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2003. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

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

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:725–736.

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

Holmes, S. 2003. Bootstrapping phylogenetic trees: Theory and methods. Stat. Sci. 2:241–255.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–265.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.

Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

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

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

Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265–277.

Lin, C. P., B. N. Danforth, and T. K. Wood. 2004. Molecular phylogenetics and evolution of maternal care in membracine treehoppers. Syst. Biol. 53:400–422.

Mueller, L. D., and F. J. Ayala. 1982. Estimation and interpretation of genetic distance in empirical studies. Genet. Res. Camb. 40:127–137.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. Douady, E. C. Teeling, O. A. Ryder, M. Stanhope, W. W. De Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–68.

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

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

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution patternare violated? Syst. Biol. 50:723–729.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:15138–16143.

Swofford, D. L., J. L. Thorne, J. Felsenstein, and B. M. Wiegmann. 1996. The topology-dependent permutation test for monophyly does not test for monophyly. Syst. Biol. 45:575–579.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pages 57–86 *in* Lectures in mathematics in the life sciences, vol. 17. American Mathematical Society.

Tuffley, C., and M. Steel. 1997. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.

Waddell, P. J., H. Kishino, and R. Ota. 2002. Very fast algorithms for evaluating the stabilityof ML and Bayesian phylogenetic trees from sequence data. Genome Informatics 13:82–92.

Wilcox, T. P., D. J. Zwickl, T. A. Heath, and D. M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139:993–1005.

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

First submitted 22 July 2003; reviews returned 18 November 2003; final acceptance 18 July 2004
Associate Editor: Thomas Buckley