*Syst. Biol.* 52(5):665–673, 2003

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390235485

# Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics

PER ERIXON,[1] BODIL SVENNBLAD,[2] TOM BRITTON,[2] AND BENGT OXELMAN[1]

[1]*Department of Systematic Botany, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-75236 Uppsala, Sweden; E-mail: per.erixon@ebc.uu.se (P.E.)*
[2]*Department of Mathematics, Uppsala University, Box 480, SE-75106 Uppsala, Sweden*

*Abstract*.—Many empirical studies have revealed considerable differences between nonparametric bootstrapping and Bayesian posterior probabilities in terms of the support values for branches, despite claimed predictions about their approximate equivalence. We investigated this problem by simulating data, which were then analyzed by maximum likelihood bootstrapping and Bayesian phylogenetic analysis using identical models and reoptimization of parameter values. We show that Bayesian posterior probabilities are significantly higher than corresponding nonparametric bootstrap frequencies for true clades, but also that erroneous conclusions will be made more often. These errors are strongly accentuated when the models used for analyses are underparameterized. When data are analyzed under the correct model, nonparametric bootstrapping is conservative. Bayesian posterior probabilities are also conservative in this respect, but less so. [Bayesian inference; bootstrap; model misspecification; phylogeny; posterior probability; simulations; support; true sampling.]



During the past 10–15 years, the focus of phylogenetic research has shifted from finding the best topology by maximizing an optimality criterion to quantification of the support for certain clades. Bootstrap values (Felsenstein, 1985), Bremer support values (Bremer, 1988, 1994), jackknife values (Farris et al., 1996), and recently Bayesian posterior probabilities (Rannala and Yang, 1996; Yang and Rannala, 1997; Huelsenbeck et al., 2001) are commonly used measures. The interpretation of support could be nonprobabilistic, as for Bremer support and some interpretations of bootstrapping and jackknifing (Farris et al., 1996; Oxelman et al., 1999). A probabilistic measure of support has been considered preferable by some authors (e.g., Felsenstein, 1985; Sanderson, 1989) because it quantifies how likely a certain clade is to be correct, given the data at hand and the assumptions made. For the rest of this article, we will restrict the discussion to probabilistic interpretations of bootstrapping and Bayesian inference.

Generally speaking, phylogenetic methods should be efficient, powerful, consistent, robust, and falsifiable (Penny et al., 1992). In terms of support, this means that methods that need a minimum number of data to attain high support values for true clades are preferable. Nonparametric bootstrapping is considered conservative by some authors (Zharkikh and Li, 1992; Hillis and Bull, 1993), but Efron et al. (1996) claimed that bootstrap frequencies are not biased systematically downward but represent unbiased probability estimates in a Bayesian sense.

In a Bayesian analysis, inferences of phylogeny are based upon the posterior probabilities of phylogenetic trees. The posterior probability, following Bayes's theorem, is a function of a prior probability and the likelihood of the data under some assumed model. The equation is not feasible to solve analytically, and therefore it has until recently not been much used. With the introduction of Markov chain Monte Carlo (MCMC) algorithms to estimate posterior probabilities of phylogenetic tree topologies (Yang and Rannala, 1997), support analyses can be performed for data sets with hundreds of taxa (Huelsenbeck et al., 2001), with explicit models of evolution specified.

One of the distinct advantages with Bayesian inference in phylogenetics is claimed to be that the posterior probabilities have a clear-cut interpretation, i.e., they represent the probability that the corresponding clade is true conditional on the model, the priors, and the data (Huelsenbeck et al., 2002). Following this line of reasoning, nonparametric bootstrapping does not have such a clear-cut interpretation in terms of phylogenetic reconstruction (e.g., Larget and Simon, 1999). One interpretation of nonparametric bootstrapping that validates a comparison with Bayesian inference was proposed by Efron et al. (1996:7090): "In a Bayesian sense, the $\alpha$ [bootstrap confidence level] can be thought of as reasonable assessments of error for the estimated tree." Durbin et al. (1998:212) claimed that "the bootstrap confidence for a feature approximates the posterior probability of that feature, assuming a flat prior over trees." This would suggest that nonparametric bootstrapping and MCMC-generated posterior probabilities would be similar, provided that the prior is negligible or, as Huelsenbeck et al. (2001:2311) put it "this [Bayesian inference] is roughly equivalent to performing a maximum likelihood analysis with bootstrap resampling, but much faster." At first glance, this prediction seems to be reasonable because of the standard practice of using a flat prior distribution and because the influence of the prior distribution decreases as the amount of data (i.e., basically sequence length) increases.

Empirically, Bayesian posterior probabilities for clades in phylogenetic trees are usually found to be considerably higher than corresponding nonparametric bootstrap frequencies (e.g., Karol et al., 2001; Murphy et al., 2001; Leaché and Reeder, 2002; Whittingham et al., 2002). Objective comparisons are difficult because few studies have applied exactly the same analysis model for bootstrapping and Bayesian analysis. A common practice is to fix the parameter values for the pseudoreplicates in bootstrapping. It is not clear how this practice affects the results.

In this study, we used a simulation approach to test the hypothesis that there are no differences between nonparametric bootstrap frequencies and posterior probabilities. If there are differences, we want to explore the relative amounts of type I and type II errors of the two methods. A method that consistently gives higher probabilities may be more powerful, in the sense that fewer data are needed to reach the correct conclusion (i.e., the method is less prone to type II error). On the other hand, it is also possible that erroneous conclusions will be made more often (i.e., the method has higher type I error rate). We also briefly explored the effect of model misspecification on the results and investigated the effect of sequence length on the difference between the two methods.

## METHODS

An unrooted five-taxon tree, ((((A:0.07, B:0.18):0.02, C:0.16):0.02, D:0.12), E:0.38), was used for all simulation of data except for the investigation of the effect of sequence length (see below). The numbers in the tree notation represent expected number of substitutions per site. These numbers were specified to give some rate heterogeneity over the tree and relatively short internal branches, resulting in clades of medium to high support for a sequence length of 1,000 bases. This, we believe, is a reasonably realistic shape of a tree. All data sets were generated using the software Seq-Gen 1.2.5 (Rambaut and Grassly, 1997). DNA sequences were generated under the Jukes–Cantor model (JC69), where all base frequencies and substitution rates are equal, and under a more parameter-rich model: the general time reversible model with rate variation among sites assumed to follow a discrete gamma distribution (i.e., GTR+$\Gamma$) with a shape parameter of 0.5 and four rate categories. The base frequencies were set to give AT-rich sequences (frequency of A = T = 0.3, C = G = 0.2) to have some deviation from equal frequencies used in the JC69 model. The rate matrix values were chosen arbitrarily by simulating one data set with a transition:transversion ratio of 2.0 and then estimating the rate matrix by analyzing that data set with maximum likelihood under the GTR+$\Gamma$ model, thereby making transitions more frequent than transversions ($r_{AC} = 0.7156$, $r_{AG} = 3.7054$, $r_{AT} = 0.8245$, $r_{CG} = 1.0219$, $r_{CT} = 3.1090$, $r_{GT} = 1.0$). Initially, 50 data sets (the $5_{50}$ data sets) were generated for the JC69 model and 50 were generated for the GTR+$\Gamma$ model. For each data set, statistical inference was performed using both nonparametric bootstrapping with maximum likelihood (MLBOOT) as well as analyses within a Bayesian framework using MCMC algorithms (BAYES) in order to identify differences between the two methods. The JC69 data sets were also analyzed with the GTR+$\Gamma$ model and vice versa, i.e., with an incorrect model, to compare BAYES and MLBOOT under model misspecification. The expectation is that analysis with incorrect model would give lower support for both methods. There are nine more parameters in the GTR+$\Gamma$ model than in the JC69 model, and this case of model misspecification is severe in the context of models commonly available in software used for phylogenetic reconstruction. For both BAYES and each pseudoreplicate in MLBOOT, all model parameter values were estimated from the data except the shape parameter in the GTR+$\Gamma$ model, which was set to 0.5.

MLBOOT was implemented by using PAUP* 4.0b10 (Swofford, 2002) with heuristic search (tree bisection–reconnection branch swapping, random addition sequence with one replicate, multrees not in effect, and LCollapse = no) and 300 pseudoreplicates for each data set. The model parameters were reestimated in every pseudoreplicate. This is important in order to make a fair comparison with Bayesian inference. Some of the pseudoreplicates will give ML trees with one of the internal branches of length zero. When the default settings in PAUP* are used, these trees will be collapsed and they will not contribute to the bootstrap frequencies, which in turn will not sum to 1. This process would result in a systematic difference between the bootstrap frequencies and MCMC-generated posteriors because the latter will always sum to 1 with the software we used. By setting LCollapse = no, this pitfall is avoided.

MrBayes 2.01 (Huelsenbeck and Ronquist, 2001) was used to estimate posterior probabilities. Model parameters were assigned uniform priors. The default priors for the branch lengths were random values between 0 and 10 substitutions per site in the sequence. The analyses were initiated with random starting trees and were run with four separate Markov chains for 500,000 generations, with a tree saved each 100th generation. The first 500 trees were discarded as a conservative generalization of the "burn-in" phase. Usually, the chains reach stationarity with respect to likelihood score after only a few hundred generations. The 4,500 remaining trees were used to calculate the posterior probabilities (BAYES) for the two internal branches of the prespecified tree topology for each data set.

Statistical evaluation of differences between the two sets of support values from BAYES and MLBOOT were performed with the nonparametric Wilcoxon signed-rank test. There is a stochastic error in the MLBOOT values because of the limited number of replicates. The confidence interval, based on a binomial distribution, is largest for support values of 50%. In the case of 300 pseudoreplicates, the 95% confidence interval is $\pm 5.6\%$ units (Hedges, 1992). Because this error is random, we expect the mean of 50 such errors (in the 50 data sets) to be small enough to be negligible for the conclusions we want to draw.

The support for the two internal branches was also calculated by using true sampling (i.e., a parametric bootstrap). For both models, 10,000 independent data sets were simulated on the five-taxon tree. The ML tree was calculated for each of these data sets. The frequencies (MLtrue) for the two internal branches were obtained by counting their occurrence in the ML trees of the 10,000 data sets.

To identify the type I error of BAYES and MLBOOT, an additional 2,000 data sets (the $5_{2000}$ data sets) with 1,000 characters each were simulated under the GTR+$\Gamma$ model on the five-taxon tree. The large number of data

sets is needed because only a small fraction of the high support values are for wrong clades, e.g., 5% of support values >95% are expected to be wrong. These data sets were analyzed under the GTR+Γ model in the same way as described above but using ML bootstrapping with 100 pseudoreplicates and Bayesian inference with the MCMC running for 60,000 generations giving 600 trees of which the first 100 were discarded. To check whether the MCMC chains reached stationarity, the first 1,000 of the $5_{2000}$ data sets were analyzed twice with different random starting priors. Results were strongly correlated ($R^2 = 0.993$), implicating stationarity of the topology parameter. To explore how the type I error of two methods is affected by underparameterization, these data were also analyzed with the JC69 model. The aim here is not primarily to compare BAYES under a correct model versus an incorrect model or MLBOOT under a correct model versus an incorrect model but rather to investigate the relative behavior of BAYES and MLBOOT when underparameterized compared with analysis the correct model, i.e., to investigate if one of the methods is more affected by underparameterization.

The results from the analyses with the correct model of the $5_{2000}$ data sets were also used to make a logistic regression where we modeled the probability $\pi$ that the true clade has been found as a function of the support value, $x$. Let $Y$ be a binary response variable indicating whether the clade is true or not ($Y = 1$ if the clade is in the true tree and 0 otherwise), then $P(Y_i = 1) = \pi(x_i)$. We applied, assuming the observations $(x_i, y_i)$ to be independent, a logit model where $\alpha$ and $\beta$ are estimated from data. Because $\pi(x_i) = x_i$ (meaning that the support value, as intended, equals the probability that the clade is in the true tree) is not part of this model, we also applied the logit model with transformed support values: $\log\{\pi(x_i)/[1 - \pi(x_i)]\} = \alpha + \beta \log[x_i/(1 - x_i)]$. The corresponding probit models were also applied to data (Agresti, 1990).

There is a conceptual difference between the methods in that Bayesian inference aims at estimating the frequencies of different topologies in the posterior distribution, whereas ML seeks the single topology with the highest likelihood. Theoretically, a non-ML topology can have a higher posterior probability than the ML topology if the ML topology has very specific parameter values (i.e., the ML topology represents a narrow peak in parameter space) but would have only slightly higher likelihood than another topology with a much broader range of parameter values. We applied a standard $\chi^2$-test to the results of the analyses of the $5_{2000}$ data sets under the correct model to check this hypothesis. The test variable was the frequency with which BAYES and MLBOOT found the correct topology.

We chose to use a single tree for simulations because we were primarily interested in testing the idea that nonparametric bootstrap frequencies and Bayesian posterior probabilities are generally roughly interchangeable (Efron et al., 1996; Durbin et al., 1998; Huelsenbeck et al., 2001). Because the methods can be assumed to be consistent under the correct model, we did not expect that the shape of the tree would affect the generality of the results whenever the topology and branch length parameters were within a realistic range (but see Cummings et al., 2003, for a more thorough exploration of parameter space). Our null hypothesis is that when BAYES and MLBOOT are performed using the same inconsistent model, they are based on the same likelihood function and should be inconsistent in the same way.

To investigate the effect of sequence length, a large number of four-taxon data sets were simulated and analyzed under the JC69 model and two different setups. First, 500 data sets were generated for eight different sequence lengths (1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 8,000, and 10,000 bases) using a four-taxon tree with one short internal branch: (V:0.04, X:0.15):0.01, Y:0.04),Z:0.15). Second, to minimize the problem of support values reaching 100% when sequence length increases, the two long terminal branches were made 0.15 expected substitutions longer when sequence length was doubled, starting at 1,000 bases using the function $Y = \alpha + \beta \ln(X)$, where Y is branch length and X is sequence length. The two constants $\alpha$ and $\beta$ were set to −1.3448676 and 0.2164043, respectively, to give the desired result of keeping the expected support (as determined by parametric bootstrapping) approximately constant in the sequence length interval 1,000–64,000 bases. This process was used for the same eight sequence lengths used above plus three additional longer sequences (16,000, 32,000, and 64,000 bases) and 500 data sets for each sequence length. All data sets were analyzed with ML bootstrapping with 200 pseudoreplicates and Bayesian inference with the MCMC running for 60,000 generations giving 600 trees, of which the first 100 were discarded.

## RESULTS

Bayesian inference yielded significantly higher support values than ML bootstrap values, with mean differences of 4.1 and 3.3 percent units, respectively, for the two internal branches under the GTR+Γ model and mean differences of 3.9 and 2.7 percent units under the JC69 model for the $5_{50}$ data sets (Table 1). The support values were substantially higher under the JC69 model. The "true sampling" frequencies, using ML, were higher than both MLBOOT and BAYES, except in one case (JC69, BAYES clade AB). The BAYES value in that case was not significantly different from the MLtrue value (unpaired comparison), because the MLtrue value (98.9) is well within the 95% confidence interval for the BAYES value (98.6–99.8).

Mean support values for true clades were generally lower when analyzing the JC69 $5_{50}$ data sets with the GTR+Γ model and vice versa, i.e., with an incorrect model. However, in the case of underparameterization BAYES gave slightly higher mean support compared with analysis under correct model (clades AB and ABC combined: 75.5 vs. 73.6, $P = 0.044$). A more detailed comparison of the support value distributions from BAYES and MLBOOT was therefore performed on the $5_{2000}$ data

TABLE 1. Comparison of support values from phylogenetic analysis with Bayesian inference (BAYES) and ML bootstrapping (MLBOOT). Fifty data sets were generated and analyzed under two evolutionary models (GTR+Γ and JC69) using a five-taxon tree. Probabilities ($P$) that BAYES and MLBOOT are incorrectly rejected as being equal were calculated with the nonparametric Wilcoxon signed-rank test. Frequencies for true sampling (MLtrue) based on 10,000 data sets are presented together with mean values for BAYES and MLBOOT.

| Model | Clade | Method | Mean (%) | Median (%) | $P$ |
|---|---|---|---|---|---|
| GTR+Γ | AB | MLtrue | 89.8 | | |
| | | BAYES | 80.3 | 88.0 | |
| | | MLBOOT | 76.2 | 82.5 | <0.0001 |
| | ABC | MLtrue | 76.5 | | |
| | | BAYES | 66.9 | 76.4 | |
| | | MLBOOT | 63.6 | 66.8 | 0.0025 |
| JC69 | AB | MLtrue | 98.9 | | |
| | | BAYES | 99.2 | 100 | |
| | | MLBOOT | 95.3 | 98.6 | <0.0001 |
| | ABC | MLtrue | 93.2 | | |
| | | BAYES | 79.9 | 97.3 | |
| | | MLBOOT | 77.2 | 89.8 | 0.0002 |

sets, analyzed both with the correct model and with underparameterization. There was a substantial difference between the shape of the support value distributions for BAYES and MLBOOT (Fig. 1). The analysis with the correct model showed that low support values (0–25%) for true clades were proportionally more frequent for BAYES, than for MLBOOT, but all support categories in the range of 35–95% were more frequent for MLBOOT. Support values >95% were more than twice as frequent for BAYES compared with MLBOOT. When underparameterized, this pattern was accentuated, giving BAYES a strongly bimodal distribution (Fig. 1D). There was no significant difference in how often BAYES and MLBOOT retrieved the true tree topology (support disregarded) among the $5_{2000}$ data sets analyzed with the correct model (68.7% vs. 68.1%, $P = 0.35$).

The analysis of the $5_{2000}$ data sets showed that the risk of making an erroneous conclusion (type I error) is higher with BAYES than with MLBOOT. Of support

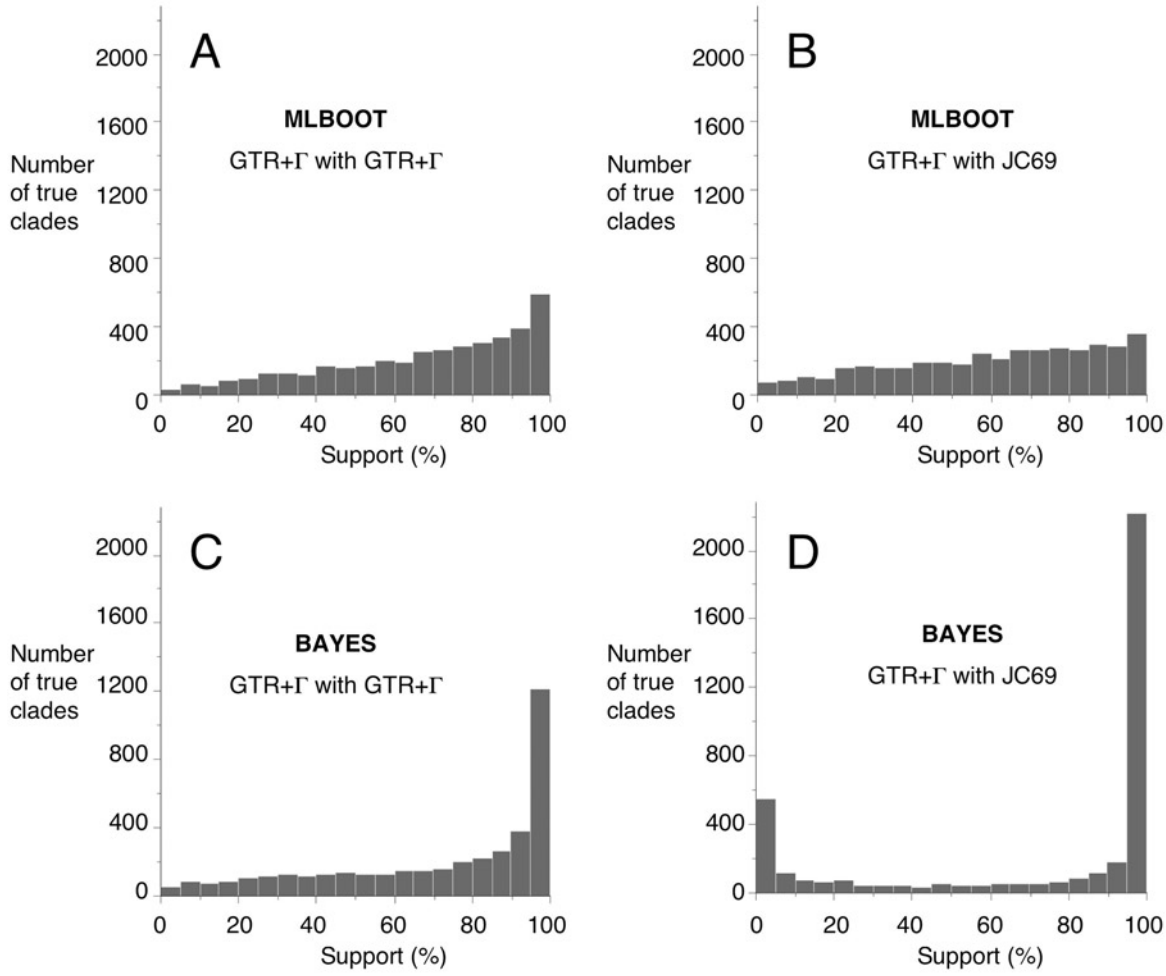


FIGURE 1. Histograms of support value distributions for true clades in the $5_{2000}$ data sets (simulated under the GTR+Γ model). The $y$-axis represents number of true clades in each support category. Total number of support values for true clades is 4,000. (a) MLBOOT analyzed with GTR+Γ, correct model; (b) MLBOOT analyzed with JC69, underparameterized model; (c) BAYES analyzed with GTR+Γ, correct model; (d) BAYES analyzed with JC69, underparameterized model.

TABLE 2. Error rates of BAYES and MLBOOT based on 4,000 clades from 2,000 data sets generated under the GTR+Γ model and analyzed with the GTR+Γ and JC69 models, respectively. The support values were put in four categories. The total number of clades (N), in each category and the proportion of these that were wrong (i.e., not AB or ABC) are shown.

| | BAYES | | | | MLBOOT | | | |
|---|---|---|---|---|---|---|---|---|
| | Correct model | | Incorrect model | | Correct model | | Incorrect model | |
| Support category (%) | N | % wrong | N | % wrong | N | % wrong | N | % wrong |
| 70.1–80.0 | 418 | 17.3 | 262 | 54.6 | 612 | 10.8 | 634 | 15.3 |
| 80.1–90.0 | 542 | 10.2 | 337 | 37.4 | 713 | 4.5 | 622 | 8.5 |
| 90.1–95.0 | 409 | 6.2 | 292 | 40.4 | 399 | 2.0 | 301 | 4.7 |
| 95.1–100.0 | 1,216 | 1.7 | 2,622 | 15.6 | 505 | 0.2 | 293 | 2.0 |

values >95%, only 0.2% were for wrong groups with MLBOOT compared with 1.7% with BAYES (Table 2) when GTR+Γ-generated data were analyzed with the correct model. When data were analyzed with an underparameterized model (GTR+Γ data analyzed with the JC69 model), the type I error rate increased for both BAYES and MLBOOT, but much less for MLBOOT. The type I error with BAYES increased from 1.7% to 15.6% for support values >95% under this kind of model misspecification. The number of clades with support values >95% increases more than twofold (Table 2), indicating that the absolute number of high support values for true groups also increases (see also Fig. 1).

Because the residual sum of squares for the logit models for transformed support values was smaller than that for any of the probit models (both for BAYES and for MLBOOT), we chose the logit model. Given the support value $x_i$, $\pi(x_i)$ is the probability that the estimated clades are the true clades. In this logit model, the relationship between $x_i$ and $\pi(x_i)$ is modeled by $\log[\pi(x_i)/(1-\pi(x_i)] = \alpha + \beta \log[x_i/(1-x_i)]$. The parameters $\alpha$ and $\beta$ are estimated from the data. In Figure 2, the estimated functions of the logit model for our simulations are shown. The simplest relationship between $x_i$ and $\pi(x_i)$ is $\pi(x_i) = x_i$, with the natural interpretation that the support value equals the probability of having the true clade. This is a special case of the model, with $\alpha = 0$ and $\beta = 1$, but this hypothesis is rejected, in this particular study ($P < 0.0001$, for both BAYES and MLBOOT, using a likelihood ratio test). This is also seen in the figure. For example, a probability of 95% corresponds to a BAYES value of 91% and an MLBOOT value of 84%.

In the four-taxon case with varying sequence length under the same simulation tree, support values quickly approach 100% for both methods, but BAYES needs shorter sequence length than does MLBOOT to reach a certain support level (Fig. 3a). BAYES also needs a smaller relative increase in sequence length than MLBOOT to increase the support value by the same amount (e.g., from 95% to 99%). The ratio of nonsupport of the two methods, i.e., (1-MLBOOT)/(1-BAYES), is not only always larger than 1 (indicating that BAYES supports are larger), but in fact the ratio increases with sequence length, indicating that the BAYES support has a higher rate of convergence to complete support (Fig. 3b).

When the simulation tree is modified to keep support values roughly constant (Fig. 4), the mean support value of MLBOOT initially approaches that of BAYES with increased sequence length. For 3,000 bases and more, the difference seems to stabilize at a level significantly different from zero, even for very long sequences (Fig. 5). The 95% confidence interval of the difference for 64,000 bases is large, indicating that problems with substitution saturation (the long terminal branches have 1.05 expected substitutions per site) are becoming apparent.

## DISCUSSION

Our results reject the idea that nonparametric bootstrap frequencies for ML estimates and Bayesian posterior probabilities for clades in phylogenetic trees are universally equivalent. The results also show that Bayesian posterior probabilities, on average, are substantially higher than corresponding bootstrap frequencies. These findings together with the observations generally made based on empirical data (e.g., Karol et al., 2001; Murphy et al., 2001; Leaché and Reeder, 2002; Whittingham et al., 2002) strongly indicate that there is, contrary to theoretical claims (Efron et al., 1996; Durbin et al., 1998), a systematic difference between nonparametric bootstrap frequencies and Bayesian posterior probabilities.

The explanation for the observed difference between MLBOOT and BAYES is unknown. It could be that Bayesian posterior probabilities and nonparametric bootstrap frequencies for ML estimates are approximately equivalent in other applications but that this is not true for the complexity of phylogenetic reconstruction. Durbin et al. (1998) pointed out that there must be enough data if bootstrapping is to be a good approximation of the posterior probability. With an infinite amount of data both methods will give absolute support (100%) because they are both consistent, given that the model used is correct. To keep support values roughly constant, we lengthened two of the terminal branches in the four-taxon tree with increased sequence length. The problem with this approach is that substitutions on those branches eventually get saturated. An alternative approach would be to shorten the internal branch, but for a four-taxon tree this means that the internal branch has to be shortened relatively more than the increase in sequence length. The result of this is that the number of expected substitutions on the internal branch quickly approaches zero.

Huelsenbeck et al. (2002) argued that one of the distinct advantages of Bayesian inference is that the posterior

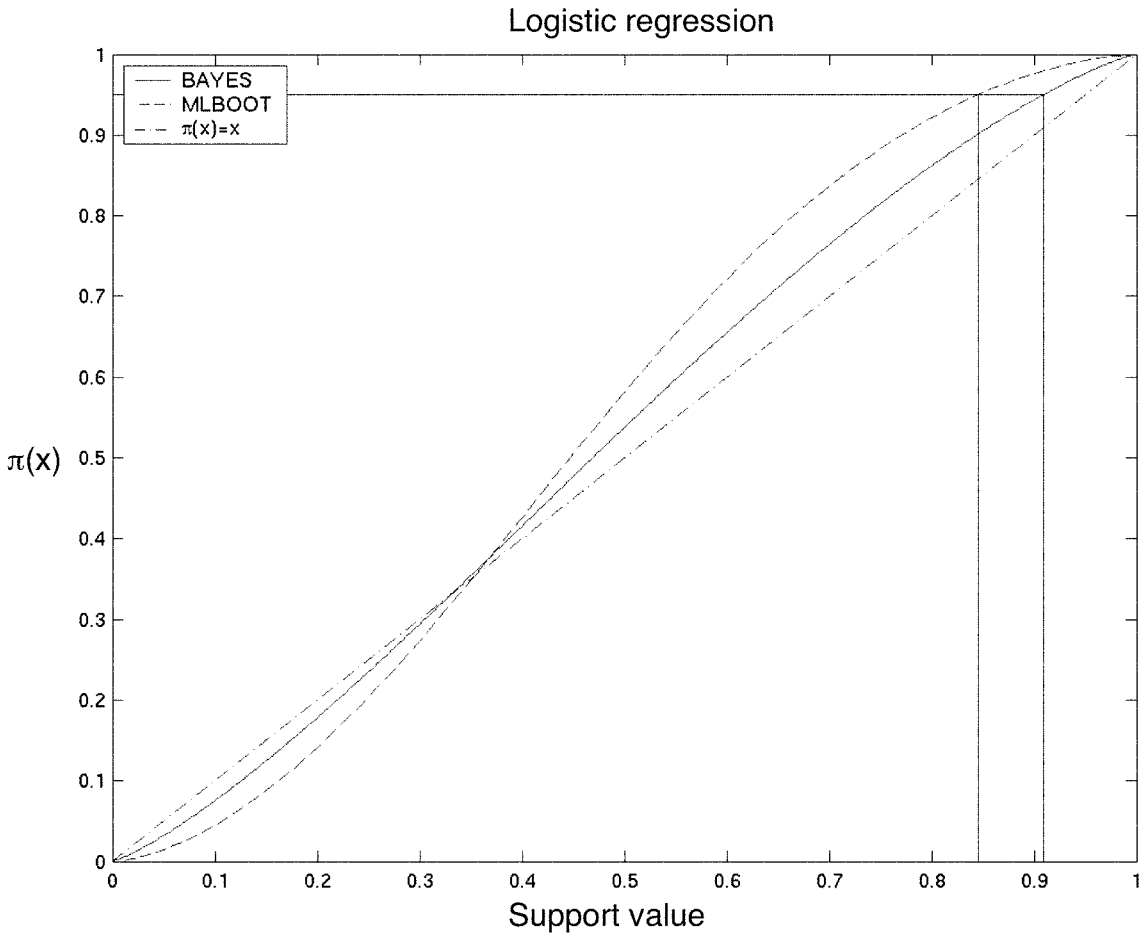


FIGURE 2. Estimated probability $\pi$ that the true clades are found as a function of support values $x_i$ for MLBOOT (dashed line, $\log\{\pi[x_i]/[1-\pi(x_i)]\} = 0.3282 + 1.5396\log[x_i/(1-x_i)]$) and BAYES (continuous line, $\log\{\pi(x_i)/[1-\pi(x_i)]\} = 0.1508 + 1.2118\log[x_i/(1-x_i)]$) using the $5_{2000}$ data sets and logistic regression with logit model and $\log[x_i/(1-x_i)]$ as explanatory variable. Help line indicates support values of BAYES and MLBOOT that correspond to 95% probability of the clade being true.

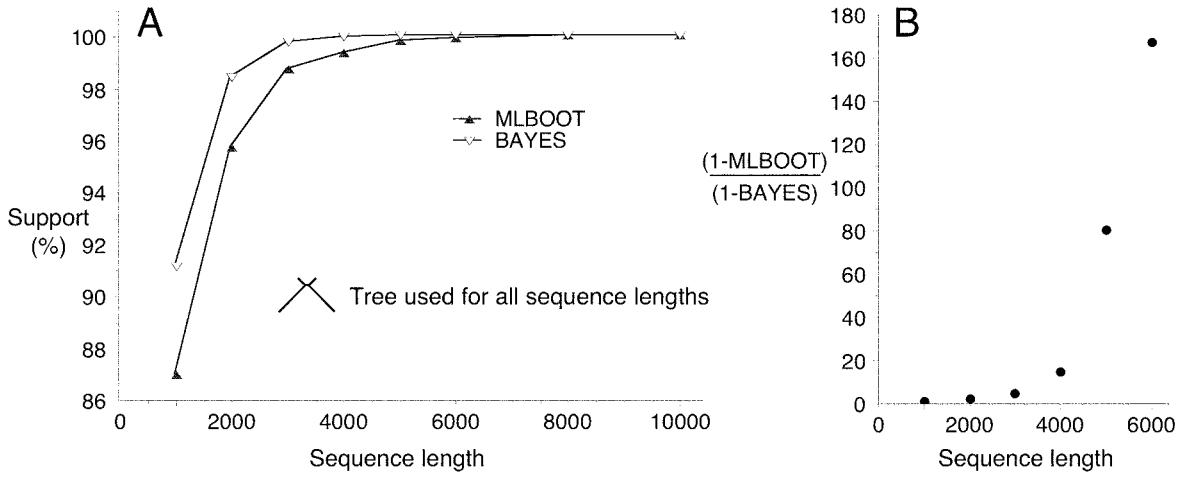


FIGURE 3. Comparison of BAYES and MLBOOT when sequence length is increased. Five hundred data sets for each sequence length were simulated on the four-taxon tree, (V:0.04, X:0.15):0.01,Y:0.04),Z:0.15), with the JC69 model. Data were analyzed under the correct model. (a) Support value as a function of sequence length; (b) increase of the ratio (1 − MLBOOT)/(1 − BAYES) with increased sequence length. BAYES has a higher rate of convergence than does MLBOOT.

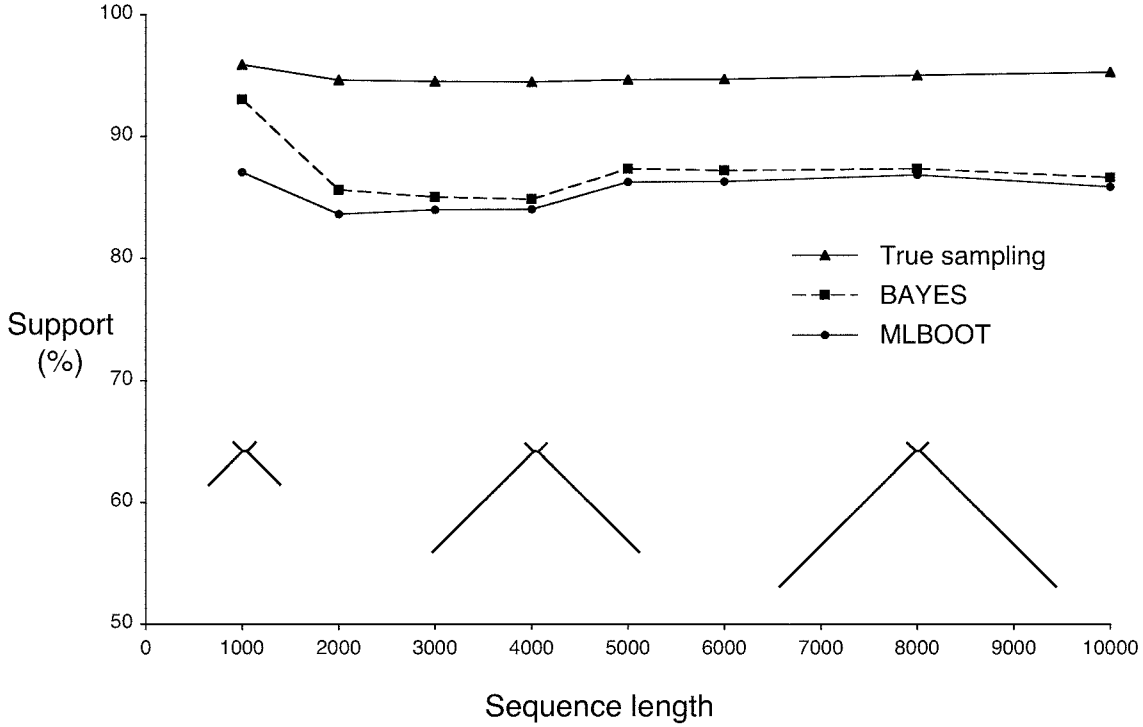


FIGURE 4. Mean support of MLBOOT, BAYES, and true sampling as a function of sequence length, when support is kept roughly constant. The length of two terminal branches were set using the function $Y = -1.3448676 + 0.2164043\ln(X)$, where Y is branch length and X is sequence length, i.e., branch length increases with 0.15 expected substitutions per site when sequence length is doubled. The internal branch and the other two terminal branches were held constant. For BAYES and MLBOOT, 500 data sets were simulated and analyzed with the JC69 model for each four-taxon tree/sequence length. Ten thousand data sets were analyzed for True sampling at each sequence length. The 1,000-bases tree, (V:0.04, X:0.15):0.01, Y:0.04),Z:0.15), the 4,000-bases tree, (V:0.04, X:0.45):0.01, Y:0.04),Z:0.45), and the 8,000-bases tree, (V:0.04, X:0.6):0.01, Y:0.04),Z:0.6), are shown.

probabilities have a clear-cut interpretation, i.e., they represent the probability that the corresponding clade is true given the model, the priors, and the data. This statement is not in agreement with the results of our study. With a reasonable agreement between model of simulation and model of analysis, our results show that BAYES is conservative, although less so than MLBOOT, in this particular case. Even if evolution followed the JC69 model, we would have model misspecification every time an erroneous topology is evaluated because this results in a different set of branch length parameters. Thus, the interpretation of posterior probabilities is not clear-cut in the context of phylogenetic reconstruction. Different nonnested models are compared in both BAYES and MLBOOT. It is far from clear how this difference should affect the interpretations of the results from the methods (Yang, 1997).

BAYES performs better than MLBOOT by getting closer to the actual probability that the clade exists in the true phylogeny, because the actual probability is underestimated more, on average, by MLBOOT. In other words, BAYES has a lower type II error. However, BAYES has a much higher type I error rate, especially in the case of model misspecification. BAYES values appear, on average, to be higher than MLBOOT values for well-supported true clades, and it takes fewer characters to get a certain support value with BAYES than with MLBOOT.

Huelsenbeck et al. (2002) suggested three potential explanations for the observed discrepancies between BAYES and MLBOOT. First, Bayesian analysis could be more sensitive to model misspecification, and because we do not generally know the correct model, this could explain the discrepancy. We have shown, for our particular example, that BAYES is more sensitive to underparameterization. However, this does not explain why we see differences in the analyses with correct model. Their second suggestion, that the corrected bootstrap method (Efron et al., 1996) should reduce the discrepancy, is odd because the correction was actually invented for those who wanted to use the bootstrap in a non-Bayesian sense (i.e., frequency probability). The third suggestion points to the methodological difference, that nonparametric bootstrapping simulates a stochastic distribution of the parameter space using profile likelihoods, whereas Bayesian methods use marginal likelihoods. We have shown that there can be a discrepancy between the support values of the two methods even when there is no detectable difference in how often they find the correct topology. Further, Efron et al. (1996) argued that nonparametric bootstrapping in phylogenetic trees accurately estimate posterior probabilities, although they gave no proof for this statement. Alfaro et al. (2003) argued that because data are parameterized differently in BAYES and MLBOOT, the methods are expected to give

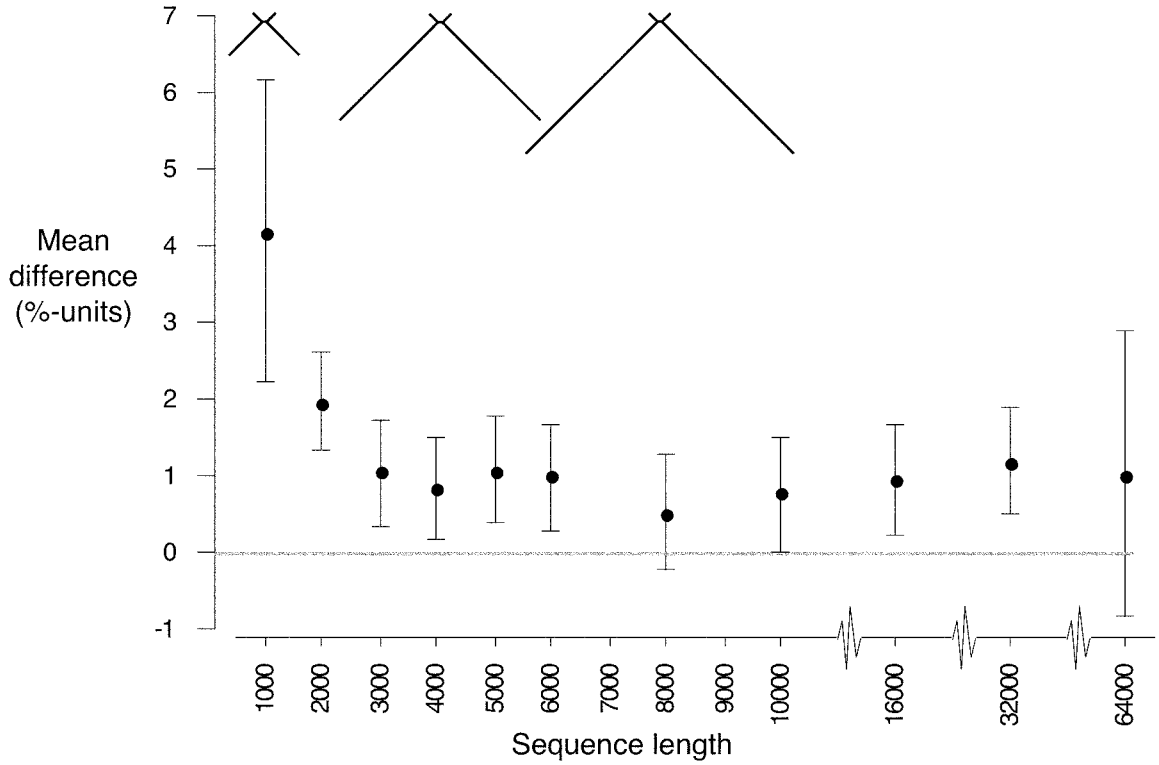


FIGURE 5. Mean difference of support values from BAYES and MLBOOT for different sequence lengths, with 95% confidence intervals (see Fig. 4 for details on tree modification with sequence length). Five hundred data sets were simulated and analyzed with the JC69 model for each four-taxon tree/sequence length. The 1,000-bases tree, the 4,000-bases tree, and the 8,000-bases tree are shown. The right part of the *x*-axis is not a linear scale.

different results. In MLBOOT, the site pattern frequencies are the only parameters. In BAYES, branch length, substitution rate matrix, tree topology, and base frequency are all parameterized, and the site patterns are not. This area of research clearly needs scrutiny, especially the relationship to sample size.

Nonparametric bootstrapping was originally introduced to phylogenetic reconstruction to give probabilities of clades representing their occurrence in an estimate based on many characters ($N \rightarrow \infty$) from the underlying pool of characters (Felsenstein, 1985). Our results show that both BAYES and MLBOOT probabilities are lower than those given by true sampling. In other words, both BAYES and MLBOOT are biased estimators of the probability of recovering a certain (true) topology if we were to sample the same number of characters over and over again.

Poorly supported clades are unreliable because those clades may have been recovered by chance, for example if the sample size is small. But if they are true, we have committed a type II error by not acknowledging a true relationship. Further data collection can be a remedy to this type of error. A high support value makes researchers believe in that clade relationship, but it could nevertheless be wrong (type I error). This is generally a more serious problem because researchers may not have any motivation to collect further data when there already exists a well-supported phylogeny. The behavior of BAYES is especially cumbersome in this context. Douady et al. (2003) found that slight differences in taxon or character sampling could give strongly supported conflicting topologies with BAYES. Suzuki et al. (2002) showed that BAYES often gives high support values for groups, even with completely uninformative data. They also observed that this behavior is accentuated when analyses were underparameterized, which corroborates our results. The nonparametric bootstrap is a long-used support measure, and even though its statistical interpretation is not clear, it is conservative, a property that we find desirable, at least when the model of evolution is unknown and probably more complex than any of the models available for analyses. It is tempting to use the higher support values generated by Bayesian inference as if they were equivalent. The effect of this would be a more frequent acceptance of false phylogenetic hypotheses. Model misspecification can, as we have shown, accentuate this problem. It is important to note that in our particular case of underparameterization, even though the type I error was very high for BAYES, the total number of true clades with a support value >95% increased much more. The behavior of BAYES under model misspecification is clearly an area in need of more research.

Because the true relationship between taxa rarely, if ever, can be observed, any phylogenetic hypotheses will always run the risk of being falsified in the light of new data. High support values cannot guarantee correct

conclusions, only well supported conclusions. More data, if sampled from the same phylogeny, make conclusions more reliable and robust.

## ACKNOWLEDGMENTS

We are grateful to Johan Nylander, Magnus Popp, and Fredrik Ronquist for discussions, and to Jack Sullivan and two anonymous reviewers for constructive criticism of an earlier version of this manuscript. This work was supported by a grant from the Linnaeus Centre for Bioinformatics, Uppsala University, and partly by a grant from the Swedish Science Foundation to B.O.

## REFERENCES

AGRESTI, A. 1990. Categorical data analysis. Wiley-Interscience, New York.

ALFARO, M. E., S. ZOLLER, AND F. LUTZONI. 2003. Bayes or bootstrap? A aimulation 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.

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.

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.

DURBIN, R., S. EDDY, A. KROGH, AND G. MITCHISON. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge Univ. Press, Cambridge, U.K.

EFRON, B., E. HALLORAN, AND S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:7085–7090.

FARRIS, J. S., V. A. ALBERT, M. KÄLLERSJÖ, D. LIPSCOMB, AND A. G. KLUGE. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

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

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. Mol. Biol. Evol. 9:366–369.

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.

HUELSENBECK, J. P., B. LARGET, R. E. MILLER, AND F. RONQUIST. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

HUELSENBECK, J. P., AND F. RONQUIST. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

HUELSENBECK, J. P., F. RONQUIST, R. NIELSEN, AND J. P. BOLLBACK. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

KAROL, K. G., R. M. MCCOURT, M. T. CIMINO, AND C. F. DELWICHE. 2001. The closest living relatives of land plants. Science 294:2351–2353.

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

LEACHÉ, A. D., AND T. W. REEDER. 2002. Molecular systematics of the eastern fence lizard *Sceloporus undulatus*: A comparison of parsimony, likelihood, and Bayesian approaches. Syst. Biol. 51:44–68.

MURPHY, W. J., E. EIZIRIK, S. J. O'BRIEN, O. MADSEN, M. SCALLY, C. J. DOUADY, E. TEELING, O. A. RYDER, M. J. STANHOPE, W. W. DE JONG, AND M. S. SPRINGER. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

OXELMAN, B., M. BACKLUND, AND B. BREMER. 1999. Relationships of the Buddlejaceae s.l. investigated using parsimony jackknife and branch support analysis of chloroplast *ndh*F and *rbc*L sequence data. Syst. Bot. 24:164–182.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73–79.

RAMBAUT, A., AND N. C. GRASSLY. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

RANNALA, B., AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

SANDERSON, M. J. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5:113–129.

SUZUKI, Y., G. V. GLAZKO, AND M. NEI. 2002. Overcredibility of molecular phylogenetics obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.

SWOFFORD, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland, Massachusetts.

WHITTINGHAM, L. A., B. SLIKAS, D. W. WINKLER, AND F. H. SHELDON. 2002. Phylogeny of the tree swallow genus *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol. Phylogenet. Evol. 22:430–441.

YANG, Z. 1997. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 14:105–108.

YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

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

*First submitted 13 December 2002; reviews returned 2 April 2003; final acceptance 15 May 2003*
*Associate Editor: Jack Sullivan*