Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence

Michael E. Alfaro, Stefan Zoller,† and François Lutzoni†*

*Evolution and Ecology, University of California, Davis; and †Department of Biology, Duke University

Bayesian Markov chain Monte Carlo sampling has become increasingly popular in phylogenetics as a method for both estimating the maximum likelihood topology and for assessing nodal confidence. Despite the growing use of posterior probabilities, the relationship between the Bayesian measure of confidence and the most commonly used confidence measure in phylogenetics, the nonparametric bootstrap proportion, is poorly understood. We used computer simulation to investigate the behavior of three phylogenetic confidence methods: Bayesian posterior probabilities calculated via Markov chain Monte Carlo sampling (BMCMC-PP), maximum likelihood bootstrap proportion (ML-BP), and maximum parsimony bootstrap proportion (MP-BP). We simulated the evolution of DNA sequence on 17-taxon topologies under 18 evolutionary scenarios and examined the performance of these methods in assigning confidence to correct monophyletic and incorrect monophyletic groups, and we examined the effects of increasing character number on support value. BMCMC-PP and ML-BP were often strongly correlated with one another but could provide substantially different estimates of support on short internodes. In contrast, BMCMC-PP correlated poorly with MP-BP across most of the simulation conditions that we examined. For a given threshold value, more correct monophyletic groups were supported by BMCMC-PP than by either ML-BP or MP-BP. When threshold values were chosen that fixed the rate of accepting incorrect monophyletic relationship as true at 5%, all three methods recovered most of the correct relationships on the simulated topologies, although BMCMC-PP and ML-BP performed better than MP-BP. BMCMC-PP was usually a less biased predictor of phylogenetic accuracy than either bootstrapping method. BMCMC-PP provided high support values for correct topological bipartitions with fewer characters than was needed for nonparametric bootstrap.

Introduction

Confidence measures play an important role in phylogenetics, especially when trees serve as the conceptual framework for the study of trait evolution. These measures allow workers to identify trees or parts of a tree that are well supported by the data and thus adequate to serve as the basis for evolutionary inference of biological systems (Huelsenbeck, Rannala, and Masly 2000; Lutzoni et al. 2001; Pagel and Lutzoni 2002). Arguably the most commonly used confidence method in phylogenetics has been nonparametric bootstrapping, a statistical technique invented by Efron (1979) and first applied to the phylogeny problem by Felsenstein (1985). Phylogenetic nonparametric bootstrapping involves the random resampling (with replacement) of characters from the original data to generate pseudoreplicate data matrices identical in size to the original matrix. These pseudoreplicates are then subjected to the same phylogenetic searches as the original data set. Bootstrap support for a group of interest is calculated as the proportion of times that the group is obtained in the pseudoreplicates.

The rationale for the resampling of the original matrix is that the distribution of the pseudoreplicates around the observed data is a valid approximation of the distribution of observed data sets on the true, unknown process that generates the data sets (Efron 1979; Efron Halloran, and Holmes 1996). In phylogenetic terms, this suggests that a monophyletic group that receives a high bootstrap proportion would be expected to be recovered by other analyses of new data sets that were generated by the same underlying process (Felsenstein 1985), and it is for this reason that the bootstrap is sometimes described as a measure of repeatability (Berry and Gascuel 1996; Felsenstein 1985; Hillis and Bull 1993).

Hillis and Bull (1993) examined the performance of nonparametric bootstrapping as a measure of phylogenetic accuracy, that is, the probability that a given monophyletic group appears on the true tree. Their finding, that bootstrap proportions greater than 50% underestimated phylogenetic accuracy, sparked a flurry of papers that sought to clarify the interpretation of bootstrap '*P*' values (Felsenstein and Kishino 1993; Li and Zharkikh 1994; Sanderson 1995; Zharkikh and Li 1995; Berry and Gascuel 1996; Newton 1996). An important point that emerged from this scrutiny was that phylogenetic accuracy, *sensu* Hillis and Bull (1993), is not a quantity that bootstrapping typically tests and, furthermore, that bootstrapping may overestimate or underestimate phylogenetic accuracy depending on the condition under which the data were generated (e.g., Efron, Halloran, and Holmes 1996; Felsenstein and Kishino 1993). In addition, type I error, which is the quantity that many workers desire the bootstrap to reflect, is only approximated by the conventional bootstrap procedure (Efron, Halloran, and Holmes 1996). A more complex, two-step bootstrapping procedure is necessary to transform bootstrap proportions into standard frequentist confidence intervals. Despite these studies, conventional nonparametric bootstrapping is still widely viewed as providing a measure of phylogenetic accuracy (e.g., Murphy et al., 2001).

Thus, nonparametric bootstrapping has been used to measure three quantities (Berry and Gascuel 1996): *repeatability,* the probability of observing a given result in future repeated sampling of the same underlying

Key words: Bayesian Markov chain Monte Carlo, bootstrap, maximum parsimony, maximum likelihood, posterior probability, phylogenetic confidence, simulation.

E-mail: malfaro@ucdavis.edu.

Mol. Biol. Evol. 20(2):255–266. 2003
DOI: 10.1093/molbev/msg028

character distribution; *accuracy* (Hillis and Bull 1993), the probability that a given monophyletic group is present on the true tree; and *type I error rate* (Felsenstein and Kishino 1993), assuming a null model of nonmonophyly. The theoretical justification for interpreting nonparametric bootstrap values as measures of repeatability is quite strong (Efron and Tibshirani 1993, 1998; Efron, Halloran, and Holmes 1996), and most of the debate over the bootstrap has focused on whether and how the bootstrap proportion can be meaningfully related to phylogenetic accuracy and frequentist testing (e.g., Sanderson 1995; Berry and Gascuel 1996). Threatening to add to the confusion over the interpretation of bootstrap values in phylogenetics is the increasingly widespread use of Bayesian methods to calculate the Bayesian confidence limits (posterior probabilities) for monophyletic relationships.

Bayesian Confidence Methods

Bayesian inference in phylogenetics has become increasingly common since its development in the late 1990s (see reviews in Huelsenbeck et al. 2001; Lewis 2001). Broadly speaking, in Bayesian inference one makes use of Bayes's theorem to condition inferences about the value of some parameter of interest on the observed data. Bayesian inference focuses on the quantity known as the posterior probability, defined as the probability of some hypothesis conditional on the observed data. The posterior probability is proportional to the product of the likelihood of the data, given that the hypothesis is correct and the prior probability of the hypothesis before any data have been collected

In Bayesian phylogenetics, parameters such as the tree topology, branch lengths, and substitution parameters, are modeled as probability distributions. Using Bayes's theorem, the posterior probability of any of one of these parameters may be expressed as the marginal distribution of those remaining. Solving analytically for the posterior probability requires the integration of the likelihood function over all possible values of the remaining parameters, which is effectively intractable for even moderately complex problems. Modern Bayesian methods use Markov chain Monte Carlo methods to approximate this integration by simulating draws from the joint posterior distribution of all model parameters. Posterior probabilities for the parameters of interest are calculated using the Markov chain samples. For example, the posterior probability of a tree or bipartition in a tree is determined simply by examining the proportion of all of the Markov-chain samples that contain the topological bipartition of interest.

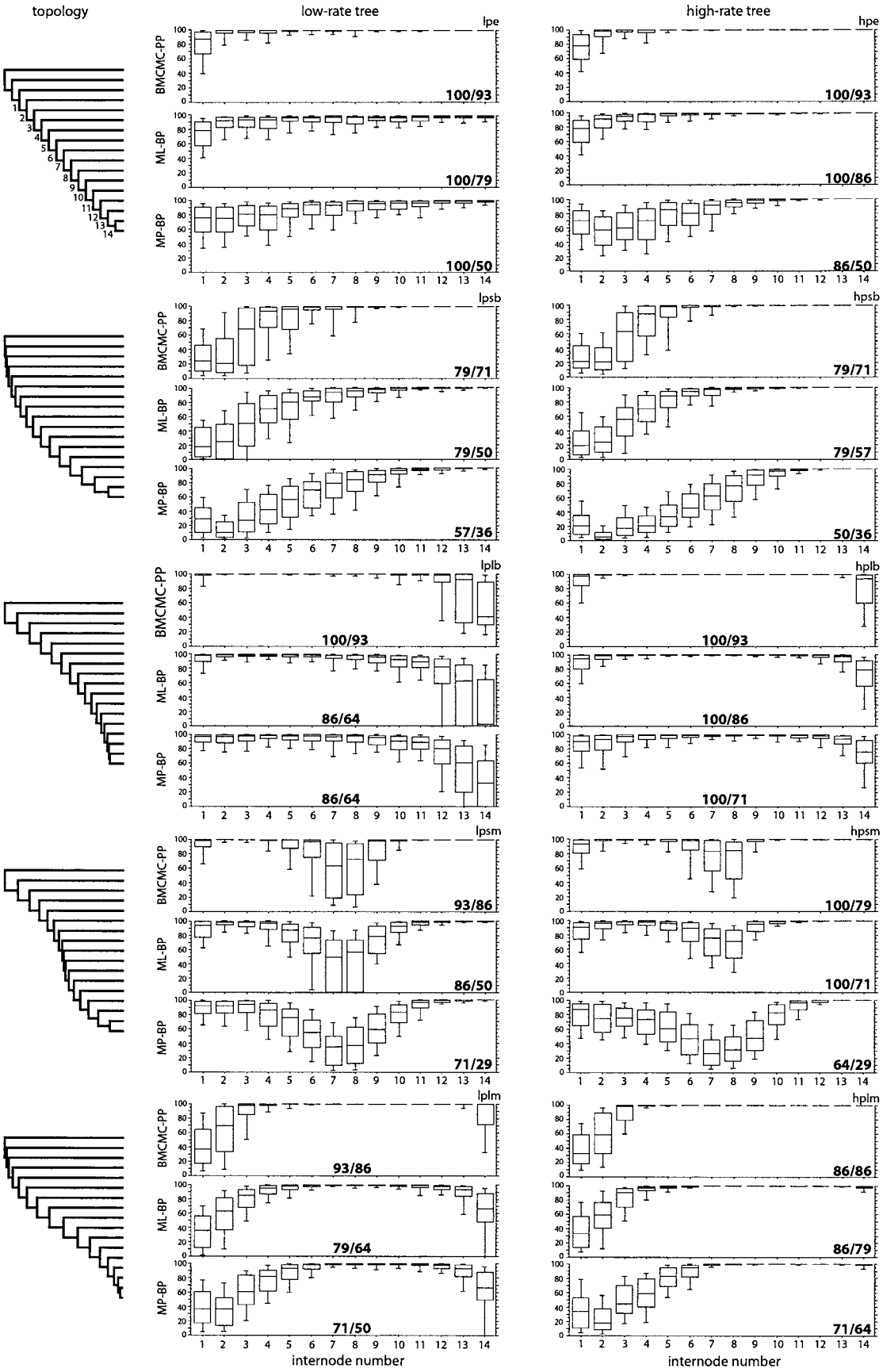
The Meaning and Measure of Confidence Values

Despite the growing popularity of Bayesian methods in phylogenetics (see Lewis 2001), there is no current consensus of how posterior probabilities should be interpreted relative to more traditional support measures such as the bootstrap. Efron, Halloran, and Holmes (1996) pointed out that bootstrap values correspond closely to posterior probabilities calculated under a multinomial model of site pattern frequency, and some workers have also implied that posterior probabilities derived from standard likelihood models of sequence evolution (i.e., those calculated by programs such as MrBayes [Huelsenbeck 2000] and BAMBE [Larget and Simon 1999]) are also closely equivalent to likelihood bootstrap proportions (e.g., Larget and Simon 1999). Others have noted that posterior probabilities are often much higher than the associated bootstrap proportion and have cited this as evidence that the Bayesian posterior probabilities do not suffer from the conservative bias that has been attributed to bootstrap values with regards to phylogenetic accuracy (Murphy et al. 2001).

The purpose of the current study is to investigate the comparative behavior of nonparametric bootstrapping and Bayesian Markov chain Monte Carlo (BMCMC) methods in assigning confidence to phylogenetic results. Simulations are powerful tools for evaluating the performance of phylogenetics methods because the true tree and generating model are known a priori (e.g., Hillis, Allard, and Miyamoto 1993). For this study, we chose to explore the performance of these methods under evolutionary scenarios that were designed to approximate a single gene study (1,000 base pairs) of a moderate number of taxa (17) rather than a simple four-taxon case in order to obtain a better understanding of how these methods perform under conditions more typical of real data sets. We quantified performance of support methods on a range of these topologies to address several fundamental questions about Bayesian posterior probabilities and bootstrap proportions. First, we compare how bootstrap and BMCMC assign confidence to the same correct internodes on a tree to determine if they are essentially equivalent techniques. Second, we compare the width of confidence envelopes for these two kinds of confidence by adopting the traditional interpretation of the bootstrap as a measure of repeatability and the posterior probability as the probability that a monophyletic group is correct. Third, we investigate the performance of these methods in estimating phylogenetic accuracy and explore the consequence of constructing decision rules from support values on rates of type I error and on other performance benchmarks that we derive from our simulations. Finally, we compared the sensitivity

→

FIG. 1.—Comparison between Bayesian and nonparametric bootstrap methods in assigning confidence to the same correct internodes on pectinate topologies. Shown are box plots, which indicate the 10%, 25%, median, 75%, and 90% interval boundaries of support for each of the 14 internodes on the indicated topology. Results from low-rate trees (0.08 expected changes per site as measured from the root of the tree to any tip) are in the first column, and results for high-rate trees (0.30 expected changes per site from root to tip of the tree) are in the second column. For each scenario, Bayesian Markov chain Monte Carlo posterior probabilities (BMCMC-PP) are shown in the top plot, followed by maximum likelihood bootstrap proportion values (ML-BP) and maximum parsimony bootstrap proportion values (MP-BP). Numbers in bold in lower right or lower middle of each graph indicate the median percentage of correct nodes (out of 14 possible) that received support ⩾70%/⩾95% over the 100 replicates.



of these methods to phylogenetic signal by describing the effects of increasing character number on support value.

Methods

Simulations

We simulated data on fully pectinate and fully symmetric topologies of 17 taxa (details of the topological conditions examined in this study are available online at http//:www.molbiolevol.org.) We varied branch lengths of the pectinate trees in the following ways: all internodes equal, basal internodes short, basal internodes long, middle internodes short, middle internodes long (fig. 1). Symmetric topologies were varied in the same ways except that we did not examine cases where the middle internodes were short or long (fig. 2). In addition, we constructed one symmetric topology that did not assume a molecular clock (fig. 2). We examined each of these nine topologies at low (0.08 expected changes per site) and high (0.30 expected changes per site) rates of character evolution. In total, our simulation universe was made up of $2 \times (5 \text{ pectinate} + 4 \text{ symmetric trees}) = 18$ evolutionary scenarios. For each scenario we used Seq-Gen (Rambaut and Grassly 1997) to evolve 100 data sets with 1,000 base pairs of sequence, each under a Kimura 2-parameter model with a transition: transversion ratio of 2 (K2P with ti:tv = 2). This relatively simple model was chosen to provide a fair comparison between model-based (maximum likelihood and Bayesian) methods and equally weighted parsimony, without being completely unrealistic.

Our study largely focused on differences between likelihood nonparametric bootstrapping and Bayesian posterior probabilities. As currently implemented, both methods require a model of evolution in a fundamentally similar way and so a comparison between likelihood bootstrapping and BMCMC is informative about their relative performance in estimating confidence. Due to the time required to perform likelihood bootstrapping, parsimony bootstrapping is sometimes the only confidence measure reported, even in cases where both maximum parsimony and maximum likelihood topologies are obtained (e.g., Koulianos and Schmid-Hempel 2000; Rodriguez-Robles and De Jesus-Escobar 2000). We chose to include parsimony bootstrapping in our study to investigate whether parsimony bootstrap values provided a reasonable estimate of the likelihood bootstrap proportion under our simulation conditions. However, it is useful to point out that we expect parsimony bootstrapping to perform more poorly than either of the model-based methods; the deck is effectively stacked against it because the underlying model of sequence evolution is also the model used by likelihood bootstrapping and BMCMC to assess confidence.

For each of the 100 replicates, we calculated Bayesian posterior probabilities and bootstrap proportions for all internodes. To calculate posterior probabilities of internodes, we used MrBayes 1.1 (Huelsenbeck 2000) to run a 100,000-generation Markov chain under a Kimura 2-parameter model, sampling every 100 generations. We used the default (flat) priors for the transition-transversion ratio (uniform 0–100), branch length (uniform 0–10), and tree topology. Base frequencies were fixed to be equal. We ran one cold and three heated chains simultaneously (Geyer 1991; Huelsenbeck and Ronquist 2001). Visual inspection of samples from each scenario suggested that the Markov chain reached stationarity within 5,000 generations, but we discarded the first 30,000 generations to ensure that stationarity was reached. Posterior probabilities were calculated from the 700 remaining trees by examining the frequency of occurrence of correct bipartitions in the MCMC sample. Maximum parsimony and maximum likelihood bootstrap proportions were calculated using the full heuristic search option in PAUP* (Swofford 1998) version 4b8 with 200 total pseudoreplicates saving all equally optimal trees, Tree Bisection-Reconnection branch swapping, and two random addition sequences per pseudoreplicate. The maximum likelihood bootstrapping analyses were performed the same way as the parsimony bootstrapping analyses, except that searches were conducted under the model used to generate the data (K2P with ti:tv = 2), whereas maximum parsimony searches on bootstrapped data sets were performed assuming that all costs for changes among nucleotides were equal to one step.

Performance Benchmarks

We used a number of benchmarks to assess the performance of maximum parsimony bootstrap proportions (MP-BP), maximum likelihood bootstrap proportions (ML-BP), and Bayesian Markov chain Monte Carlo posterior probabilities (BMCMC-PP). First, we examined the correlation of support values among all three methods across all scenarios to examine the degree of correspondence between bootstrapping and Bayesian analysis. Second, we compared support assigned by each method to a tree by calculating median support of each of the 14 internodes across all 100 replicates. We also examined the median percentage of correct internodes across the tree (# supported/14) that each method supported above arbitrary cut-off values of 70% and 95% over the 100 replicates. This provided a heuristic indication of the relative capability of these methods to assign support to correct internodes. One may also use this measure to determine the type I error rate (the rate of rejecting the null model when it is true, which in the case of our simulations is the rate of rejecting correct topological bipartitions) for basing a decision rule on either of these cut-off values. Type I error rate for a particular decision rule can be calculated as 1 minus the proportion of total supported correct internodes.

We compared the relative tendencies of these methods to assign high support to wrong internodes in two ways. First, we examined the raw number of incorrect internodes that each method supported above 70% and 95%. Second, we found the 95% threshold value for each method by following the procedure outlined below.

1. For each method and each scenario, we assembled all incorrect topological bipartitions that received support values of 1% or greater. Support values for wrong

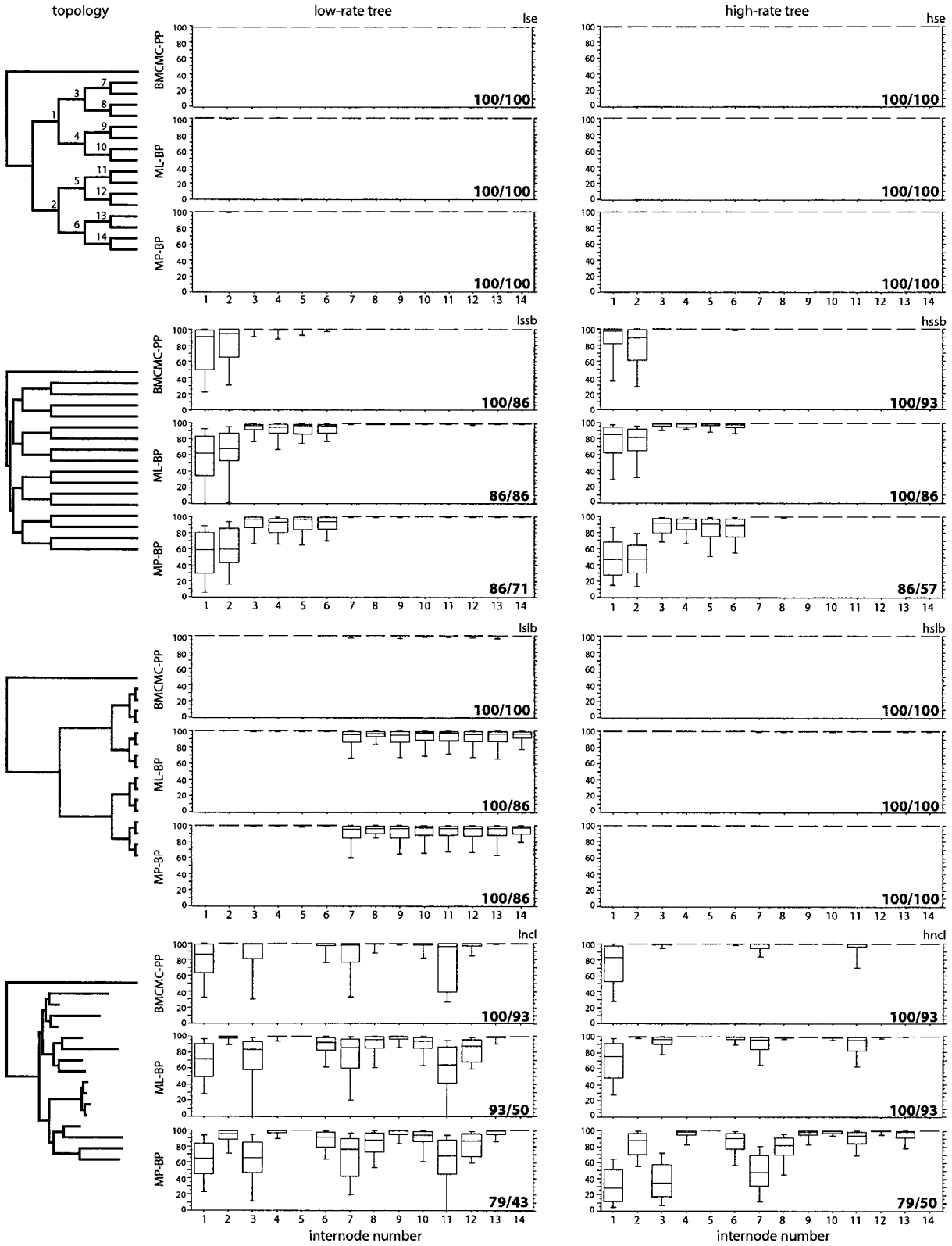


FIG. 2.—Comparison between Bayesian and nonparametric bootstrap methods in assigning confidence to the same correct internodes on symmetric topologies. Shown are box plots that indicate the 10%, 25%, median, 75%, and 90% interval boundaries of support for each of the 14 internodes on the indicated topology. Results from low-rate trees (0.08 expected changes per site as measured from the root of the tree to any tip) are in the first column, and results for high-rate trees (0.30 expected changes per site from root to tip of the tree) are in the second column. For each scenario, Bayesian Markov chain Monte Carlo posterior probabilities (BMCMC-PP) are shown in the top plot, followed by maximum likelihood bootstrap proportion values (ML-BP) and maximum parsimony bootstrap proportion values (MP-BP). Numbers in bold in lower right hand of each graph indicate the median percentage of correct nodes (out of 14 possible) that received support ≥70%/≥95% over the 100 replicates.

bipartitions across replicates of the same scenario were pooled.
2. For each method and each scenario, we created a frequency histogram ordered by support values. These plots were always heavily skewed to the left because most of the incorrectly supported topological bipartitions had support values of 10% or less.
3. Moving along the x-axis from 0% to 100% support value of the frequency histogram, we determined the support value that accommodated 95% of the incorrect bipartitions. Less than 5% of all of the incorrect, supported topological bipartitions received support higher than this threshold value. Methods that tend to assign high support to a larger number of internodes should have a higher threshold value than those that do not. We compared the 95% threshold value for each method.
4. Finally, we compared the number of correct topological bipartitions that were supported at or above the 95% threshold value for incorrect bipartitions.

The number of topological bipartitions supported at the 95% threshold value thus provides a standard for comparison among the three support methods. It represents a trade-off between accepting correct internodes while simultaneously ensuring that 95% of the incorrect, supported internodes are rejected.

We also calculated the phylogenetic accuracy of these methods. To determine the relative performance of Bayesian and bootstrapping methods in estimating accuracy (*sensu* Hillis and Bull 1993), we calculated this parameter as a function of increasing support value. For a given support category, accuracy equaled the number of correct internodes divided by the number of correct and incorrect internodes across the 100 replicates of the scenario, multiplied by 100.

To examine the effects of increasing data set size on confidence level, we performed two simulations in which the number of characters was gradually increased using the clocklike and nonclocklike symmetric topologies (lse and lncl trees; fig. 2). We used Seq-Gen (Rambaut and Grassly 1997) to evolve 5,000 base pairs of sequence on each topology under the same model (K2P, ti:tv = 2) used in the earlier simulations. Confidence analyses were performed as outlined above on an initial data set size of 50 characters for the lse tree and 100 characters for the lncl tree. Data set size was gradually increased until all methods assigned 95% support values or the 5,000-character limit was reached. Simulations were repeated 25 times for each topology to calculate the median support value for each data set size.

Results

Correlation Among Support Values from Bayesian MCMC and Nonparametric Bootstrap Methods

We excluded three scenarios (lse, hse, and hslb) from our calculations of correlations between Bayesian and bootstrap methods, because all or nearly all of the support values were 100%. For 12 of the 15 remaining scenarios, BMCMC-PP and ML-BP values were strongly correlated ($r^2 > 0.8$, $P \leqslant 0.05$) (correlations for topological scenarios lpe, lssb, and hssb were only weakly correlated [$r^2 < 0.8$]). ML-BP support values showed strong correlation with MP-BP for only seven scenarios (lncl, lplb, lplm, lpsb, lpsm, lslb, and lssb), whereas BMCMC-PP and MP-BP correlated strongly in only a single scenario (lplb). For the 12 scenarios in which BMCMC-PP and ML-BP were strongly correlated, half were low-rate trees and half were high-rate trees. In contrast, ML-BP and MP-BP were strongly correlated only on low-rate trees.

Comparison of Bootstrap and Bayesian Methods in Assigning Confidence to Specific Internodes

Across all 18 scenarios, some general patterns of support were evident. All three methods tended to assign lower support to regions at the base of the tree and where relative branch length was short (figs. 1 and 2). Median BMCMC internodal support was almost always equal to or higher than ML and MP bootstrap support. In problematic regions of the tree, MP-BP was usually lower than ML-BP and was sometimes much lower. Except for the nonclock-like symmetric trees (lncl and hncl), the pectinate topologies generally contained more problematic regions than the symmetric topologies. Median nodal support was usually higher on high-rate trees than on low-rate trees for BMCMC-PP and ML-BP. In contrast, median support from MP-BP decreased on many internodes when compared with their low-rate counterpart trees, especially in scenarios where internodes at or near the base of the tree were relatively short (see hpsb, hpsm, hssb, and hncl scenarios, figs. 1 and 2). For symmetric topologies, BMCMC and ML bootstrap support values were improved when using data sets generated with high-rate trees compared with data derived from low-rate trees. As expected, the opposite behavior was observed for MP bootstrap. For the most basal internodes of pectinate topologies, all methods performed more poorly with high-rate tree data sets than with low-rate tree data sets. However, the reverse was true for the most apical internodes of pectinate topologies.

Variance in support value was usually greater for MP-BP than for ML-BP and BMCMC-PP. For the latter, variance in nodal support was generally smaller than variance for ML-BP when internode length was relatively long. Median support for internodes in these situations was also generally very high. When internode length was short, BMCMC-PP variance sometimes exceeded that of ML-BP (e.g., lpsb internode 3 and lplm internode 2, fig. 1). Median support for the shortest internodes was generally lower, although it was sometimes still quite high (e.g., lplb internode 13, fig. 1). Internodes that showed extreme fluctuation in support values were generally very short. In most cases, these internodes were so short that maximum likelihood was unable to reconstruct the internode in all 100 replicates of a particular scenario (results not shown). For any given cut-off value, BMCMC-PP always assigned support to an equal or greater average number of correct internodes than either MP-BP or ML-BP, and ML-BP

always performed as well as, or better than, MP-BP (figs. 1 and 2). Thus, for arbitrarily chosen decision rules of 70% and 95%, rates of type I error (the rate of rejecting true internodes) were lower for BMCMC-PP than for either bootstrapping method. Differences in performance among support methods were most apparent at the highest confidence levels (≥95%) and were quite striking in some instances. For example, on the low-rate pectinate topology with equal length internodes (lpe) approximately nine correct internodes received a PP of 95% compared with approximately six and four that were supported by ML-BP and MP-BP, respectively. Symmetric topologies appeared to pose less of a challenge to the reconstruction algorithms than did the pectinate topologies. In three scenarios (lse, hse, and hslb), all methods assigned 100% support to all internodes. Furthermore, differences among the three methods were generally higher on the high-rate trees than on the low-rate trees.

None of the methods assigned support to a large number of incorrect internodes, which was not surprising given the relatively favorable evolutionary conditions under which we simulated these data sets (table 1). Parsimony assigned moderate (>70%) support to the largest number of incorrect internodes (~2.3 internodes/scenario for MP-BP versus ~0.7 internodes/scenario for ML-BP and ~1.9 internodes/scenario for BMCMC-PP, averaged over all 18 scenarios). BMCMC-PP assigned high (>95%) support to more incorrect internodes than either bootstrapping method (~0.14 internodes/scenario for BMCMC-PP, ~0.03 internodes/scenario for ML-BP, and ~0.06 internodes/scenario for MP-BP, averaged over all 18 scenarios), although the overall rate of assigning high support to incorrect internodes was extremely low. However, as a result of this tendency of wrong topological bipartitions to have higher posterior probabilities than bootstrap proportions, the 95% threshold value (the support value that was greater than or equal to 95% of the support values that *wrong* internodes received) was highest for BMCMC-PP (fig. 3*A*). Using a decision rule constructed to minimize the rate of accepting incorrect bipartitions would generally allow one to recover most correct monophyletic relationships regardless of the support method (fig. 3*B*). However, ML-BP recovered slightly more correct internodes than BMCMC-PP (13.8 versus 13.6) and both model-based methods recovered more internodes than MP-BP (13.1). MP-BP also showed the greatest variance in performance across scenarios, occasionally recovering fewer than 12 correct internodes/tree.

Although all three methods assigned high support to few incorrect internodes, we identified some scenario replicates in which BMCMC-PP assigned a 95% or greater posterior probability to an incorrect internode, whereas ML-BP and MP-BP assigned much lower support (table 1). These internodes were all found in regions of low-rate trees with the shortest internodes, and maximum likelihood trees for these replicates also contained the wrongly supported internodes. Thus, sampling error associated with evolving data at a slow rate on regions of the model topology with the shortest internodes could occasionally produce data sets with signal that was incongruent with the model topology.

Table 1
Number of Wrong Topological Bipartitions Receiving Moderate and High Support

Simulation	BMCMC-PP	ML-BP	MP-BP
lpe	5[a],0[b]	2,0	9,0
lpsb	3,2	4,0	9,0
lplb	13,1	1,0	4,0
lpsm	12,0	2,0	3,0
lplm	23,2	8,0	18,1
hpe	4,0	3,0	15,1
hpsb	25,1	15,0	35,0
hplb	9,2	6,0	5,0
hpsm	16,2	5,1	27,1
hplm	18,2	12,1	31,0
lse	0,0	0,0	0,0
lssb	0,0	0,0	0,0
lslb	25,1	15,0	35,0
lncl	12,1	2,0	7,0
hse	0,0	0,0	0,0
hssb	2,0	3,0	5,0
hslb	0,0	0,0	0,0
hncl	9,1	6,0	50,4

[a] Number of times across all 100 simulation replicates that an incorrect topological bipartition received a support value greater than 70.

[b] Number of times across all 100 simulation replicates that an incorrect topological bipartition received a support value greater than 95.

Comparison of Bayesian and Bootstrap Methods in Estimating Phylogenetic Accuracy

We were unable to plot accuracy versus increasing support for four symmetric scenarios (lse, hse, lslb, and hslb) because all or nearly all of the internodes received 100% support. In the remaining 14 scenarios (fig. 4), all three methods generally underestimated the true accuracy at levels of support greater than 50%. This bias was often less pronounced for BMCMC-PP. However, the latter overestimated accuracy at moderately high support levels in one scenario (lplm). BMCMC-PP appeared to lie closest to the line of perfect correspondence between accuracy and support for most scenarios.

For any particular topology, posterior probabilities and bootstrap proportions showed the greatest disparity on the shortest internodes. When we examined the effects of branch length on support across all scenarios, we found that posterior probabilities exceeded 95% for many very short internodes (as short as 1.3 expected changes). In contrast, maximum parsimony and likelihood bootstrap proportions did not reach 95% on branches shorter than three expected changes. BMCMC-PP assigned 100% confidence to some internodes with as few as 1.3 expected changes in contrast to ML-BP, which required at least 5 expected changes and MP-BP, which required 6.7 expected changes. ML and MP bootstrap proportions of 70% or more were obtained for branch lengths as short as 1.7 expected changes.

Sensitivity to the Amount of Phylogenetic Signal

Simulation on lse and lncl topologies to investigate the effects of increasing number of characters on support values revealed that the BMCMC-PP assigned 95% support to all internodes with a smaller number of characters relative to both bootstrapping methods. On the symmetric clocklike topology lse (fig. 5*A*), tip internodes

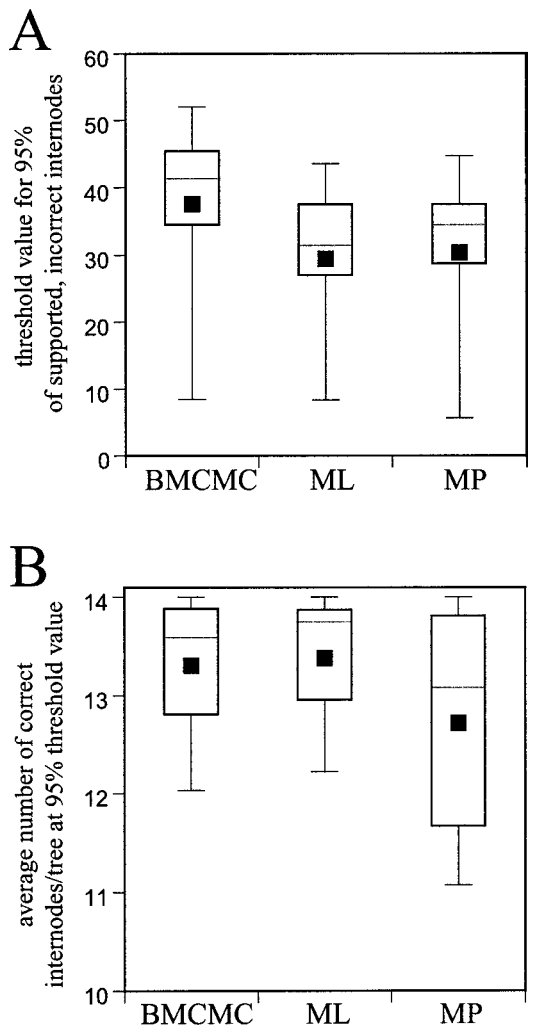


FIG. 3.—Comparison between Bayesian and nonparametric bootstrap methods in assigning support values for incorrect internodes. Shown are box plots that indicate the 10%, 25%, median, 75%, and 90% interval boundaries of 95% threshold values for incorrect, supported bipartitions from 15 scenarios. Average support values are indicated by squares within the box plot. Due to an extremely low number of incorrect topological bipartitions, it was not possible to calculate threshold values for three scenarios (lse, hse, and hslb), which were excluded from the analysis. (*A*) Distribution of 95% threshold values for incorrect, supported bipartitions from 15 scenarios. On average, all three methods assigned low support to most of the incorrectly supported bipartitions, although Bayesian Markov chain Monte Carlo posterior probability (BMCMC-PP) assigned higher support to a greater proportion of them than did either bootstrapping method. (*B*) Distribution of the number of correct internodes supported at or above the 95% threshold value from *A* for each of the 18 scenarios. See text for an explanation of how this threshold value was determined.

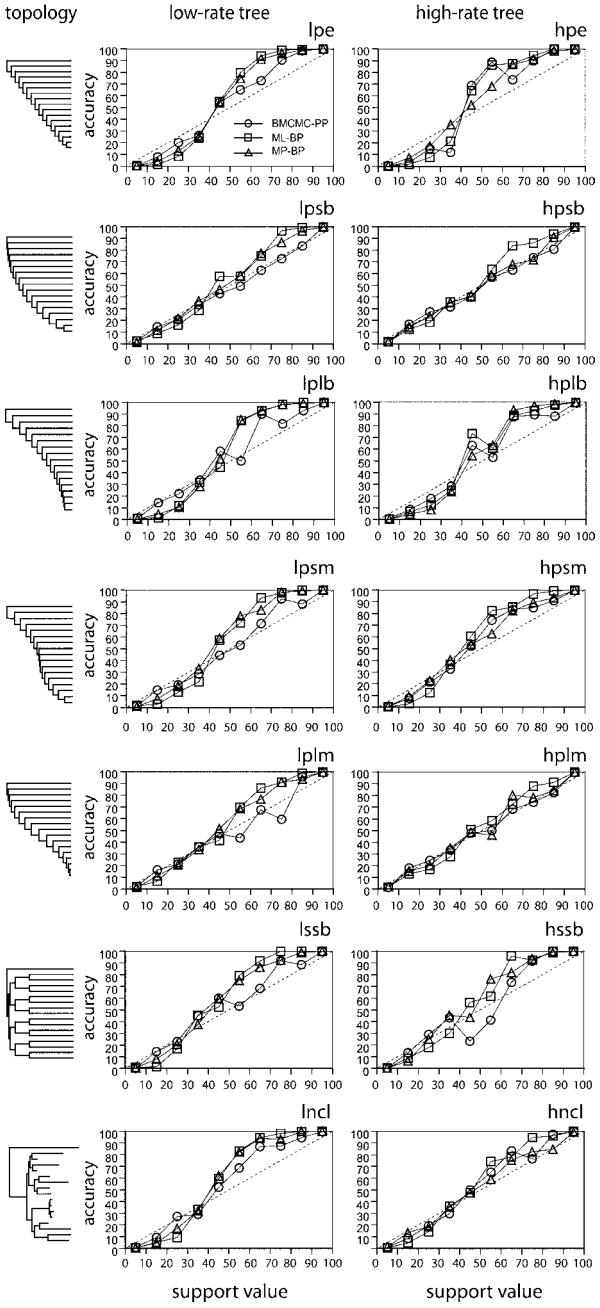


FIG. 4.—Relationship between phylogenetic accuracy and Bayesian and nonparametric bootstrap methods. Accuracy was computed at 10% intervals and is equal to the number of correct internodes divided by the number of correct and incorrect internodes times 100 across the 100 replicates. Dotted line represents perfect correspondence between support and accuracy.

(7 to 14) received a median support of 95% with 100 to 150 characters for BMCMC-PP, compare with 200 to 300 characters for ML-BP and MP-BP. All internodes received 95% support at 200 characters with BMCMC-PP, compared with 300 characters for ML-BP and 350 for MP-BP. We observed a similar pattern on the nonclocklike topology lncl (fig. 5*B*), with all internodes reaching a 95% posterior probability at 1,600 characters and a 95% likelihood bootstrap proportion at 2,000 characters. Parsimony bootstrap values decreased with increasing data set size for internodes 1, 3, and 7 on this tree. This is most likely due to maximum parsimony being inconsistent under such conditions (see *Discussion*). Even when these three internodes are excluded from this comparison, MP-BP constantly required the largest number of characters when discrepancies among methods were detected (fig. 5*B*). In several cases (e.g., lse topology, internodes 11, 12, and 13) BMCMC-PP reached support values of 95% or higher with fewer characters than MP-BP required to reach support values of 70% or higher. In the most extreme

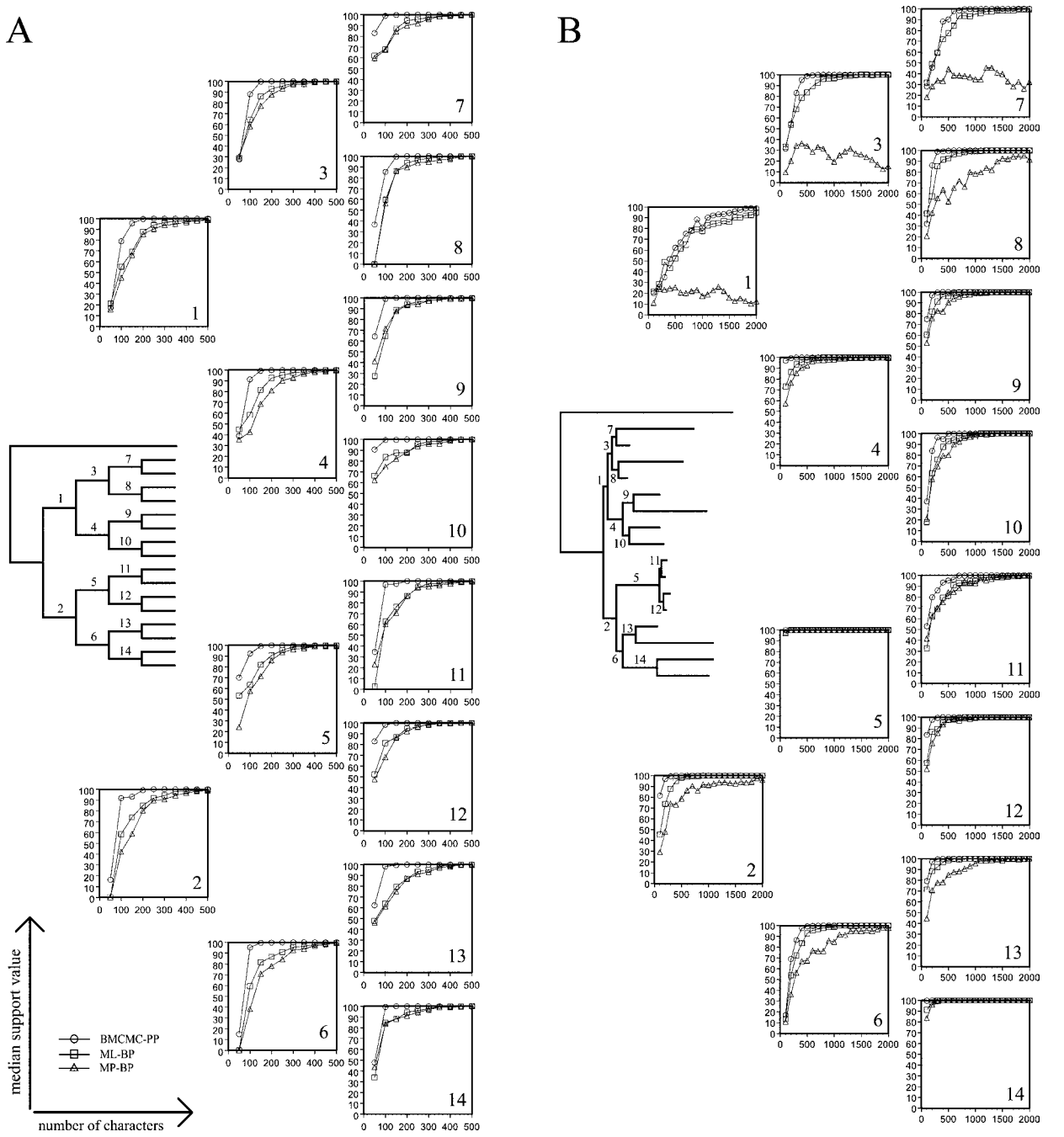


FIG. 5.—Comparison of Bayesian and nonparametric bootstrap support values with increasing number of characters on (*A*) clocklike and (*B*) nonclocklike symmetric topologies (lse and lncl trees). Each simulation was repeated 25 times for each topology to calculate the median support value for each data set size. BMCMC-PP = Bayesian Markov chain Monte Carlo posterior probability, ML-BP = maximum likelihood bootstrap proportion, MP-BP = maximum parsimony bootstrap proportion.

example, on internode 8 of the lncl topology, BMCMC-PP reached 95% support with 300 characters while BP-MP required 700 characters to reach a 70% support value.

Discussion

Behavior and Performance of Bayesian MCMC Versus MP and ML Bootstrap

On the basis of our simulations, we draw a number of conclusions about the comparative behavior and performance of Bayesian MCMC sampling and bootstrapping in assessing phylogenetic confidence. Most generally, Bayesian posterior probabilities and the bootstrap proportion are not equivalent measures of confidence. For a given data set BMCMC-PP will, on average, attach high confidence to a greater number of correct internodes than does nonparametric bootstrapping (figs. 1 and 2). As a result, type I error rates (the frequency of rejecting true monophyletic groups) for an arbitrary determined decision rule are likely to be lower for BMCMC-PP than for bootstrapping.

However, when support values are standardized by the 95% threshold value for wrong monophyletic groups, BMCMC-PP and ML-BP perform very similarly in recovering correct internodes (fig. 3*B*). MP-BP performs more poorly and has higher variance, although all methods performed well under most of our simulation conditions. Bayesian posterior probabilities are a better estimator of accuracy (*sensu* Hillis and Bull 1993) than bootstrap proportions for support values greater than 50% (fig. 4; Wilcox et al. 2002). Thus a posterior probability of 70% is likely to correspond to an accuracy near 70%, whereas a 70% bootstrap proportion is often close to a 95% accuracy (Hillis and Bull 1993). Workers who desire to interpret posterior probabilities as an accuracy indicator should take care to adopt appropriately high posterior probabilities to ensure that they are reaching the desired level of accuracy.

Confidence on Short Internodes

We found the greatest disparity in Bayesian and bootstrapping methods to occur in regions of the tree where internodes are short or the number of characters is relatively small (e.g., figs. 1, 2, and 5). It thus appears that BMCMC-PP is able to attach very high posterior probabilities to branches with very small amounts of character change (Kauff and Lutzoni 2002) and may indicate that this method has a greater sensitivity to the signal in a data set than are ML-BP and MP-BP. The disparity may be quite large in some instances, and a rather substantial increase in number of characters may be required to get the likelihood or parsimony bootstrap proportion to converge on the posterior probability (e.g., fig. 5).

This is an attractive behavior of BMCMC-PP because it offers the possibility of obtaining confidence on short internodes, which are commonly both particularly interesting and poorly supported by bootstrapping in many studies. However, the price for this increased sensitivity could be an increased tendency of BMCMC-PP to assign high confidence to incorrect internodes, especially in situations where character sampling has not been sufficient to recover the correct topology. In this scenario, analogous to the simulation replicates that contained wrong internodes with high posterior probabilities in our study (table 1), sampling error on regions of the tree with small amounts of character change will occasionally produce data matrices that produce an incorrect internode under likelihood analysis. It is possible that the number of characters supporting this incorrect internode will sometimes fall below the critical value necessary to garner a high bootstrap value (Felsenstein 1985), although the posterior probability will reach 95% or higher. In contrast, the inherent low sensitivity of character resampling methods such as bootstrapping, when very few characters contribute to a specific internal branch length, may sometimes be a virtue. Workers relying on Bayesian posterior probabilities to assess confidence should pay particular attention to the length of supported branches since our results suggest that occasionally very short, wrong branches may receive a high posterior probability.

Why Are Bootstrap Proportions and MCMC-Calculated Posterior Probabilities Different?

From a Bayesian perspective, the multinomial model of site pattern frequencies proposed by Efron (Efron 1979; Efron, Halloran, and Holmes 1996) as one explanation of the bootstrap procedure is vastly different from the model of sequence evolution employed by MrBayes. In the first, the site pattern frequencies are the only parameters. In the second, branch length, substitution rate matrix, tree topology, and base frequency are all parameterized and the site patterns are not. Given the distinct differences in how the data are modeled, it is not surprising, at least to the Bayesian, that the posterior distributions on internodes do not correspond exactly: they are formulaically different. A fairly straightforward way to interpret posterior probabilities and bootstrap proportions would be to acknowledge that these methods measure different features of the data. For example, an internode with a very high posterior probability but a moderate bootstrap value should be interpreted as an internode that has a high probability of being correct, conditional on the data that has been collected so far (and the model of evolution). It is also an internode that is highly dependent on the underlying composition of the data matrix and as such may not be observed when additional characters are gathered. Both methods may provide useful measures of the data.

Parsimony as an Alternative to Likelihood Bootstrapping

As expected, equally weighted maximum parsimony bootstrapping usually performed poorly relative to maximum likelihood bootstrapping under our simulation conditions. The difference between likelihood and parsimony bootstrapping was most dramatic in the pectinate scenarios, where parsimony usually supported less than half of all correct internodes with a bootstrap proportion greater than 80%. Furthermore, MP bootstrapping was usually more susceptible than ML bootstrapping to assigning high support values to incorrect internodes (fig 3 and table 1). We attribute this result to higher statistical inconsistency for MP (e.g., Huelsenbeck 1997, 1998) because the estimates differed most strongly in regions of the trees with a combination of relatively short internodes and long branches (e.g., internodes 1, 3, and 7 of figure 5*B*). Our results suggest that even in relatively simple scenarios with moderate amounts of data, long-branch attraction may negatively affect confidence estimates on a phylogeny when using equally weighted MP as the optimization criterion. Therefore, equally weighted MP bootstrapping should not be used to approximate ML bootstrap values.

Bayes or Bootstrap?

To answer this question, phylogeneticists must have some idea of what they would like their confidence method to measure. Nonparametric bootstrapping is appropriate if one is interested in the sensitivity of observed results to the sampling error associated with collecting characters from a hypothesized underlying character distribution. If one is

willing to specify a fully probabilistic model of character evolution and wishes to place confidence limits on the results of an analysis conditioned on the observed data and that model, Bayesian posterior probabilities are the appropriate confidence measure to use. In cases where one decides to bootstrap, it is useful to note that it may require a relatively large amount of data to obtain high confidence on short internodes (Berbee, Carmean, and Winka 2000) compared with BMCMC-PP. When assessing posterior probabilities, it is important to remember that confidence values estimated on extremely short internodes may sometimes be sensitive to the underlying stochastic process.

We suggest that BMCMC-PP is a useful method to use when systematists wish to show how well the data support the results of model-based phylogenetic analysis. The posterior probability enjoys a straightforward interpretation as the probability that a particular monophyletic group is correct, which may be how most systematists already interpret bootstrap proportions on optimal trees. Furthermore, the posterior probability appears to have some desirable frequentist properties, particularly with regard to the rate of rejecting true topological bipartitions. BMCMC-PP also appears to have increased sensitivity to phylogenetic signal, which may allow workers to achieve high confidence in a correct result with fewer characters. However, BMCMC-PP also appears to be more susceptible than likelihood bootstrapping to assigning high confidence to incorrect short internodes. This should not be interpreted as an indication that nonparametric bootstrapping is always less likely to provide high support to wrong relationships. Because MP-BP is more sensitive to long-branch attraction than are ML-BP and BMCMC-PP, nonparametric bootstrapping is more likely to assign high support values to incorrect internodes when parsimony is chosen as the optimization criterion. Additional work is needed to determine the circumstances where the more conservative nature of likelihood bootstrapping may be preferred to the increased power of BMCMC-PP.

Acknowledgments

We thank Keith Barker, Peter Wagner, Michael Sanderson, Wes Johnson, Jeff Thorne and Hirohisa Kishino for helpful discussion over the course of this project. This project was funded in part by National Science Foundation (NSF) grants DEB-9615542 and DEB-0133891 to F.L. Most simulations were conducted on a High Performance Computer Cluster acquired through a grant from NSF Major Research Instrumentation (DBI-9871374) in part to F.L.

Literature Cited

Berbee, M. L., D. A. Carmean, and K. Winka. 2000. Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough? Mol. Phylogenet. Evol. **17**:337–344.

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

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.

———. 1998. The problem of regions. Ann. Stat. **26**:1687–1718.

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

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.

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. Pp. 156–163 *in* E. M. Keramidas, ed. Computing science and statistics: proceedings of the 23rd Symposium on the Interface. Interface Foundation, Fairfax Station, Va.

Hillis, D. M., M. W. Allard, and M. M. Miyamoto. 1993. Analysis of DNA sequence data: phylogenetic inference. Methods Enzymol. **224**:456–87.

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. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. **46**:69–74.

———. 1998. Systematic bias in phylogenetic analysis: is the *Strepsiptera* problem solved? Syst. Biol. **47**:519–537.

———. 2000. MrBayes: Bayesian inference of phylogeny, version 2.01. Distributed by the author.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accomodating phylogenetic uncertainty in evolutionary studies. Science **288**:2349–2350.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. 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.

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

Koulianos, S., and P. Schmid-Hempel. 2000. Phylogenetic relationships among bumble bees (*Bombus*, Latreille) inferred from mitochondrial cytochrome b and cytochrome oxidase I sequences. Mol. Phylogenet. Evol. **14**:335–341.

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.

Lewis, P. 2001. Phylogenetic systematics turns over a new leaf. Trends Ecol. Evol. **16**:30–37.

Li, W., and A. Zharkikh. 1994. What is the bootstrap technique? Syst. Biol. **43**:424–430.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature **411**:937–940.

Murphy, W. J., E. Eizirik, S. J. O'Brien et al. (11 coauthors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294**:2348–2351.

Newton, M. A. 1996. Bootstrapping phylogenies: large deviations and dispersion effects. Biometrika **83**:315–328.

Pagel, M., and F. Lutzoni. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. Pp. 151–164 *in* M. Laessig, and A. Valleriani, eds. Biological evolution and statistical physics. Springer Verlag, Berlin.

Rambaut, A., and N. C. Grassly. 1997. SeqGen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp. Appl. Biosci. **13**:235–238.

Rodriguez-Robles, J. A., and J. M. De Jesus-Escobar. 2000. Molecular systematics of new world gopher, bull, and pinesnakes (*Pituophis*: Colubridae), a transcontinental species complex. Mol. Phylogenet. Evol. **14**:35–50.

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. **44**:299–320.

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Wilcox, T. P., D. J. Swickl, 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.

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

Nick Goldman, Associate Editor

Accepted October 11, 2002