Syst. Biol. 52(4):477–487, 2003

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390218213

Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case

MICHAEL P. CUMMINGS,[1] SCOTT A. HANDLEY,[2] DANIEL S. MYERS,[1,3] DAVID L. REED,[4] ANTONIS ROKAS,[5] AND KATARINA WINKA[6]

[1]*Center for Bioinformatics and Computational Biology, 2119 A.V. Williams, University of Maryland, College Park, Maryland 20742, USA; E-mail: mike@mbl.edu (M.P.C.)*
[2]*Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri 63110-1093, USA; E-mail: sahandle@artsci.wustl.edu*
[3]*Pomona College, Claremont, California 91711-7004, USA; E-mail: dmyers@pomona.edu*
[4]*Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA; E-mail: reed@biology.utah.edu*
[5]*Howard Hughes Medical Institute and Laboratory of Molecular Biology, University of Wisconsin, Madison, Wisconsin 53706, USA; E-mail: arokas@wisc.edu*
[6]*Department of Ecology and Environmental Science, Umeå University, SE-901 87 Umeå, Sweden; E-mail: katarina.winka@bmg.umu.se*

Abstract.— Assessment of the reliability of a given phylogenetic hypothesis is an important step in phylogenetic analysis. Historically, the nonparametric bootstrap procedure has been the most frequently used method for assessing the support for specific phylogenetic relationships. The recent employment of Bayesian methods for phylogenetic inference problems has resulted in clade support being expressed in terms of posterior probabilities. We used simulated data and the four-taxon case to explore the relationship between nonparametric bootstrap values (as inferred by maximum likelihood) and posterior probabilities (as inferred by Bayesian analysis). The results suggest a complex association between the two measures. Three general regions of tree space can be identified: (1) the neutral zone, where differences between mean bootstrap and mean posterior probability values are not significant, (2) near the two-branch corner, and (3) deep in the two-branch corner. In the last two regions, significant differences occur between mean bootstrap and mean posterior probability values. Whether bootstrap or posterior probability values are higher depends on the data in support of alternative topologies. Examination of star topologies revealed that both bootstrap and posterior probability values differ significantly from theoretical expectations; in particular, there are more posterior probability values in the range 0.85–1 than expected by theory. Therefore, our results corroborate the findings of others that posterior probability values are excessively high. Our results also suggest that extrapolations from single topology branch-length studies are unlikely to provide any general conclusions regarding the relationship between bootstrap and posterior probability values. [Bayesian analysis; Markov chain Monte Carlo sampling; maximum likelihood; phylogenetics.]

The goal of phylogenetic analysis is to infer the historical relationships among entities (e.g., genes, organisms) based on the information contained in their heritable characteristics (e.g. DNA sequences, morphology). These inferences are conditional on a particular statistical model, which attempts to describe relationships in the data. Along with deriving relationships, many researchers wish to assess such inferences using a statistic that is interpreted in terms of confidence, reliability, support, or robustness. This is an important step in phylogenetic reconstruction because it gives a quantitative measure of how well inferences are jointly supported by the data and the chosen model. The purpose of the research described here was to characterize the relationship between two such statistics used in phylogenetics: the proportion of bootstrap replicates and the posterior probability.

The Bootstrap

The nonparametric bootstrap (henceforth referred to simply as the bootstrap) is a computer-based statistical technique that uses data resampling to estimate values of interest (Efron, 1979; Efron and Tibshirani, 1993) and was first applied to phylogenetic analysis by Felsenstein (1985). In the context of phylogenetic analysis, the bootstrap starts by sampling, with replacement, individual characters (e.g., an individual nucleotide site across all taxa in the set) from the original data sample D, which are used to construct a new bootstrap sample of the original size, D^*. The bootstrap sample is then analyzed to infer a phylogenetic tree $\hat{T}^*$ using any method of choice (e.g., maximum likelihood, maximum parsimony, neighbor joining). The process is repeated for a specified number of replicates B and summarized. Although the bootstrap can be applied to any parameter of interest, in phylogenetic analysis the bootstrap is almost exclusively applied to tree topology, τ. The proportion of times a particular phylogenetic relationship is observed across B replicates, P_{boot}, can be interpreted as a probability, which is used as a measure of "confidence" in, or "support" for, the phylogenetic relationships inferred from the data. More explicitly, Felsenstein and Kishino (1993) suggested that $1 - P_{\text{boot}}$ is the probability of falsely favoring a relationship that is not present under the null hypothesis (i.e., type I error). Although use of the bootstrap in phylogenetic analysis has been somewhat controversial (reviewed by Sanderson, 1995), it has been the most widely applied statistical assessment of inferred phylogenetic relationships.

Posterior Probabilities

An important development in phylogenetic analysis has been the application of Bayesian methods, particularly when coupled with Markov chain Monte Carlo (MCMC) methods (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999;

Newton et al., 1999). The Bayesian approach involves determining the posterior probability of interest given a prior probability, likelihood function, and data. The intractability of direct mathematical determination of posterior probabilities has lead to the application of stochastic estimation procedures such as MCMC. A MCMC approach can be generally described as an algorithm-led trip through parameter space, where parameter space is defined in terms of topology, branch lengths, substitution rates, and other parameters. The algorithm involves a series of proposed parameter modifications, each of which is accepted or rejected as a function of the resulting change in likelihood. When the process reaches stationarity, the frequency with which individual parameter values are observed is an accurate estimation of the posterior probability.

In contrast, methods based on an objective criterion (e.g., maximum likelihood, maximum parsimony, minimum evolution) are little concerned with parameter values observed during the journey through parameter space. Instead, they use a particular algorithm (e.g., complete enumeration, branch-and-bound, various heuristics) to reach a final destination, the best tree as determined by the objective criterion. In the Bayesian/MCMC approach there is no particular destination. Rather, it is the journey itself that is of greatest interest. More detailed explanations of Bayesian analysis and MCMC methods as applied to phylogenetic inference have been presented elsewhere (e.g., Larget and Simon, 1999; Huelsenbeck et al., 2001; Lewis, 2001).

Application of Bayesian/MCMC methods in phylogenetic analysis results in posterior probability values, which give the probability of a tree, $P(\hat{T})$, given the data, D, or $P(\hat{T} \mid D)$. A tree, T, is characterized by the topology, τ, and associated branch lengths, β; thus, $T = (\tau, \beta)$. Our focus is on the posterior probability values, $P(\tau \mid D)$, which are interpreted as the support of the data for a specific topology.

Bootstrap versus Posterior Probability Values

The exact relationship of bootstrap values, P_{boot}, to posterior probability values, $P(\tau \mid D)$ is an open and important question in phylogenetic analysis. While the theory of each measure is largely independent, it has been posited that they should be equivalent (Efron et al., 1996). However, several observations suggest that these statistics may differ (e.g., Buckley et al., 2002; Leaché and Reeder, 2002; Suzuki et al., 2002; Whittingham et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Douady et al., 2003). Therefore, a need exists for additional studies focusing on the behavior and relationship of the bootstrap and posterior probability measures. Unfortunately, an analytical solution is not readily apparent. Here, we describe our work to characterize the relationship of P_{boot} to $P(\tau \mid D)$ using simulation and hypothesis testing. Our experimental design, described in more detail below, compares the means from many replicate P_{boot} and $P(\tau \mid D)$ estimates, and hence we compared the expected values of these statistics. Specifically, we tested the null hypothesis that the expectations of the two measures are equal,

$$H_0\text{: } E(P_{\text{boot}}) = E[P(\tau \mid D)],$$

compared with the general alternative,

$$H_1\text{: } E(P_{\text{boot}}) \neq E[P(\tau \mid D)].$$

We also examined an additional topological model, a star topology, as has been done by others (e.g., Gaut and Lewis, 1995; Suzuki et al., 2002). The motivation was to compare bootstrap and posterior probability values with a theoretical expectation and to corroborate or not the finding of Suzuki et al. (2002) that $P(\tau \mid D)$ values are excessively high.

METHODS

Model Space

We examined a model space that has been useful in exploring other aspects of phylogenetic analysis: the so-called four-taxon case (Huelsenbeck and Hillis, 1993; Gaut and Lewis, 1995). A tree with four terminal taxa has five branches (four external branches and one internal branch) and can be viewed as a two-dimensional matrix, with the abscissa representing the lengths of three branches (two external and the internal) and the ordinate representing the lengths of two branches (the other two external) (Fig. 1). A tree with four terminal taxa has three possible resolved unrooted topologies: the model topology (τ_1) from which the data are simulated and two alternative (incorrect) topologies (τ_2 and τ_3).

A single internal branch divides (defines) the nontrivial partition (Fig. 2). Hence, P_{boot} and $P(\tau \mid D)$ for the internal branch (partition) are of interest. In our study, the topology τ_2 is the result of the long-branch attraction problem. The model space design is similar to that of Gaut and Lewis (1995), with branch lengths in units of proportion of expected nucleotide substitutions ranging from 0.02 to 0.74 in increments of 0.02. This design yields a model space of $37^2 = 1{,}369$ elements (Fig. 1). All analysis parameters were chosen to achieve a balance among realism, relevance, variance and statistical power, and computational complexity.

Sequence Simulations

For each element in the model space, we generated 1,000 replicate data sets, each containing four simulated DNA sequences with branch lengths corresponding to the parameters of the four-taxon case. Sequence length was 1,000 nucleotides, a length similar to that considered in many molecular systematics studies. Sequences were generated using a general time reversible model (Tavaré, 1986) with substitution rates across sites following a continuous Γ distribution (Wakeley, 1993; Yang, 1993). The parameter values were based on maximum likelihood estimates from 10 vertebrate mitochondrial genomes (Cummings et al., 1995, 1999; Otto et al., 1996):

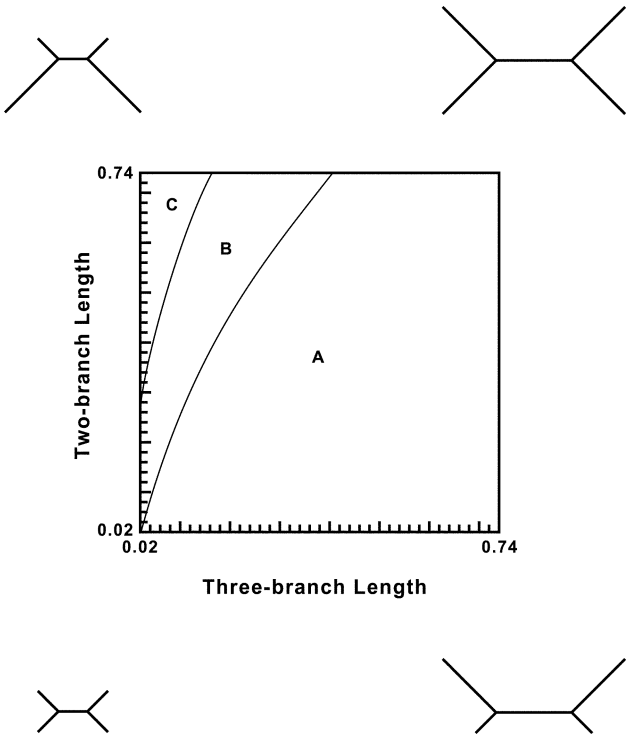


FIGURE 1. Analytical model space presented as a two-dimensional diagram. The abscissa represents the lengths of three branches (two external and the internal: branches 1, 3, and 5), and the ordinate represents the lengths of two branches (the other two external: branches 2 and 4) following the numbering described in Figure 2. Branch lengths are proportional to the expected difference accumulated along the branch. The regions referred to as the neutral zone (A), the near two-branch corner region (B), and the two-branch corner region (C) are shown. Branch lengths are not drawn to scale.

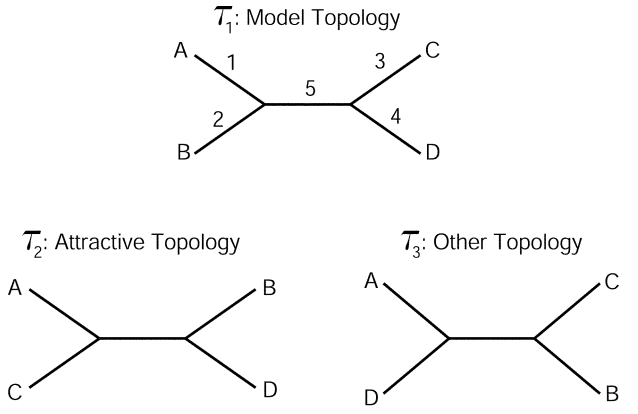


FIGURE 2. Three possible fully resolved topologies from four taxa. τ_1 is used in the simulations (model topology); τ_2 is the result of long-branch attraction overwhelming the signal for the model topology (attractive topology); and τ_3 is the other topology. The numerals denote branches defining the model space: three branches (1, 3, 5) and two branches (2, 4).

rate parameters T ↔ C = 1.93, T ↔ A = 0.42, T ↔ G = 0.28, C ↔ A = 0.58, C ↔ G = 0.13, and A ↔ G = 1.00; Γ distribution shape parameter $\alpha = 0.44$; base frequencies T = 0.25, C = 0.28, A = 0.34, and G = 0.13. To generate data sets for a star topology, we simulated sequences with expected external branch lengths equal and an internal branch length of zero. Simulated DNA sequences were generated using a slightly modified version of the program evolver, part of the PAML package (Yang, 1997).

Phylogenetic Analyses

We used each set of simulated sequences as data to infer the phylogenetic relationships by two methods: maximum likelihood bootstrap using the program PAUP* 4.0b9 and 4.0b10 (Swofford, 2002) and Bayesian analysis using the program MrBayes 2.01 (Huelsenbeck and Ronquist, 2001). The phylogenetic analyses were paired so that both maximum likelihood and Bayesian analyses were performed on the same set of sequences for each replicate in each element of the model space. With the paired design of the experiment, we eliminated one possible source of variance, i.e., variance due to different simulated sequences between maximum likelihood and Bayesian analyses. The likelihood model parameters included six nucleotide substitution categories and an among-site rate heterogeneity parameter, which used a Γ distribution model with four discrete rate classes represented by the mean of each class. Yang (1994) showed that four discrete rate classes produces a near-optimum fit when approximating a continuous Γ distribution rate model for DNA sequences.

For the maximum likelihood analyses, values of the Γ distribution shape parameter α, nucleotide substitution rates, and base frequencies were estimated from the data for the original simulated sequence D, and the estimated values were then applied in the analysis of the bootstrap samples, D^*. Branch-and-bound searching was used in all cases, ensuring that the tree with the highest likelihood was evaluated and thus eliminating another possible source of variance, i.e., variance due to not finding the tree of maximum likelihood. The number of bootstrap replicates, B, for each simulated sequence set was 2,000, a number giving a precision of 0.01 at $P_{\text{boot}} = 0.95$ (Hedges, 1992), and the percentage cutoff value in PAUP* was set to 0.

The decision to use parameter estimates from the tree with the highest likelihood from the original data in subsequent bootstrap replicates was a thoughtful design decision by the authors. We believe that this method represents the best current practice procedure and is based on the observation that likelihood surfaces are flat and that likelihood values are relatively insensitive to moderate changes in parameter values. Furthermore, the maximum likelihood tree represents the best point estimate of the phylogeny for the data, and the estimates of model parameters values associated with the maximum likelihood tree represent the best estimates of these values for the data. However, it is also possible to reestimate the likelihood model parameters for each individual bootstrap replicate, a procedure that requires much more computation. We investigated the difference between using the estimate of the model parameters for the original

data and reestimating these parameters for each individual bootstrap replicate. The experiment used a paired design to compare the bootstrap values obtained using these two procedures for 1,000 simulated samples for a single point in the model space, the element in the upper left corner of Figure 1.

For the Bayesian analyses, values of the Γ distribution shape parameter α, nucleotide substitution rates, and base frequencies were estimated from the data as necessary at each chain step. Initial exploratory investigations were conducted to determine appropriate chain length, burn-in, and sampling frequency of the Bayesian analyses using the program BOA 1.0.0 (Smith, 2001) in conjunction with the R statistical computing system (Ihaka and Gentleman, 1996). The exploration examined multiple simulated sequence sets from the extremes of the model space (each of the four corners in Fig. 1) and a broad range of chain lengths and burn-in levels. The negative log (−ln) likelihood values were evaluated for convergence of the joint posterior probability using the Brooks, Gelman, and Rubin diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998), which is based on within-chain and between-chains variance. Values of the diagnostic corrected scale reduction factor were <1.01 in $<10^4$ chain steps for sequence sets representing all model extremes, which is consistent with sampling from a stationary distribution. The sampling frequency was evaluated graphically by examining autocorrelation plots of the joint posterior probability. With a lag of ≤ 20 steps, autocorrelation values for −ln likelihood were stationary. Each final Bayesian analysis was performed with four chains (one cold and three heated) of length 5×10^4 steps, with the last 4×10^4 steps sampled at a frequency of every 20 steps, yielding 2,000 observations, equal to the number of bootstrap replicates. The value of the temperature parameter for heated chains was 0.2. Summary statistics were calculated from the sampled observations using a Perl program.

Theoretical Expectations for a Star Topology

Bootstrap and posterior probability values across all three possible resolved topologies have some simple properties in the case of a star topology, which we used as a null model. For the bootstrap, the null model expectations include $\bar{x}(P_{\text{boot}}) = 0.3333$ and $\sum_{i=1}^{3} \tau_i P_{\text{boot}} = 1$. Similarly, for posterior probability $\bar{x}[P(\tau \mid D)] = 0.3333$ and $\sum_{i=1}^{3} P(\tau_i \mid D) = 1$. However, the maximum values of each measure among the possible resolved topologies for each replicate analysis, $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$, are also of interest and are the basis for the values reported by Suzuki et al. (2002). Therefore, it is necessary to know the theoretical expectations for $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ under the null model. There are obvious constraints on the theoretical maximum values, $0.3333 < \max < 1$, but the distribution of these maximum values is less obvious. To determine the distribution, we referred to the relevant theory, which was first described in terms of harmonic analysis by Fisher (1929). In more general terms, an equivalent problem is picking $n - 1$ points at random along the unit interval (0,1) and determining the distribution of the longest interval between points (David, 1981:98–100), with $n = 3$ in the present case. We applied this theory to simulate 10^4 sets of 1,000 expected maximum values each using a Perl program and compared these values with $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ based on analyses of 1,000 simulated data sets representing sequences for star topologies with an expected proportion of nucleotide substitutions along each branch of 0.74.

Hypothesis Testing

We chose to test the null hypothesis H_0: $E(P_{\text{boot}}) = E[P(\tau \mid D)]$, using the statistic d, the magnitude of the differences between the mean proportion of bootstrap replicates $\bar{x}(P_{\text{boot}})$, and the mean posterior probability values supporting the internal partition $\bar{x}[P(\tau \mid D)]$. Under the null hypothesis the expectation for this difference, $E(d)$, is 0. We used a permutation test (Manly, 1991; Good, 1994; Maritz, 1995) to assess the significance of this difference. For each set of 1,000 paired values at each element of the model space, we randomly exchanged values at a probability of 0.5 and determined the sum of one group. The distribution of these values over 10^4 replicates (original observation + 9,999 permutations) was used to determine the probability of the observed value. Permutation tests were conducted using a Perl program based on the Fortran program of Manly (1991).

We tested the null hypotheses associated with the analyses of star-topologies, H_0: distribution of $\max(P_{\text{boot}}) =$ theoretical distribution for maximum values and H_0: distribution of $\max[P(\tau \mid D)] =$ theoretical distribution for maximum values, compared with the corresponding two-sided alternative hypotheses, H_1: distribution of $\max(P_{\text{boot}}) \neq$ theoretical distribution for maximum values and H_1: distribution of $\max[P(\tau \mid D)] \neq$ theoretical distribution for maximum values. We used the simulated sets of expected maximum values to determine the 0.0250 and 0.9750 critical values over a distribution of 1,000 observations. Values in the distribution of $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ outside the region defined by the critical values are considered significantly different from theoretical expectations.

Computation

For the part of the study involving the four-taxon model space, the experiment required a relatively large number of phylogenetic analyses: 1,369 model space elements × 1,000 replicates per element × 2 phylogenetic methods = 2.738×10^6 separate phylogenetic analyses, ignoring the individual bootstrap replicates and MCMC steps for each simulated sequence set. At this level of atomization, each analysis in the problem space is wholly independent of any other, and parallelization is an obvious computational model. Similarly, the 1,369 hypothesis tests can also be performed in parallel. Parallel computation has been successfully used in previous large-scale computer-based studies in phylogenetics (Cummings et al., 1995, 1999; Otto et al., 1996). We developed our own

Grid computing system using standard software components (Myers and Cummings, 2003) to distribute the independent analyses across multiple computer systems and geographic locations. The principal part of the simulations took 15.68 CPU years (wallclock) to complete, the comparison of bootstrap procedures required an additional 88.08 CPU days, and the star topology analyses required 5.35 CPU days.

RESULTS

Comparing Bootstrap Procedures

We compared the difference between two methods for the bootstrap procedure: using the estimate of the model parameter values for the original data, and reestimating the parameter values for each individual bootstrap replicate. The absolute difference in $\bar{x}$ values from 1,000 paired analyses and the associated significance based on permutation tests were 0.0095 ($P = 0.4524$) for τ_1, 0.0231 ($P = 0.0070$) for τ_2, and 0.0059 ($P = 0.4949$) for τ_3. Although the $\bar{x}$ values for the two different bootstrap procedures were not significantly different for the topology of primary interest, τ_1, the difference in computation required was substantial. Reestimating the likelihood model parameters for each individual bootstrap replicate required $\bar{x} = 124.94$ times more CPU time. The remaining results presented for P_{boot} are based on the procedure where likelihood model parameters were estimated from the data for the original simulated sequence, D, and the estimated values were then applied in the analysis of the bootstrap samples, D^*.

Comparing Bootstrap and Posterior Probability Values

The mean values of P_{boot} and $P(\tau \mid D)$ at points in the model space are the consequence of a combination of factors, including length of simulated sequence, branch lengths, choice of likelihood model, and underlying characteristics of the analytical procedure. There are three relatively distinct regions in the model space that can be differentiated based on the difference between mean bootstrap and mean posterior probability values, $d = \bar{x}(P_{\text{boot}}) - \bar{x}[P(\tau_1 \mid D)]$ (Fig. 3). First, there is a broad region encompassing most of the model space from lower left corner toward the upper right corner, referred to as the neutral zone (region A in Fig. 1). The neutral zone, as its name implies, is mixed with respect to which measure has higher or lower values. The second region is adjacent to the neutral zone, closer to the two-branch corner (upper left), referred to as the near two-branch corner (region B in Fig. 1). The third area is the two-branch corner (upper left; region C in Fig. 1), where the effects of long-branch attraction become evident in phylogenetic analyses. The size of the two-branch corner is smaller in our study than that observed by Gaut and Lewis (1995), because we used both longer sequences and a more parameter-rich likelihood model. However, in all features of our results are fully consistent with those of Gaut and Lewis (1995).

The significance of d for each element in the model space is given in the results of the permutation tests (Fig. 4). As might be expected, d is not significant for τ_1 in the neutral zone. However, in the near two-branch corner (region B in Fig. 1) $\bar{x}[P(\tau_1 \mid D)]$ is significantly greater than $\bar{x}(P_{\text{boot}})$, but in the two-branch corner (region C in Fig. 1) the opposite pattern is observed with $\bar{x}(P_{\text{boot}})$ significantly greater than $\bar{x}[P(\tau_1 \mid D)]$.

Over the three possible topologies, both P_{boot} and $P(\tau \mid D)$ each sum to 1. Thus, the difference between the means of the two measures, d, for τ_1 is reflected in the value of d for the alternative topologies, τ_2 and τ_3. Examining $\bar{x}(P_{\text{boot}})$ and $\bar{x}[P(\tau \mid D)]$ for the other two topologies leads to further understanding of the differences observed for τ_1. For the attractive topology τ_2,

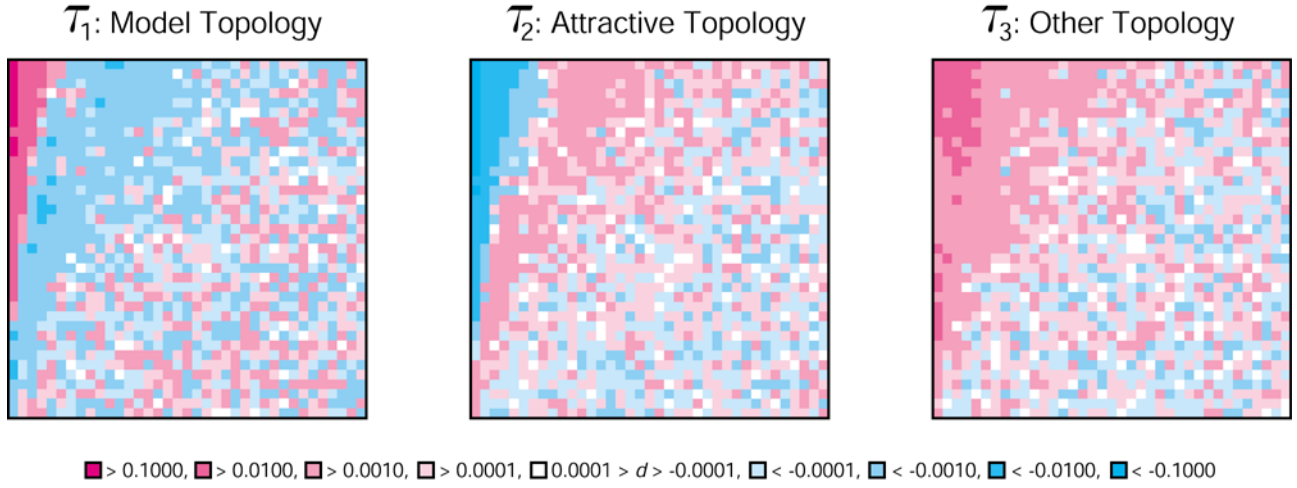


FIGURE 3. Difference, d, between the mean proportion of bootstrap replicates $\bar{x}(P_{\text{boot}})$ and the mean posterior probability values supporting the internal partition $\bar{x}[P(\tau \mid D)]$ for each element in the simulation model space (Fig. 1). Scale for the magnitude of $d = \bar{x}(P_{\text{boot}}) - \bar{x}[P(\tau \mid D)]$ is presented below plots with shades of magenta depicting those cases where $\bar{x}(P_{\text{boot}}) > \bar{x}[P(\tau \mid D)]$ and shades of cyan depicting those cases where $\bar{x}(P_{\text{boot}}) < \bar{x}[P(\tau \mid D)]$.

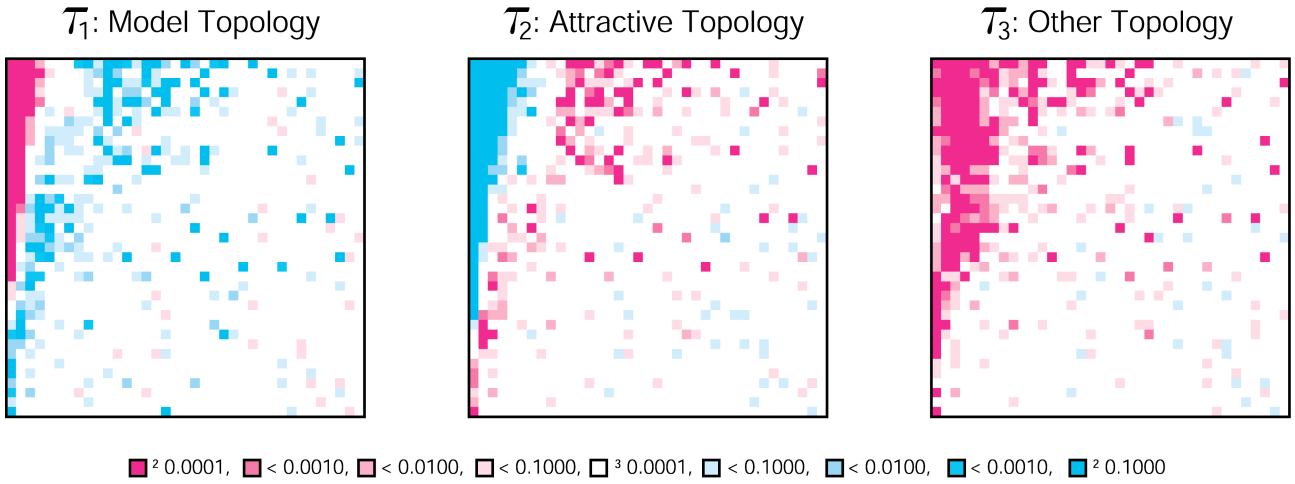


FIGURE 4. Results of permutation tests of H_0: $E(P_{\text{boot}}) = E[P(\tau \mid D)]$ over the simulation model space (Fig. 1). Scale for the significance values of results (uncorrected for multiple tests) is presented below plots, with shades of magenta depicting those cases where the magnitude of $\bar{x}(P_{\text{boot}}) - \bar{x}[P(\tau \mid D)] > 0$ and shades of cyan depicting those cases where $\bar{x}(P_{\text{boot}}) - \bar{x}[P(\tau \mid D)] < 0$.

the neutral zone is again mixed with respect to which measure has higher or lower values (Fig. 3). However, the values of $d = \bar{x}(P_{\text{boot}}) - \bar{x}[P(\tau_1 \mid D)]$ in the other regions are the opposite of that observed for τ_1. In the near two-branch corner region (region B in Fig. 1), $\bar{x}(P_{\text{boot}}) > \bar{x}[P(\tau_2 \mid D)]$, and deep in the two-branch corner $\bar{x}(P_{\text{boot}}) < \bar{x}[P(\tau_2 \mid D)]$. The results of the permutation tests show that few differences between $\bar{x}(P_{\text{boot}})$ and $\bar{x}[P(\tau \mid D)]$ are significantly different in the neutral zone (Fig. 4). In the near two-branch corner, $\bar{x}(P_{\text{boot}})$ is often significantly greater than $\bar{x}[P(\tau_2 \mid D)]$, and deep in the two-branch corner $\bar{x}[P(\tau_2 \mid D)]$ is significantly greater than $\bar{x}(P_{\text{boot}})$.

For the other topology, τ_3, the neutral zone is again mixed with respect to which measure has higher or lower values (Fig. 3). For much of both two-branch corner regions (regions B and C in Fig. 1), $\bar{x}(P_{\text{boot}}) > \bar{x}[P(\tau_2 \mid D)]$, and many of these differences are significant; ($P < 0.0010$) (Fig. 4).

Star Topology

For the null model of a star topology, the values for $\bar{x}(P_{\text{boot}})$ and $\bar{x}[P(\tau \mid D)]$ for each topology were very close to the expected value of 0.3333 (data not shown). The means of the maximum values among all three topologies across replicates were also close to theoretical expectations as determined by simulation, $\bar{x} = 0.6111, n = 10^7$: $\max(P_{\text{boot}})$ $\bar{x} = 0.6348$, $s^2 = 0.0214$; $\max[P(\tau \mid D)]$ $\bar{x} = 0.6272$, $s^2 = 0.0277$. However, means and variances provide for only limited comparison. Therefore, we generated quantile–quantile plots to compare the distributions (Fig. 5). Quantiles are like percentiles except they range from 0 to 1 rather than 0 to 100. In the quantile–quantile plots presented in Figure 5, the quantiles of the theoretical distribution as determined by simulation are compared with the corresponding quantile of the measure of interest, $\max(P_{\text{boot}})$ or $\max[P(\tau \mid D)]$. In this way the two distributions can be compared in their entirety. Both $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ differ significantly from theoretical expectations over much of their distributions. These differences are the specific focus of the plots in Figure 6, where the difference between theoretical expectation and the measure of interest, $\max(P_{\text{boot}})$ or $\max[P(\tau \mid D)]$, is plotted against the theoretical expectation. In such a plot, identical distributions would be represented by a straight line parallel to the abscissa with a value of 0 difference. Values for $\max(P_{\text{boot}})$ are significantly greater than the theoretical expectation for most of the distribution, particularly in the vicinity of the median, the 0.50 quantile, $\max(P_{\text{boot}}) = 0.6365$, and again in some later quantiles (Figs. 5, 6). The departures from theoretical expectations are particularly pronounced for $\max[P(\tau \mid D)]$ from the 0.456 quantile, $\max[P(\tau \mid D)] = 0.5865$, and above, which are significantly greater than theoretical expectations with the exception of the very highest quantile (Figs. 5, 6). In the interval of most interest, values in the range 0.95–1.0, the distribution of $\max(P_{\text{boot}})$ is much closer to the distribution for the theoretical expectations, with none of the values of $\max(P_{\text{boot}})$ in this interval significantly exceeding theoretical expectations (Fig. 6a). In contrast, the distribution for $\max[P(\tau \mid D)]$ values in this interval are all significantly higher than the theoretical expectation, with the exception of the last quantile, 1.0 (Fig. 6b). To put it another way, the observed proportion of $P_{\text{boot}} > 0.95 = 0.014$, which is close to theoretical expectation as determined by simulation, 0.009, but the proportion of $P(\tau \mid D) > 0.95 = 0.063$, ~ 7.6 times the theoretical expectation.

DISCUSSION

Given sufficient data, the appropriate likelihood model, and the appropriate search strategies, the values of P_{boot} and $P(\tau \mid D)$ are similar. Over most of the model space, the means of the two measures were not significantly different (Fig. 4). However, in the absence of other information it is difficult to draw general

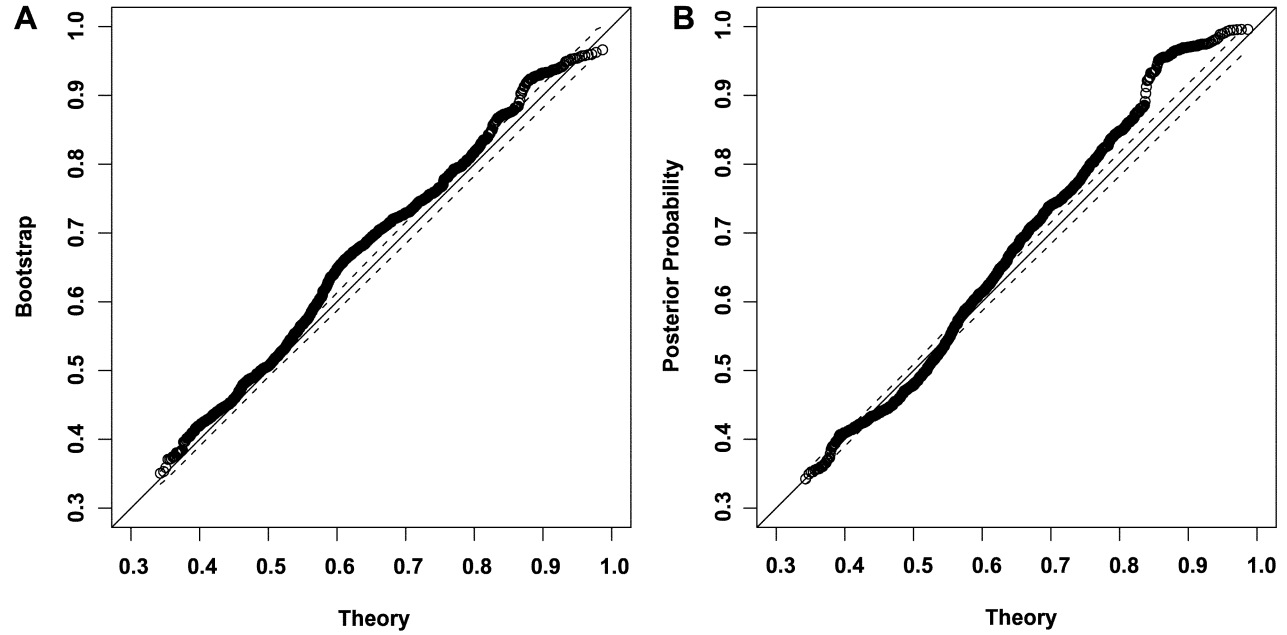


FIGURE 5. Quantile–quantile plots comparing the distributions of the theoretical maximum values with maximum bootstrap values, $\max(P_{\text{boot}})$ (A), and maximum posterior probability values, $\max[P(\tau \mid D)]$ (B), for star topologies. Distributions for $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ are based on 1,000 separate paired analyses. Distribution of theoretical maximum values are based on the means of 10^4 samples, each representing 1,000 ordered values derived by simulation. Solid lines represent the null hypothesis that the two distributions are equal; dashed lines represent the 0.0250 and 0.9750 critical values for the two-sided alternative hypothesis.

conclusions when the two measures differ significantly. We observed significant differences in the two-branch corner of the model space where long-branch attraction affects both measures, albeit differently. In the near two-branch corner (region B in Fig. 1), there is a general trend for $\bar{x}(P_{\text{boot}}) < \bar{x}[P(\tau_1 \mid D)]$, whereas deep in the two-branch corner (region C in Fig. 1), there is a general trend for $\bar{x}(P_{\text{boot}}) > \bar{x}[P(\tau_1 \mid D)]$. The reason why

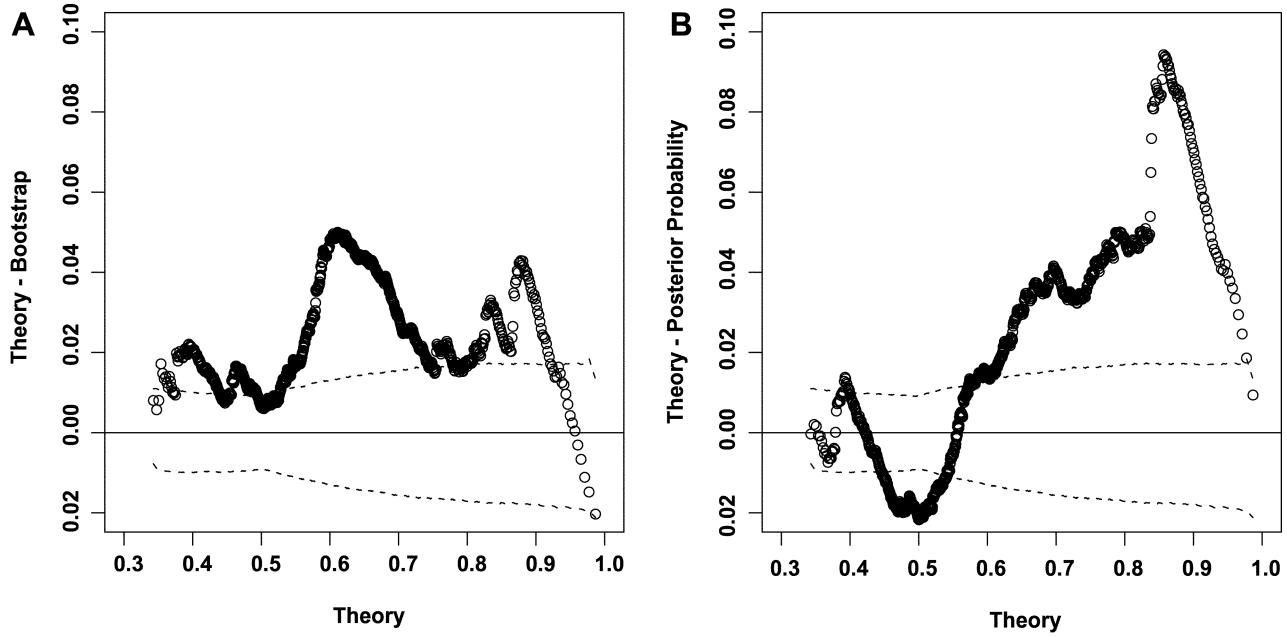


FIGURE 6. Plots depicting the relationship of the distribution of maximum values expected under theory to the difference of maximum values expected under theory − maximum bootstrap values, $\max(P_{\text{boot}})$ (A) and maximum values expected under theory − maximum posterior probability values, $\max[P(\tau \mid D)]$ (B). Straight lines represent the null hypothesis that the two distributions are equal; dashed lines represent the 0.0250 and 0.9750 critical values for the two-sided alternative hypothesis.

$\bar{x}(P_{boot}) > \bar{x}[P(\tau_1 \mid D)]$ deep in the two-branch corner is because the branch-and-bound search with the maximum likelihood optimality criterion always found the topology with the highest likelihood, which was most often the model topology. The MCMC-guided journey through parameter space in the Bayesian analysis often visited the attractive topology in addition to the model topology, which led to higher values for $P(\tau_2 \mid D)$ and correspondingly lower values for $P(\tau_1 \mid D)$.

The relationship of $\bar{x}(P_{boot})$ to $\bar{x}[P(\tau_1 \mid D)]$ when plotted with $\bar{x}(P_{boot})$ along the abscissa and $\bar{x}[P(\tau_1 \mid D)]$ along the ordinate follows a sigmoidal curve (Fig. 7). Previous studies, each based on a single empirically derived data set and/or simulation based on an empirically derived tree show a similar relationship between P_{boot} and $P(\tau \mid D)$ (Leaché and Reeder, 2002; Whittingham et al., 2002; Wilcox et al., 2002). However, the points where $P_{boot} = P(\tau \mid D)$ differ across studies. In our study, the change point is in the vicinity of $\bar{x}(P_{boot}) = \bar{x}[P(\tau \mid D)] = 0.93$; for most points $\bar{x}(P_{boot}) > \bar{x}[P(\tau \mid D)]$ below and $\bar{x}(P_{boot}) < \bar{x}[P(\tau \mid D)]$ above this value (Fig. 7b). Among the reasons for the higher change point in our study compared with those of the other studies may be differences with respect to attractive alternative topologies and the amount of information supporting specific relationships. For the star topology null model, a generally similar sigmoidal curve, albeit with a lower change point, is evident in the comparison of the distribution of $\max(P_{boot})$ and the $\max[P(\tau \mid D)]$ (Fig. 8).

Extrapolation from our four-taxon simulation analyses to the more complex phylogenetic analyses of empirical data sets suggests that over most of the parameter space the differences between bootstrap and posterior probability values are expected to be nonsignificant. In contrast, when long branches are separated by a very short internode (two-branch corner of Fig. 1), bootstrap and posterior probability values may differ significantly. However, the extremes in branch length differences in our study may rarely be encountered in empirical data sets. Our results contrast with those of the few empirical analyses, which consistently find $P_{boot} < P(\tau \mid D)$ (Leaché and Reeder, 2002; Whittingham et al., 2002; Wilcox et al., 2002; Douady et al., 2003). Instead, our results show that rather than the values of one metric being consistently greater than those of the other, the values are most often not statistically different or that the direction of difference depends on the branch lengths involved. However, our results based on analyses of star topologies suggest that the consistent observations of $P_{boot} < P(\tau \mid D)$ in studies of empirical data may largely be the result of $P(\tau \mid D)$ values being excessively high. Although there are advantages of basing studies on empirical data sets, these data sets do not represent the diversity of branch length differences found in our simulated data. Furthermore, our results suggest that extrapolation from these empirical studies are unlikely to provide any general conclusions regarding the relationship between bootstrap and posterior probability values. Simple comparisons between metrics are insufficient for fully assessing absolute performance; such an assessment requires a theoretical standard with which to compare results, as we have done in our analyses of

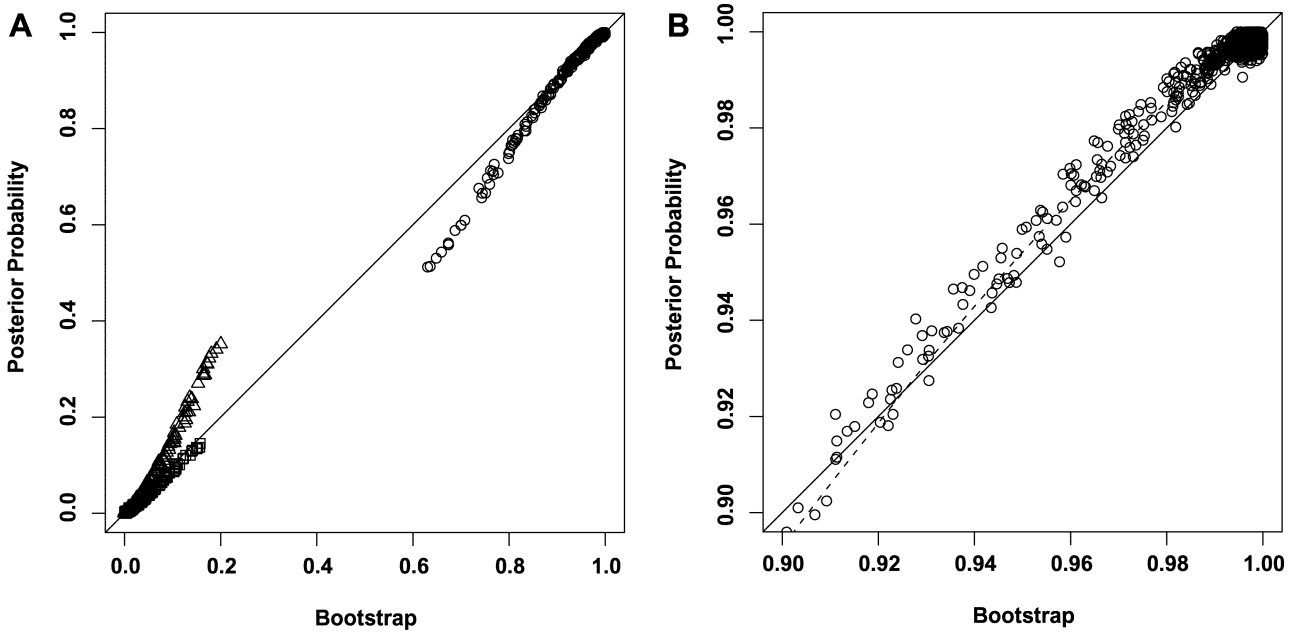


FIGURE 7. Plots depicting the relationship of bootstrap (abscissa) to posterior probability (ordinate) from the four-taxon model space for the entire range of values (A) and for the intervals 0.90–1.0 (B). Points represent paired $\bar{x}(P_{boot})$ and $\bar{x}[P(\tau \mid D)]$ values, $n = 1{,}000$, from each element of the model space (Fig. 1) and each of the three topologies: τ_1, circles; τ_2, triangles; τ_3, squares. Solid lines represent H_0: $E(P_{boot}) = E[P(\tau \mid D)]$. Points below the line depict values where $\bar{x}(P_{boot}) > \bar{x}[P(\tau \mid D)]$; points above the line depict where values $\bar{x}(P_{boot}) < \bar{x}[P(\tau \mid D)]$. The dashed line represents a robust locally weighted regression fit to the points depicted and emphasizes the nonlinear relationship between $\bar{x}(P_{boot})$ and $\bar{x}[P(\tau \mid D)]$.

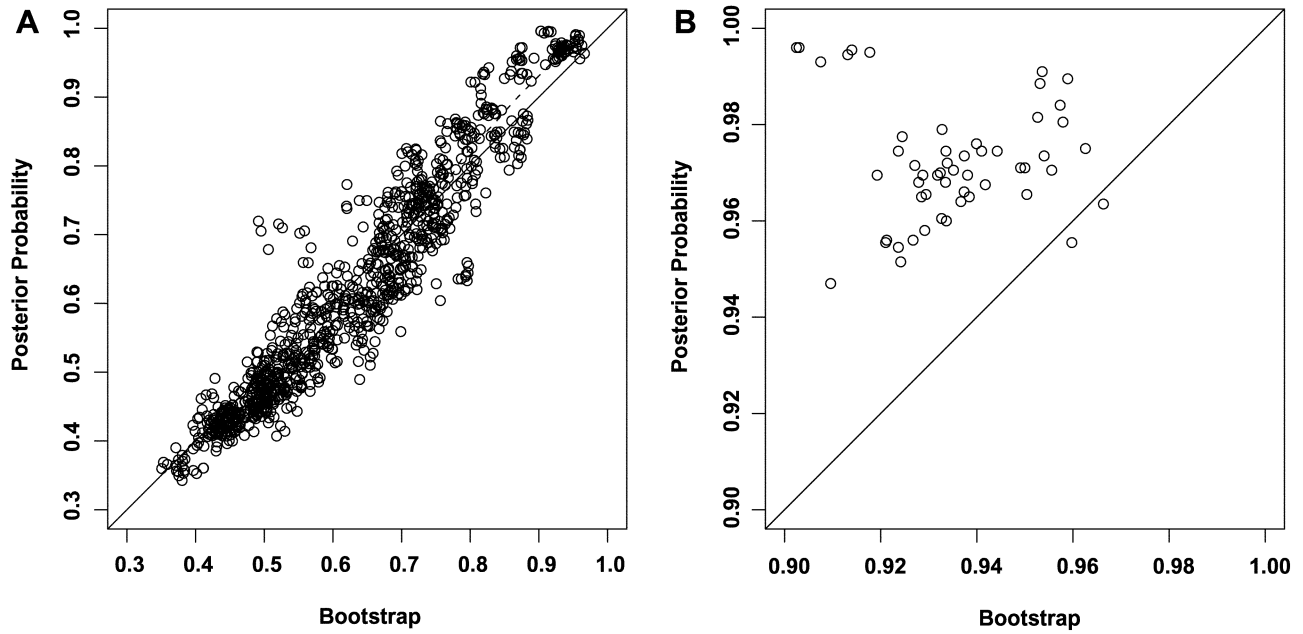


FIGURE 8. Plots depicting the relationship of bootstrap (abscissa) to posterior probability (ordinate) from the star topologies for the entire range of values (A) and for the intervals 0.90–1.0 (B). Points represent paired $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$ values, $n = 1{,}000$. Solid lines represent H_0: $E(P_{\text{boot}}) = E[P(\tau \mid D)]$. Points below the line depict values where $\bar{x}(P_{\text{boot}}) > \bar{x}[P(\tau \mid D)]$; points above the line depict where values $\bar{x}(P_{\text{boot}}) < \bar{x}[P(\tau \mid D)]$. The dashed line represents a robust locally weighted regression fit to the points depicted and emphasizes the nonlinear relationship between $\max(P_{\text{boot}})$ and $\max[P(\tau \mid D)]$.

star topologies. Clearly, more empirical and simulation studies are needed to fully understand the relationship between bootstrap and posterior probabilities in phylogenetic analyses and the relationships of both measures to theoretical expectations.

Regarding extrapolation from studies of trees of four taxa to analyses of more taxa, it is important to realize that four taxa represent the smallest number of taxa that provide for a nontrivial tree (i.e., a tree with more than one possible unrooted topology). The utility of the four-taxon case in exploring problems in phylogenetic analysis is a consequence of this nontriviality, the simplicity of exploring specific analytical problems, and the ability to do so in a way that is both easy to convey and to comprehend with two-dimensional representations. Felsenstein (1978) and others (e.g., Huelsenbeck and Hillis, 1993; Gaut and Lewis, 1995; Suzuki et al., 2002) recognized these properties of the four-taxon case in problems involving differences in relative branch lengths. Trees with more than four taxa add only additional partitions to the single partition of the four-taxon case. Therefore, although most often more than four taxa are considered in phylogenetic studies, there is no property of the four-taxon case that is not, by extension, applicable to studies of more than four taxa. However, the problems of increased analytical complexity and decreased representational effectiveness make the use of more than four taxa in studies such as that presented here both superfluous and obfuscating. In other words, everything demonstrated for a tree of four taxa applies to trees of more taxa, but the increased tree space is much more difficult to analyze and visualize effectively.

Suzuki et al. (2002) addressed the relationship between bootstrap and posterior probability values using simulated data sets for four taxa. Part of their experimental design involved simulating sequences for a star topology under a Kimura model (Kimura, 1980) and analyzing the data using a simpler Jukes–Cantor model (Jukes and Cantor, 1969). From analyses of star topologies, Suzuki et al. showed that posterior probability values were excessively high compared with bootstrap values based on neighbor joining and maximum likelihood. However, the experimental design used confounded the effects of analyzing data with less well-fit models with the effects attributable to the general properties of the underlying analytical methods. We separated these two confounded factors by using an experimental design where the models used to simulate and analyze the data were very similar, thus focusing attention on properties of P_{boot} and $P(\tau \mid D)$ values. Such a design is in keeping with the best current practice procedure where the fit of models is explicitly assessed and the best-fit model is applied to subsequent analyses. Although maximum likelihood phylogenetic analysis is at least moderately robust to model assumptions (e.g., Fukami-Kobayashi and Tateno, 1991; Kuhner and Felsenstein, 1994; Yang et al., 1994; Gaut and Lewis, 1995), both simulation studies (Yang et al., 1994; Gaut and Lewis, 1995) and studies based on empirical data (Buckley and Cunningham, 2002) demonstrate the advantages of using models that better fit the data. In the present context, the results of Buckley and Cunningham (2002) are particularly relevant because they demonstrate improved bootstrap results when more realistic models are used. The results from our star topology

evaluations demonstrate that max[$P(\tau \mid D)$] departs significantly from theoretical expectations, particularly in the interval 0.85–1. Therefore, our results provide additional evidence consistent with the results of Suzuki et al. (2002), which show that $P(\tau \mid D)$ values appear to be excessively high.

It appears that the exact relationship between P_{boot} and $P(\tau \mid D)$ for any data set is complex and involves the nature of the underlying tree space, the fit of the likelihood model to the data, and the details of how the bootstrap/maximum likelihood search is conducted and the MCMC/Bayesian analysis is performed.

ACKNOWLEDGMENTS

This project was conceived and initiated during preparation for the 2001 Workshop on Molecular Evolution (Marine Biological Laboratory, Woods Hole). We thank our workshop colleagues Josephine Babin, Sheri Church, and Ellen Pritham for discussions at the beginning of the project. We also thank our friends and colleagues for providing computing resources used in this project: Robert Campbell, Monica Riley, Laura Shulman, and Mitchell Sogin (Marine Biological Laboratory); Kenneth Halanych and Arthur Newhall (Woods Hole Oceanographic Institution); Dale Clayton and the Center for High Performance Computing, (University of Utah); the Computer Science Clinic Program (Harvey Mudd College); James Marshall (Pomona College); David Danziger (Oberlin College); Michael Fister (Intel Corporation); Joe Felsenstein and Mary Kuhner (University of Washington); Ernest Retzel, Kevin Silverstein, and the Computational Biology Centers (University of Minnesota); Lawrence Leung (University of California, Berkeley); and Aaron Grogan. Andrew Solow provided some statistical advice, Masatoshi Nei answered questions regarding Suzuki et al. (2002), and Bruce Rannala, Joanna Mountain, and an anonymous reviewer provided comments on our manuscript. A.R. is funded by a Human Frontier Science Program Long Term Fellowship, and D.R. is funded by a National Science Foundation Postdoctoral Fellowship. Partial funding for the Workshop on Molecular Evolution was provided by NSF grant DEB-9980563 and NASA grant NAG5-9415.

REFERENCES

ALFARO, M. E., S. ZOLLER, AND F. LUTZONI. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

BROOKS, S. P., AND A. GELMAN. 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7:434–455.

BUCKLEY, T. R., P. ARENSBURGER, C. SIMON, AND G. K. CHAMBERS. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.

BUCKLEY, T. R., AND C. W. CUNNINGHAM. 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. Evol. 19:394–405.

CUMMINGS, M. P., S. P. OTTO, AND J. WAKELEY. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12:814–822.

CUMMINGS, M. P., S. P. OTTO, AND J. WAKELEY. 1999. Genes and other samples of DNA sequence data for phylogenetic inference. Biol. Bull. 196:345–350.

DAVID, H. A. 1981. Order statistics, 2nd edition. John Wiley & Sons, New York.

DOUADY, C. J., F. DELSUC, Y. BOUCHER, W. F. DOOLITTLE, AND E. J. P. DOUZERY. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

EFRON, B. 1979. Bootstrapping 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 93:7085–7090.

EFRON, B., AND R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman & Hall, New York.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

FELSENSTEIN, J. 1985. Confidence intervals 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.

FISHER, R. A. 1929. Tests of significance in harmonic analysis. Proc. R. Soc. Lond. A 125:54–59.

FUKAMI-KOBAYASHI, K., AND Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. J. Mol. Evol. 32:79–91.

GAUT, B. S., AND P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

GELMAN, A., AND D. B. RUBIN. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–511.

GOOD, P. 1994. Permutation tests, a practical guide to resampling methods for testing hypotheses. Springer-Verlag, New York.

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.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

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. BOLLPACK. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

IHAKA, R., AND R. GENTLEMAN. 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5:299–314.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

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.

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

MANLY, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall, London.

MARITZ, J. S. 1995. Distribution-free statistical methods, 2nd edition. Monographs on Statistics and Applied Probability 17. Chapman & Hall, London.

MAU, B., M. NEWTON, AND B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

MYERS, D. S., AND M. P. CUMMINGS. 2003. Necessity is the mother of invention: A simple grid computing system using commodity tools. J. Parallel Distr. Com. 63:578–589.

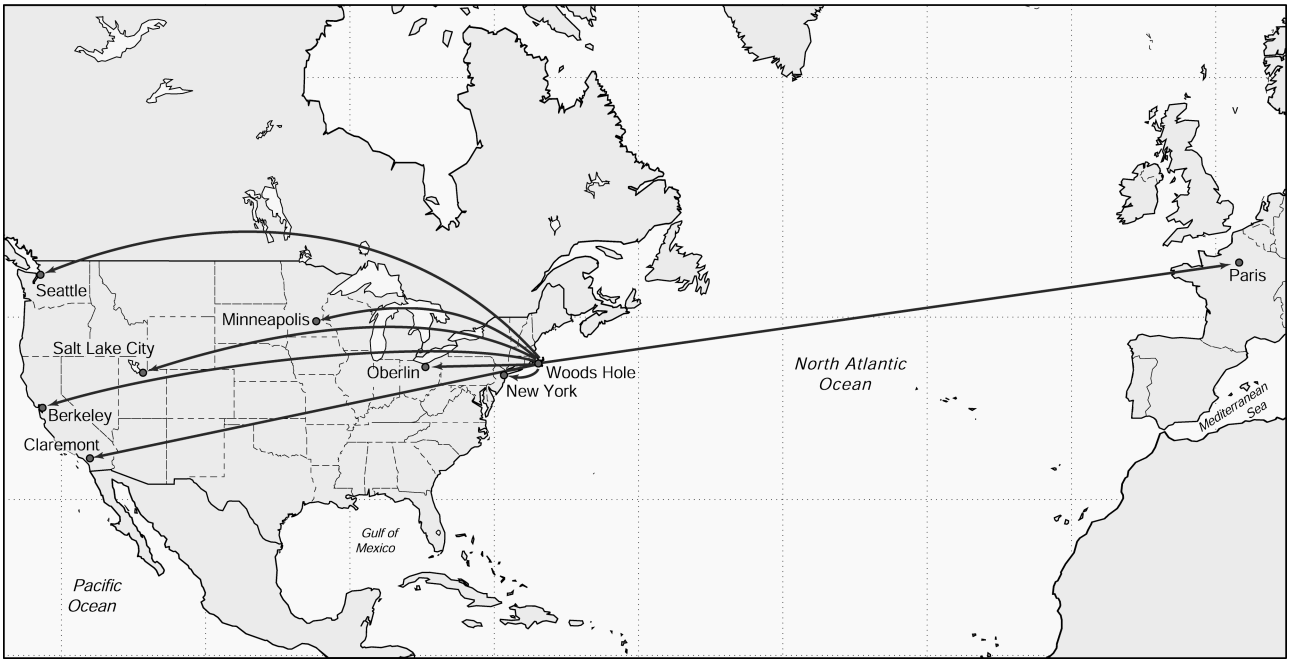
NEWTON, M., B. MAU, AND B. LARGET. 1999. Markov chain Monte Carlo for Bayesian analysis of evolutionary trees from aligned molecular sequences. IMS Lecture Notes—Monograph Series 33:143–162.

OTTO, S. P., M. P. CUMMINGS, AND J. WAKELEY. 1996. Inferring phylogenies from DNA sequence data: The effects of sampling. Pages 103–115 *in* New uses for new phylogenies (P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, eds.). Oxford Univ. Press, New York.

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. 1995. Objections to bootstrapping phylogenies: A critique. Syst. Biol. 44:299–320.
SMITH, B. J. 2001. Bayesian output analysis program (BOA), version 1.0.0. Program distributed by author, http://www.public-health.uiowa.edu/boa/.
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 4.0. Sinauer, Sunderland, Massachusetts.
TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37:613–623.
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.
WILCOX, T. P., D. J. ZWICKL, T. A. HEATH, AND D. M. HILLIS. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.
YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1369–1401.
YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.
YANG, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 15:555–556.
YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.
YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

First submitted 23 August 2002; reviews returned 10 November 2002; final acceptance 15 March 2003
Associate Editor: Bruce Rannala



A map of a portion of the northern hemisphere showing the general geographic distribution of server and clients composing the Grid system (Myers and Cummings, 2003) used in the study. The Grid system spanned multiple continents and administrative domains, and included 163 unique clients. At times, computer programs for the project ran on over 135 processors simultaneously. Each client executed Perl code and architecture-specific binaries. Communication between clients and server used XML-RPC (eXtensible Markup Language-Remote Procedure Call) via TCP/IP (Transmission Control Protocol/Internet Protocol). Analyses were coordinated through a relational database management system abstracted behind a Java interface on a server at the Marine Biological Laboratory in Woods Hole, Massachusetts.