How Meaningful Are Bayesian Support Values?

Mark P. Simmons, Kurt M. Pickett,†[1] and Masaki Miya‡*

*Department of Biology, Colorado State University, Fort Collins; †Department of Entomology, The Ohio State University, Columbus; and ‡Department of Zoology, Natural History Museum and Institute, Chiba, Japan

In this study, we used an empirical example based on 100 mitochondrial genomes from higher teleost fishes to compare the accuracy of parsimony-based jackknife values with Bayesian support values. Phylogenetic analyses of 366 partitions, using differential taxon and character sampling from the entire data matrix of 100 taxa and 7,990 characters, were performed for both phylogenetic methods. The tree topology and branch-support values from each partition were compared with the tree inferred from all taxa and characters. Using this approach, we quantified the accuracy of the branch-support values assigned by the jackknife and Bayesian methods, with respect to each of 15 basal clades. In comparing the jackknife and Bayesian methods, we found that (1) both measures of support differ significantly from an ideal support index; (2) the jackknife underestimated support values; (3) the Bayesian method consistently overestimated support; (4) the magnitude by which Bayesian values overestimate support exceeds the magnitude by which the jackknife underestimates support; and (5) both methods performed poorly when taxon sampling was increased and character sampling was not increases. These results indicate that (1) the higher Bayesian support values are inappropriate (in magnitude), and (2) Bayesian support values should *not* be interpreted as probabilities that clades are correctly resolved. We advocate the continued use of the relatively conservative bootstrap and jackknife approaches to estimating branch support rather than the more extreme overestimates provided by the Markov Chain Monte Carlo–based Bayesian methods.

Introduction

Bayesian analyses have recently been proposed as a new method for phylogenetic inference (Rannala and Yang 1996; Yang and Rannala 1997; Larget and Simon 1999; Mau, Newton, and Larget 1999; Huelsenbeck et al. 2001). In maximum-likelihood analyses (Felsenstein 1973), the most likely trees are searched for. Branch support may then be assessed secondarily using the bootstrap (Felsenstein 1985) or jackknife (Farris et al. 1996). In contrast, Bayesian analyses estimate the posterior probability of each clade based on the frequency at which that clade is resolved among sampled trees once stationary log-likelihood values have been reached.

The bootstrap has been variously interpreted as a means of assessing confidence levels (Felsenstein 1985), a conservative measure of accuracy (Zharkikh and Li 1992; Hillis and Bull 1993), a ''test of monophyly'' (Rodrigo 1993, p. 508), or simply as a means of assessing evidential support (Sanderson 1995). Efron, Halloran, and Holmes (1996) noted that, in a Bayesian framework, bootstrap confidence levels may be comparable to posterior probabilities, given a uniform prior (i.e., the prior probabilities for the values of interest are equal). Hillis and Bull (1993) described bootstrap values as downwardly biased estimates of accuracy (see also Felsenstein and Kishino [1993]). Efron, Halloran, and Holmes (1996) noted that the bootstrap estimate may be higher or lower than expected, depending on which direction the boundary between alternative trees is curved in tree-space. Iterated bootstrapping has been proposed to accommodate the curved boundary between alternative trees in tree-space so that bootstrap values may better reflect type 1 error (Rodrigo 1993; Zharkikh and Li 1995; Efron, Halloran, and Holmes 1996). Alternatively, Berry and Gascuel (1996) proposed utilizing Robinson and Fould's (1981) distance to reduce error.

Huelsenbeck and co-authors have stated that ''The posterior probability of a tree can be interpreted as the probability that the tree is correct'' (Huelsenbeck et al. 2001, p. 2310) and ''... the numbers on the branches of the tree now represent the posterior probability that the clade is true'' (Huelsenbeck et al. 2002, p. 675). Worded so broadly, these statements are misleading, as posterior probability assessments are all updates of a prior view, conditional on the data, the model, and the information content asserted in that prior view. So, although it is appropriate to consider a posterior probability the probability of truth given these conditions, it is not appropriate to consider them universal probabilities of truth. As such, we will refer to ''phylogenetic accuracy'' or the ''probability that a clade is correctly resolved'' to mean correct in this universal sense and reserve the term posterior probability to refer to probability conditional on the data, the model, and the prior.

Bayesian support values have been found to be consistently higher than bootstrap values for the same clades in empirical analyses (e.g., Rannala and Yang 1996; Leaché and Reeder 2002; Whittingham et al. 2002). Two recent papers comparing Bayesian posterior probabilities with bootstrap values have presented conflicting results based on simulations. Wilcox et al. (2002, p. 369) concluded that:

> Our simulation analysis indicates that these higher levels of [Bayesian] support [values] are appropriate, and that Bayesian support values provide much closer estimates of phylogenetic accuracy (even though they are still somewhat conservative) than the estimates provided by corresponding bootstrap proportions. Therefore, we recommend that when available, Bayesian posterior probabilities should be used in preference to bootstrap proportions to assess support for estimated clades in phylogenetic trees.

In contrast, Suzuki, Glazko, and Nei (2002, p. 16140) ''demonstrated that the posterior probability in Bayesian

[1] Present address: Division of Invertebrates, American Museum of Natural History, New York, NY.

Key words: branch support, bootstrap support, jackknife support, posterior probabilities, ayesian phylogenetic inference, parsimony.

E-mail: psimmons@lamar.colostate.edu.

Mol. Biol. Evol. 21(1):188–199. 2004
DOI: 10.1093/molbev/msh014
Molecular Biology and Evolution vol. 21 no. 1

phylogenetics can be excessively high in the analysis of concatenated sequences even when the same model as that for generating each gene sequence was used.'' Suzuki, Glazko, and Nei (2002) questioned the reliability of phylogenies inferred using Bayesian methods and asserted that their simulations were more realistic than those used by Wilcox et al. (2002), in which the same model used to generate sequences was used for phylogenetic inference.

Like more common resampling-based support measures, such as the bootstrap and the jackknife, Bayesian clade support has been interpreted as an indicator of phylogenetic accuracy. Although several studies have compared the accuracy of Bayesian clade support values with the bootstrap, no study as yet has compared Bayesian support with jackknife support. In this study, we used an empirical example based on 100 mitochondrial genomes from higher teleost fishes to determine whether Bayesian clade support is a better indicator of accuracy than the jackknife. Phylogenetic analyses of 366 partitions, using differential taxon and character sampling from the entire data matrix of 100 taxa and 7,990 characters, were performed for both phylogenetic methods. The tree topology and branch-support values from each partition were compared with the tree inferred from all taxa and characters. Using this approach, we quantified the reliability of the branch-support values assigned by the jackknife and Bayesian methods, with respect to each of 15 basal clades. Our results indicate that (1) both measures of support differ significantly from an ideal support index; (2) the jackknife underestimated support values; (3) the Bayesian method consistently overestimated support; (4) the magnitude by which Bayesian values overestimate support exceeds the magnitude by which the jackknife underestimates support; and (5) both methods performed poorly when taxon sampling was increased but character sampling was not.

Materials and Methods

Data Matrix

Miya et al.'s (2003) data set based on 100 mitochondrial genomes from higher teleost fishes was selected for this study because of its extensive taxon and character sampling and because many basal clades were well supported. For the purposes of this study, well-supported clades were defined as those with both 63% or greater jackknife support and posterior probability for the parsimony and Bayesian analyses, respectively. (The lowest support values for the clades examined here were 68% jackknife support and 98% Bayesian support.) Sixty-three per cent jackknife support corresponds to the expected jackknife frequency for a clade supported by a single uncontradicted synapomorphy (Farris et al. 1996). Jackknife values, when the removal probability is set to e^{-1}, as performed here, are equivalent to bootstrap values when parsimony-uninformative characters are eliminated before calculating bootstrap support and an infinite number of bootstrap replicates are performed (Farris et al. 1996). Therefore, theoretically, the jackknife should be a superior measure of support with more desirable statistical properties than the bootstrap. Nevertheless, a general correspondence between jackknife and bootstrap values has been demonstrated by Mort et al. (2000) and Salamin et al. (2003). Also, Harshman (1994, p. 421) demonstrated that the inclusion of uninformative characters results in only minor decreases in bootstrap-support values, ''and can generally be ignored.'' Therefore, although we examine the jackknife, we expect that our results can be generalized to the bootstrap as well.

Trees were rooted with *Sardinops melanostictus*, following Miya, Kawaguchi, and Nishida (2001) and Miya et al. (2003), based on Inoue et al.'s (2001) analysis of the basal teleost fishes. Basal clades, each consisting of multiple terminals, were examined in this study so that differential taxon sampling could be used among the runs from the second and third sampling strategies (see below). For the purposes of this study, basal clades were arbitrarily defined as those clades consisting of eight or more taxa in the trees inferred from the complete matrix. Based on these definitions, there are 16 well-supported basal clades, of which 15 are indicated in figures 1 and 2 (also available as Supplementary Material online at http://www.molbiolevol.org/). The 16th clade was excluded because it was not applicable (i.e., could not be nontrivially resolved because it only included one terminal) in 22 of the 47 runs for each set (see below).

Sampling Strategies

Three alternative sampling strategies were performed. Strategy 1 was based on increased character sampling without any increase in taxon sampling. In contrast, strategy 2 entailed increased taxon sampling without any increase in character sampling. In strategy 3, effort was split between increased taxon and character sampling. The numbers of taxa and characters sampled for each run within each strategy are shown in table 1. Runs were terminated for strategies 1 and 2 when all taxa or characters, respectively, had been sampled. This resulted in a partial final run for strategy 2 (run 5.55). Each whole-numbered run consists of a multiple of 14,382 nucleotides (e.g., run 10, in which the second and third strategies each consist of 143,820 nucleotides).

The strategies 1 and 2 represent the two extremes of how data may be sampled. There are a multitude of ways to split one's effort between adding both taxa and characters. The approach that was taken for strategy 3 was to sample taxa and characters in the ratio used for the complete matrix of 100 taxa and 7,990 characters by Miya et al. (2003), based on the equation $y = mx + b$ (wherein $m = 87.695$ and $b = -779.51$) and rounding to whole numbers. The 47 Bayesian and parsimony runs for each set were composed of the initial run of 18 taxa and 799 characters, nine runs for strategy 1, nine runs for strategy 2 (using the ''A'' and ''B'' taxon-sampling approaches [see below]), and 28 runs for strategy 3 (using the ''A'' and ''B'' taxon-sampling approaches).

Character and Taxon Sampling

Miya et al.'s (2003) character sampling, including their exclusion of third codon positions, was followed. The

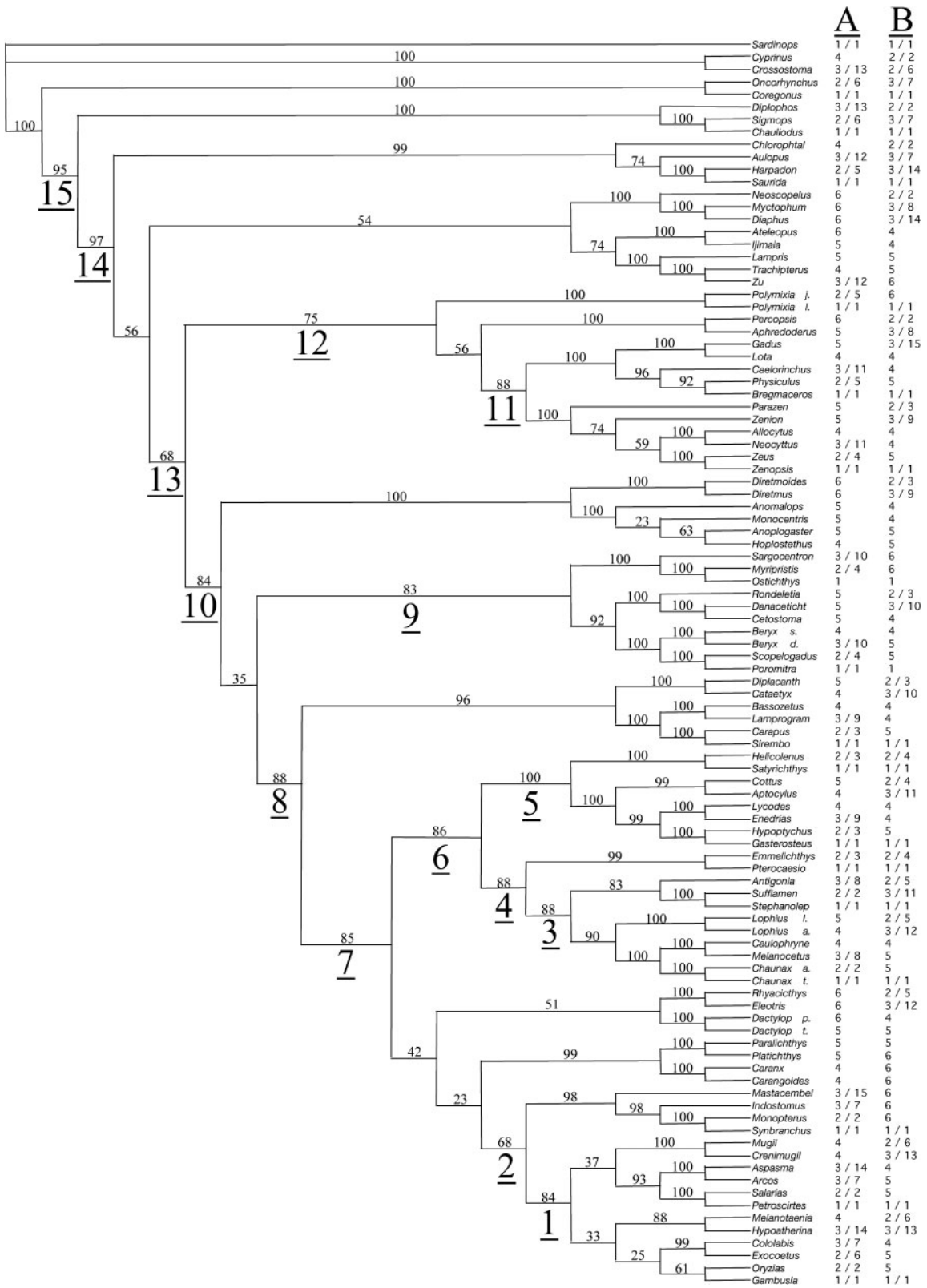

FIG. 1.—Single most-parsimonious tree of 45,230 steps (CI = 0.16; RI = 0.37) for the complete matrix of 100 taxa and 7,990 nucleotide characters. Jackknife support values are listed above branches; the 15 well-supported basal clades are listed below branches. The order of taxon addition for sampling strategies 1 and 2 (table 1) is indicated to the right for both the "A" and "B" sampling approaches. For example, "3/11" indicates that the taxon was added to the third run for strategy 2, and the 11th run for strategy 3.

Table 1
Number of Taxa and Characters Sampled for Each of the Three Sampling Strategies from the 15 Runs

	Strategy 1		Strategy 2		Strategy 3	
Run	Taxa	Characters	Taxa	Characters	Taxa	Characters
1	18	799	18	799	18	799
2	18	1,598	36	799	23	1,251
3	18	2,397	54	799	27	1,598
4	18	3,196	72	799	30	1,918
5	18	3,995	90	799	33	2,179
5.55	—	—	100	799	—	—
6	18	4,794	—	—	36	2,397
7	18	5,593	—	—	39	2,581
8	18	6,392	—	—	41	2,806
9	18	7,191	—	—	43	3,010
10	18	7,990	—	—	45	3,196
11	—	—	—	—	47	3,366
12	—	—	—	—	49	3,522
13	—	—	—	—	51	3,666
14	—	—	—	—	53	3,799
15	—	—	—	—	54	3,995

data matrix includes characters from first and second codon positions of 12 protein-coding genes and 21 tRNA genes. The two rDNA genes were excluded because unambiguous alignment based on secondary structure models was not possible (Miya and Nishida 2000), and ND6 was excluded following Miya, Kawaguchi, and Nishida (2001) and Miya et al. (2003). Additionally, 397 nucleotide positions from ambiguously aligned regions were excluded. The final matrix is composed of 7,990 nucleotide characters, 749 of which contain gaps or missing data in one or more taxa. To eliminate locus-specific effects on the results, characters were sampled randomly. This was deemed important because some mitochondrial genes in animals have been described as particularly problematic for phylogenetic inference (e.g., Baker, Wilkinson, and DeSalle 2001; Shevchuk and Allard 2001).

For the initial run, 18 taxa were sampled such that each of the 15 well-supported, basal clades for the complete matrix included two or more taxa. A minimum of two taxa within each clade was necessary for nontrivial resolution of the clade. Within each of the 15 clades, the two taxa were selected such that their inferred most-recent common ancestor is the same as the most-recent common ancestor of the entire clade. This was done so as not to have the results confounded by different clades being examined in the different runs, as described by Sanderson and Wojciechowski (2000).

In this study, two approaches to taxon addition were utilized for the second and third sampling strategies in an attempt to bracket the extremes of how additional taxa may be sampled. The ''A'' approach entailed adding one taxon immediately above each of the initial 18 taxa from the tree based on the complete matrix, as the trees were drawn in figures 1 and 2. Because branches on the trees were rotated such that the sister group that included more terminals was placed on the lower part of the tree, this had the general effect of adding taxa that were inferred to be closely related to those taxa already included in the analysis. The ''B'' approach entailed adding one taxon immediately below each of the initial 18 taxa. This had the general effect of adding taxa that were inferred to be distantly related to those taxa already included in the analysis. For both approaches, taxa were added by successively scrolling up or down the tree, respectively, until the appropriate number of taxa was added.

One-tenth of the total number of characters (799) was sampled for the initial run. This is roughly equivalent to the number of first and second codon positions sampled in many phylogenetic analyses based on single genes. Four independent sets of successive runs were performed for each sampling strategy. Each set was based on the same taxa sampled for a different group of randomly selected characters.

Tree Searches

The parsimony jackknife is a common method of assessing support that corresponds generally to the parsimony bootstrap (see above) but as yet has not been compared with Bayesian clade support. This would seem to make our comparison of jackknife and Bayesian support timely. However, in this study, parsimony jackknife analyses were substituted for parsimony bootstrap analyses primarily for efficiency of analysis because jackknife tree searches require optimization of fewer characters on each of the alternative trees. Maximum-likelihood bootstrap analyses were not performed because of their much higher computational cost (Sanderson and Kim 2000). Although maximum-likelihood analyses are arguably more comparable to Bayesian analyses, thousands of maximum-likelihood bootstrap analyses of data sets with as many as 100 taxa would be computationally prohibitive. Therefore, in this paper, the jackknife serves as a surrogate for the bootstrap. All parsimony-based tree searches were performed using equally weighted parsimony. When searching for the most parsimonious tree(s), 1,000 tree searches were performed with random taxon addition, tree-bisection-reconnection branch swapping, and a maximum of 10 trees held per search using NONA version 2.0 (Goloboff 1993). Jackknife tree searches were performed using WinClada version 1.00.08 (Nixon 2002), running NONA as a daughter process. One thousand jackknife replicates were performed, with each replicate running one search and up to 10 trees held per search.

Bayesian analyses were conducted with MrBayes version 2.01 (Huelsenbeck and Ronquist 2001) using the GTR + I + Γ model. The hierarchical likelihood ratio test (Huelsenbeck and Crandall 1997) and the Akaike information criterion (Akaike 1974), as implemented in Modeltest version 3.06 (Posada and Crandall 1998), selected GTR + I + Γ as the best-fitting model for the complete matrix. The rates for the GTR model and the shape of the gamma distribution (Yang 1993) estimated by Modeltest were specified as starting values for all Bayesian analyses. The GTR rates and the shape of the gamma distribution were not fixed. Nucleotide frequencies were estimated from the data. For each matrix, two independent Bayesian analyses were performed to check for convergence (as in Miller, Buckley, and Manos 2002), with four chains per analysis and trees sampled every 100

generations. Majority-rule consensus trees were calculated using PAUP* (Swofford 1998).

One million generations were performed for each Bayesian analysis of the complete matrix of 100 taxa and 7,990 characters. The analyses reached stationarity before 100,000 and 250,000 generations, respectively. The 16,500 trees sampled at stationarity were used to infer the Bayesian tree with support values. For the Bayesian analyses performed on the different partitions of the complete matrix (as listed in table 1), at least 350,000 generations were performed for each analysis. The first 250,000 generations were discarded as ''burn-in,'' and the remaining 2,000 trees that were sampled at stationarity from the combined analyses were used to produce a 50% majority-rule consensus tree with clade frequencies. When stationarity was not reached for both analyses within the first 250,000 generations, or when the same stationarity was not reached for both analyses, the analyses were rerun using 600,000 or 1,000,000 generations, as necessary.

The GTR + I + Γ model and the same set of starting values as used for the Bayesian analysis of the complete matrix were used for all Bayesian analyses (as listed in table 1). Because the characters were randomized, this approach is appropriate in that there should be no locus-specific effects for any given data matrix. This approach of using the same model and starting values to analyze matrices that include differential taxon sampling has been advocated by Hillis (1999) (but see Posada and Crandall 2001). It is also expected to be an advantage for Bayesian analyses given that the model and starting values were selected using all taxa available (Pollock and Bruno 2000). Finally, even though the model parameters may differ between the complete matrix and any given one of partitioned matrices, Bayesian analyses do not necessarily use the initial parameters once stationarity (and, hence, estimation of the posterior probabilities) has been reached (Rannala and Yang 1996; Huelsenbeck et al. 2002).

Because of topological differences between the most-parsimonious tree and the Bayesian tree (figs. 1 and 2), not all of the same taxa were sampled for any of the 47 different runs (table 1) from each set (except the run in which all 100 taxa were sampled for strategy 2). Therefore, except for the single redundant run, Bayesian analyses were also performed using the parsimony-based taxon sampling, and parsimony analyses were also performed using the Bayesian-based taxon sampling. As such, the results from the parsimony analyses are directly comparable with the results from the Bayesian analyses, and the taxon-sampling methodology used should not be biased in favor of one method over the other. A total of 366 analyses were performed for both the parsimony and the Bayesian methods.

Quantifying Results

Although it would be nice to know the true mitochondrial gene tree for the higher teleost fishes sampled, it is fundamentally impossible to know the true tree for empirical analyses, outside of those based on experimental phylogenies (e.g., Hillis et al. 1992). However, an adequate substitute is the tree that is supported by the most inclusive taxon and character sampling (following Cummings, Otto, and Wakeley 1995) and therefore represents the best-tested hypothesis of relationships (Kluge 1989; Nixon and Carpenter 1996). We term this the ''reference'' tree. Hence, we are testing for internal consistency of the parsimony and Bayesian methods, not correspondence to an unknown true tree. This is the same approach used by Miller (2003), with respect to character sampling, and Lecointre et al. (1993), Poe (1998), and Simmons and Freudenstein (2003), with respect to taxon sampling.

For each of the 366 analyses performed for the parsimony and Bayesian methods, each clade resolved was compared with the 15 well-supported basal clades from the respective reference tree constructed using the complete matrix of 100 taxa and 7,990 characters (figs. 1 and 2). If the clade in question corresponded to one of the 15 well-supported basal clades from the reference tree (with respect to the taxa sampled in the given run), it was scored as ''correctly'' resolved. If the clade in question contradicted any of the 15 well-supported basal clades from the respective reference tree, it was scored as ''incorrectly'' resolved.

An important qualification to note for any study that evaluates branch-support values for multiple clades, whether they be simulation studies (e.g., Hillis and Bull 1993; Wilcox et al. 2002) or empirically based studies, is that branch-support values from different clades are not strictly independent of one another (Faith and Ballard 1994; Gatesy 2000). The only way to bypass this issue of nonindependence would be to limit oneself to studying matrices that completely lack character conflict or those that include only four taxa and thereby have only a single internal branch.

Previous studies have compared the relative performance of the bootstrap to Bayesian posterior values (Wilcox et al. 2002; Suzuki, Glazko, and Nei 2002; Alfaro, Zoller, and Lutzoni 2003; Cummings et al. 2003; Douady et al. 2003). Suzuki, Glazko, and Nei (2002) and Cummings et al. (2003) concluded that Bayesian support values are less accurate than bootstrap values, whereas Wilcox et al. (2002) and Alfaro, Zoller, and Lutzoni (2003) have asserted the opposite, and Douady et al. (2003) concluded that the methods differ in strengths and liabilities at upper and lower bounds of support. In each of these studies, one measure is asserted to perform better than the other either generally or in a given context.

Wilcox et al. (2002, p. 366) stated, ''In our simulations, nonparametric bootstrapping significantly underestimated the probability of recovering a clade for all but the lowest support values, as has been previously reported by several authors (Hillis and Bull 1993; Rodrigo 1993; Zharkikh and Li 1995). In contrast, Bayesian support values provided much closer estimates of the true probabilities of recovering the respective clades . . .'' Here, subjective graphical proximity of a measure's distribution to an idealized line is the criterion for determining significance.

Suzuki, Glazko, and Nei (2002, p. 16138) were more cautious, pointing out only that Bayesian support values are inflated, whereas bootstrap values are conservative,

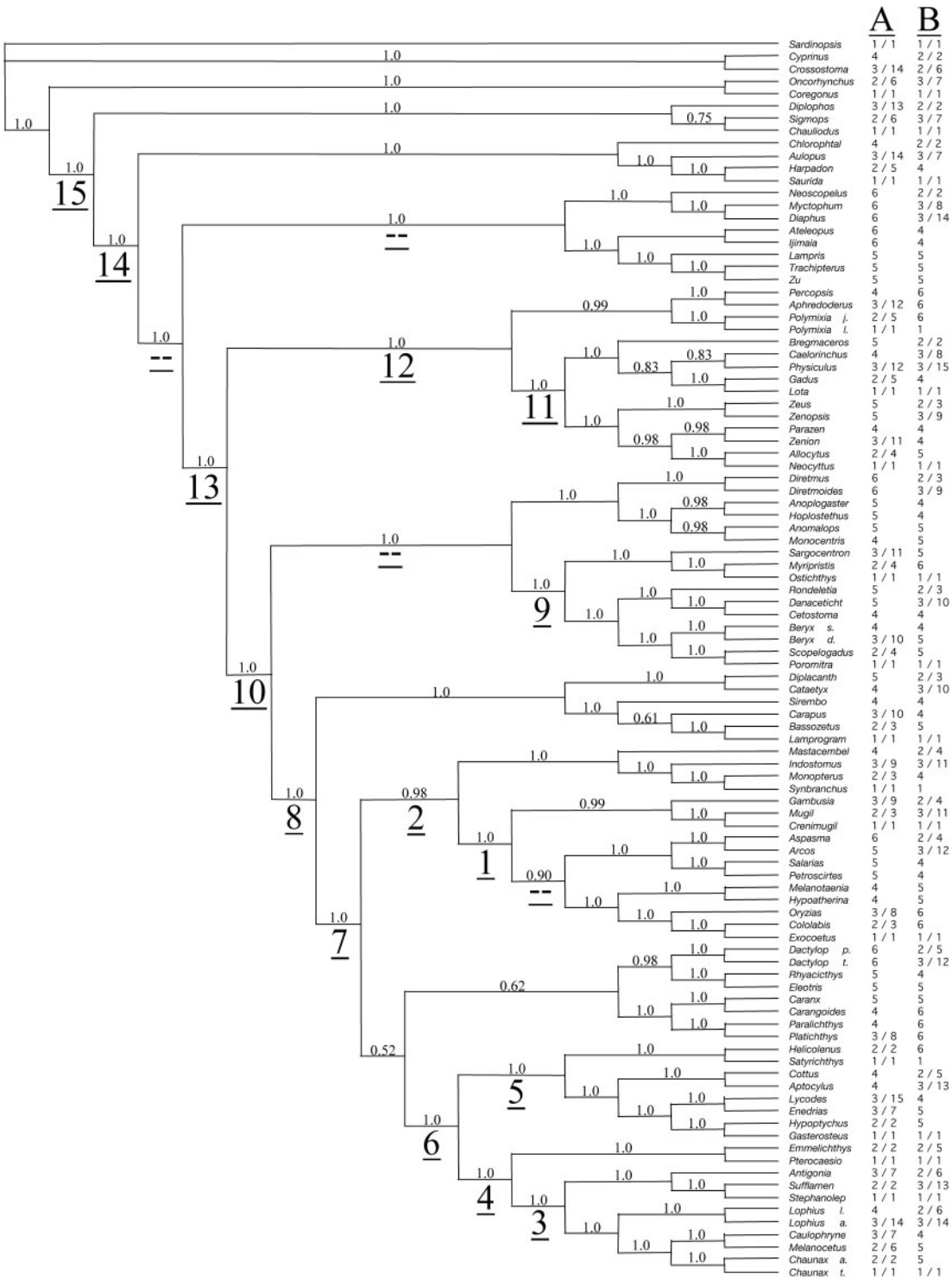

FIG. 2.—The majority-rule consensus tree of Bayesian analyses for the complete matrix of 100 taxa and 7,990 nucleotide characters under the GTR + I + Γ model of DNA substitution. Bayesian support values are listed above branches; the 15 well-supported basal clades are listed below branches. The dashes below four branches indicate those clades that were not resolved or not well supported in the parsimony analysis and, therefore, not considered here. The order of taxon addition for sampling strategies 2 and 3 (table 1) is indicated to the right for both the "A" and "B" sampling approaches. For example, "3 /11" indicates that the taxon was added to the third run for strategy 2, and the 11th run for strategy 3.

ultimately concluding that the bootstrap is "more suitable for assessing the reliability of phylogenetic trees than posterior probabilities . . ." While we agree with Suzuki, Glazko, and Nei (2002) that a conservative measure of support is, in general, preferable to an inflated one, they did not determine whether their bootstrap values underestimated the ideal less than the Bayesian values overestimated it.

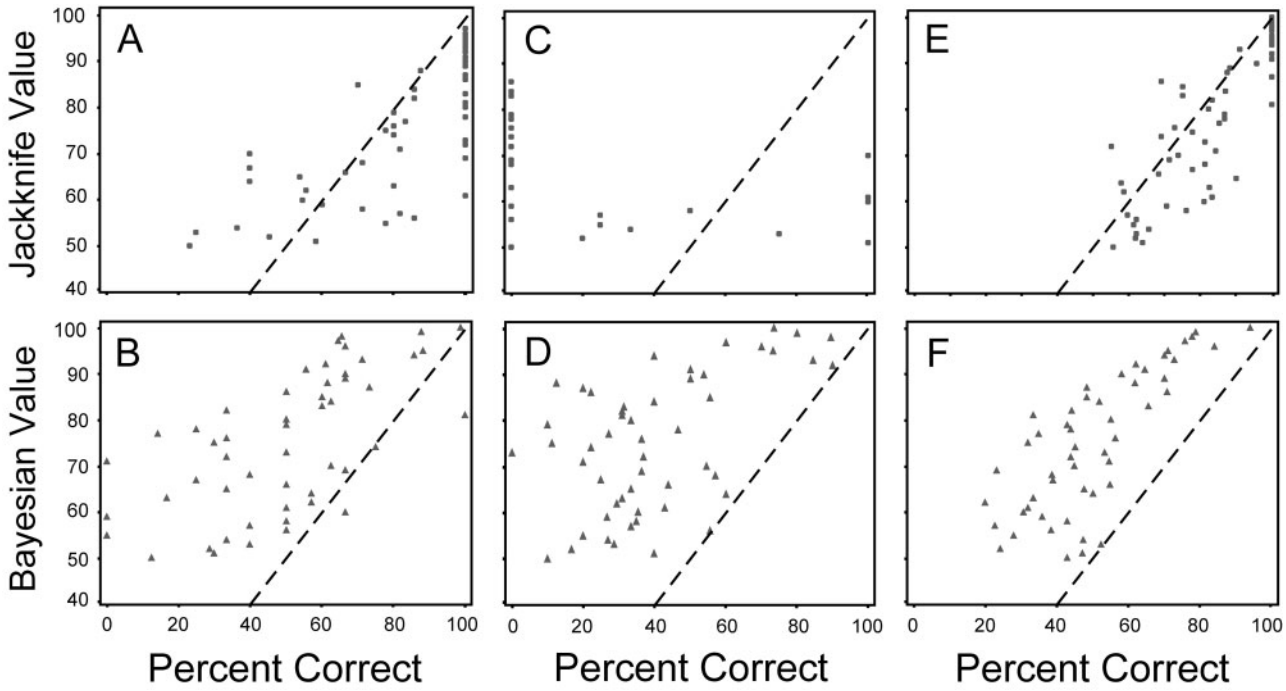

FIG. 3.—Scatter plots of jackknife and Bayesian support values plotted relative to the percent of the time that clades with that support value were correctly resolved, relative to the respective reference tree. The support values for the different sampling strategies are as follows: panel A, jackknife under strategy 1; panel B, Bayesian under strategy 1; panel C, jackknife under strategy 2; panel D, Bayesian under strategy 2; panel E, jackknife under strategy 3; panel F, Bayesian under strategy 3.

In their paper comparing Bayesian values with bootstraps under ML and parsimony, Alfaro, Zoller, and Lutzoni (2003, p. 261) asserted, ''[Bayesian MCMC posterior probability] appeared to lie closest to the line of perfect correspondence between accuracy and support for most scenarios.'' As in Wilcox et al. (2002) and Suzuki, Glazko, and Nei (2002), the decision rule is subjective visual inspection. Continuing this visual decision rule, Alfaro, Zoller, and Lutzoni (2003, p. 264) stated, ''Furthermore, [maximum parsimony] bootstrapping was usually more susceptible than [maximum likelihood] bootstrapping to assigning high support values to incorrect internodes (fig. 3 and table 1),'' and attribute the difference to long-branch attraction on the part of parsimony. However, their results showed that the Bayesian average is higher than the parsimony bootstrap average, the average parsimony bootstrap for incorrect internodes was minutely higher than that of maximum likelihood, and the ranges of all three measures were largely overlapping. Statistical treatment of these averages might well reveal no difference among any of the support values. However, because they did not test for any difference, Alfaro, Zoller, and Lutzoni (2003) effectively did not show any difference among the three methods.

In another recent paper, Douady et al. (2003) used linear regression to estimate the correlation between Bayesian support values and bootstrap values. Like Suzuki, Glazko, and Nei (2002), Douady et al. (2003, p. 252) found that Bayesian values were generally higher than bootstrap support, whether clades were correctly or incorrectly resolved, concluding, ''Thus, the more conservative [ML Bootstrap] and [bootstrapped Bayesian values] seem less subject to the behavior of strongly supporting a node when it is actually false.'' Unfortunately, the parametric linear regression, which is particularly sensitive to violations of its strict assumptions, is likely inappropriate for distributions of support value data. With these data, the assumptions of normality, standard deviation equality, or linearity are not likely met. However, the support for Douady et al.'s (2003) claims is stronger than most of the other papers on this topic because their decision rule was not subjective.

More recently, Cummings et al. (2003) compared Bayesian support values with the bootstrap by comparing their simulated data with a distribution that resulted from the randomized permutation of those data. This kind of statistic, which makes no assumption about the distributional properties of the data, is completely appropriate for support value data. As such, the conclusions drawn by this paper—that Bayesian values are inflated—are the most robust of all studies completed so far.

In each of these five papers, the authors asserted that one measure is superior to the other in a given context. However, Wilcox et al. (2002), Suzuki, Glazko, and Nei (2002), and Alfaro, Zoller, and Lutzoni (2003) did not assess the degree of disparity using a standard statistic. This lack of statistical comparison makes it difficult to interpret the relative merits of different conclusions when they contradict one another. However, the conclusions of Douady et al. (2003) and Cummings et al. (2003) are not subject to this kind of subjectivity.

We assessed the statistical departure of the jackknife and Bayesian support values from ideal by way of the nonparametric Wilcoxon Signed Ranks test. The Wilcoxon

Table 2
Data and Results for Various Statistical Comparisons of Support Value Distributions Using the Wilcoxon Signed Ranks Test[1]

Test 1			Test 2			Test 3		
Comparison	N Ranks	Statistical Results	Comparison	N Ranks	Statistical Results	Comparison	N Ranks	Statistical Results
JAC1 versus Ideal	+12		BAY1 versus Ideal	+48		BAY1* versus JAC1*	+42	
(A of figure 3)	−36	$Z = -2.472$	(B of figure 3)	−3	$Z = -5.980$		−5	$Z = -5.477$
	0	$P \cong 0.013$		0	$P \cong 0.000$		1	$P \cong 0.000$
JAC2 versus Ideal	+19		BAY2 versus Ideal	+51		BAY2* versus JAC2*	+9	
(C of figure 3)	−5	$Z = -3.300$	(D of figure 3)	−0	$Z = -6.215$		−15	$Z = -0.957$
	0	$P \cong 0.001$		0	$P \cong 0.000$		0	$P \cong 0.338$
JAC3 versus Ideal	+11		BAY3 versus Ideal	+51		BAY3* versus JAC3*	+50	
(E of figure 3)	−39	$Z = -3.944$	(F of figure 3)	−0	$Z = -6.215$		−1	$Z = -6.205$
	1	$P \cong 0.000$		0	$P \cong 0.000$		0	$P \cong 0.000$

NOTE.—The "N Ranks" columns show the number of support value points that overestimated (+), underestimated (–), and tied (no sign) Ideal. The "Statistical Results" columns show the *Z*-values and approximate *P*-values for each comparison. *Z*-values indicate the magnitude of overall departure from perfect fit ($Z = 0.0$). All *P*-values are two-tailed, testing the hypothesis that the two distributions are the same (H_0: $m_d = 0$; H_A: $m_d^1 = 0$). Test 1: *Support to Ideal* (see *Materials and Methods*) for jackknife analyses. JAC1, JAC2, and JAC3 represent the distributions of jackknife values resulting from parsimony analyses under strategies 1, 2, and 3, respectively (see *Materials and Methods*). Test 2: *Support to Ideal* (see *Materials and Methods*) for Bayesian analyses. BAY1, BAY2, and BAY3 represent the distributions of Bayesian values resulting form Bayesian (Likelihood model = GTR + I + Γ) analyses under strategies 1, 2, and 3, respectively (see *Materials and Methods*). Test 3: *Across Support Value, Within Strategy* (see *Materials and Methods*). Comparisons of Bayesian support to jackknife support for all three sampling strategies.

test is a so-called distribution-free statistic, meaning its interpretation does not depend on detailed assumptions regarding the properties of the distributions being compared (Devore and Peck 1993). The percentage of clades across replicates that correctly recovered clades found in the reference tree for a given support value were compared with an idealized support measure. Such an idealized measure would find 50% support for clades recovered correctly 50% of the time, 60% support for clades recovered correctly 60% of the time, and so forth (as in Wilcox et al. [2002]). This is phylogenetic accuracy *sensu* Hillis and Bull (1993). Support values were ranked according to their magnitude of departure from the ideal. Three kinds of statistical comparisons were conducted: (1) *Support to Ideal*—the performance of both jackknife and Bayesian support values were compared with ideal under each of the three sampling strategies described above; (2) *Across Support Value, Within Strategy*—the performance of the jackknife was compared with Bayesian support for the same sampling strategy; and (3) *Within Support Value, Across Strategy*—the performance of one support method (jackknife and Bayesian) was compared with the performance of the same method under different sampling strategies. These three statistical comparisons allowed us to test which method most accurately approximates the ideal. In other words, we tested whether either method outperforms the other and whether that performance was conditional on the sampling strategy employed. Note that the initial runs based on 18 taxa and 799 characters were not included when comparing the relative performance of jackknife and Bayesian support for each of the three sampling strategies independently of one another.

Results

A Microsoft Excel file of the raw data is available as Supplementary Material online at http://www.molbiolevol.org/.

Statistical Tests: Support to Ideal

Figure 3 shows the distribution of jackknife and Bayesian support values for all clades with support values of at least 50% for analyses under each of the three sampling strategies individually (plots A to F). By looking at the distribution of data points relative to the ideal (dashed line), it appears that the jackknife values very nearly approximated ideal in each case, except for strategy 2. Also, visual inspection indicates that the Bayesian values consistently overestimated support. However, our statistical analyses indicate that all of the jackknife support distributions and all of the Bayesian value distributions differed significantly from the ideal (table 2). The distribution of ranks above and below ideal for each comparison (table 2) show that the jackknife distributions fell above *and* below ideal, whereas the Bayesian values are almost entirely overestimates. Only in the strategy 1 did any Bayesian values underestimate support; in every other case, all the support values were overestimates (table 2, test 2). This observation is further supported by the skew of *Z*-values for Bayesian values compared with jackknife values. The departure of the *Z*-value from zero indicates the magnitude of departure from ideal exhibited by the distribution (Devore and Peck 1993), and many of the Bayesian *Z*-values show twice the departure from ideal when compared with the jackknife *Z*-values. However, there is no way to use difference in *Z*-values to validly measure the significance of difference. Although the *Z*-values indicate tendency and magnitude of difference, they themselves cannot be used alone to assess significance.

Statistical Tests: Across Support Value, Within Strategy

To determine whether the jackknife values departed from ideal *less* than the Bayesian values, as their *Z*-values suggest, the jackknife values for each sampling strategy were compared with the Bayesian values for the same strategies (see table 2, test 3). However, because under

Table 3
Data and Results for Various Statistical Comparisons of Support Value Distributions Using the Wilcoxon Signed Ranks Test

Test 4			Test 5		
Comparison	N Ranks	Statistical Results	Comparison	N Ranks	Statistical Results
JAC1* versus JAC2*	+19		BAY1* versus BAY2*	+33	
	−4	$Z = -3.337$		−16	$Z = -2.696$
	1	$P \cong 0.001$		2	$P \cong 0.007$
JAC1* versus JAC3*	+17		BAY1* versus BAY3*	+22	
	−21	$Z = -0.073$		−29	$Z = -0.473$
	10	$P \cong 0.942$		0	$P \cong 0.636$
JAC2* versus JAC3*	+5		BAY2* versus BAY3*	+16	
	−19	$Z = -3.743$		−35	$Z = -3.464$
	0	$P \cong 0.000$		0	$P \cong 0.001$

each strategy, the jackknife values tended to underestimate support, and the Bayesian values overestimated support, the raw values could not be meaningfully compared. This is because the tendencies of each support score was to the other side of the ideal, and as the distribution of support values of both methods under each strategy differed significantly from the ideal, comparing the raw distributions would necessarily result in significant differences. Therefore, adjusted distributions, calculated from the absolute value of the difference between each support datum and the ideal, were used for comparison (indicated by asterisks [*] in table 2, test 3). In each case except for strategy 2, the distribution of Bayesian values differed significantly from the distribution of jackknife values. This, coupled with the *Z*-values from the first two sets of comparisons (table 2, tests 1 and 2), demonstrates that although both the jackknife and Bayesian values differed from the ideal under every sampling strategy, jackknife values, with the exception of those for strategy 2, differed significantly less from ideal than did the Bayesian values. However, jackknife values under strategy 2 did not perform more poorly than did Bayesian values under strategy 2 (table 2, test 3, $P \cong 0.338$). Neither method performed well when taxa were increased and characters were not, resulting in the situation where the characters could not resolve the relationships reliably.

Statistical Tests: Within Support Value, Across Strategy

To address which sampling strategy gave the best results for each optimality criterion, the Wilcoxon Signed Ranks test was again used to compare the performance of each sampling strategy with ideal for both the jackknife and Bayesian analyses. As in test 3 (table 2), the absolute values of the difference between each support datum and ideal were used for comparisons (indicated by *). The distribution of jackknife values for sampling strategy 1 differed significantly from those of strategy 2 but did not differ from that of strategy 3 (table 3, test 4). The same phenomenon was observed for the Bayesian analyses (table 3, test 5). We have already established that the jackknife values from sampling strategies 1 and 3 more closely approximated ideal than did the Bayesian values for strategies 1 and 3 (table 2, test 3). In addition, the equally poor performance of both the jackknife and Bayesian values under strategy 2 was established (table 2, test 3). Given this, the results in table 3 further support the observation that the jackknife is a superior measure of ideal support. The jackknife values for strategies 1 and 3 appear to differ from the poorly performing values of strategy 2 more than the Bayesian values for strategies 1 and 3 differ from the poorly performing Bayesian values from strategy 2 (see more extreme *P*-values and *Z*-values for jackknife values in table 3).

Discussion

Proponents of Bayesian phylogenetics have pointed out that many of the details of these analyses remain to be elucidated (Huelsenbeck et al. 2002; Holder and Lewis 2003). One of these is the central claim that the frequency of clades summarized by the majority-rule consensus of trees generated under this procedure reflects the probability that the clade is true, given the priors, the model, and the data (but see Huelsenbeck et al. [2002, p. 675], who have suggested that other consensus methods may be valid). Recently, the phylogenetics community has begun to explore the limits of this assertion using both simulated (Suzuki, Glazko, and Nei 2002; Wilcox et al. 2002; Alfaro, Zoller, and Lutzoni 2003; Cummings et al. 2003; Douady et al. 2003) and empirical (Douady et al. 2003) data. Thus far, conclusions are split among the view that Bayesian support values are more reliable than the bootstrap as indicators that clades are correctly resolved (Wilcox et al. 2002; Alfaro, Zoller, and Lutzoni 2003), the opposite view (Suzuki, Glazko, and Nei 2002; Cummings et al. 2003), and the view that Bayesian values may form a reliable upper bound, whereas bootstrap values may form a more valid lower bound (Douady et al. 2003). However, the conclusions of all of these studies cannot be viewed equally, and conflicts among the conclusions of the various papers are likely the result of the various methods by which inferences were drawn from the data (i.e., statistical comparison vs. visual inspection).

Wilcox et al. (2002) and Suzuki, Glazko, and Nei (2002) tested the general performance of the two methods, drawing opposite conclusions. Suzuki, Glazko, and Nei (2002) and Wilcox et al. (2002) did agree that Bayesian

support values are higher than bootstrap overall. Indeed, many other studies have shown that Bayesian support values are high relative to bootstrap values (e.g., Rannala and Yang 1996; Leaché and Reeder 2002; Whittingham et al. 2002). So, although the authors of these studies have made different inferences about which measure is preferable, all agree on one property of Bayesian support values: they are comparatively high. Unfortunately, neither of these studies statistically established that bootstrap support differed significantly from Bayesian support or that either approximated an ideal measure better than the other. Cummings et al. (2003) analyzed statistically the general accuracy of Bayesian and bootstrap support in a simulated star topology case and concluded that bootstrap support is generally lower than Bayesian support because the latter is inflated. This statistically rigorous study supported the conclusions of Suzuki, Glazko, and Nei (2002).

Alfaro, Zoller, and Lutzoni (2003) and Douady et al. (2003) investigated support in a more specific, partitioned way. These two studies examined the performance of Bayesian and bootstrap support values when clades are correctly resolved and when they are incorrectly resolved. Both Alfaro, Zoller, and Lutzoni (2002) and Douady et al. (2003) showed that Bayesian support values are higher than bootstrap values when clades are resolved both correctly *and* incorrectly. These authors explained this relative inflation by asserting that Bayesian support performs better than the bootstrap when clades are correct, but this pattern would also emerge if Bayesian support were inflated regardless of character support. Even if this alternate explanation is incorrect, high support for incorrect clades is clearly worse than low-support values for correct clades.

In their four-taxon simulation study, Cummings et al. (2003) analyzed data that, when plotted, appear similar to the data of Alfaro, Zoller, and Lutzoni (2003) and Douady et al. (2003). However, after the application of an appropriate statistic, Cummings et al. (2003) concluded that in most cases, the bootstrap and Bayesian values were not significantly different. This accords with Douady et al.'s (2003, p. 250) conclusion that Bayesian support and ML bootstrap values are ''moderately correlated.''

In this study, we used a reference tree based on mitochondrial genomes of higher teleost fishes—the topology of which was the same for both Bayesian and parsimony methods of reconstruction for the 15 clades examined—to assess the statistical differences among Bayesian support values, parsimony jackknife values, and an ideal support measure. Our results show that Bayesian support values performed more poorly overall than do jackknife support values. Furthermore, our results demonstrated that Bayesian support values significantly overestimated support by a magnitude greater than the jackknife, and hence the bootstrap, underestimated support. These results were not dependent on the taxon and character sampling strategy employed (except for jackknife support for strategy 2, in which there were limited data available).

Our empirically based results statistically indicate that jackknife values are a better, albeit conservative, approximation of ideal, and that Bayesian values consistently overestimate support. These findings agree with the conclusions of Suzuki, Glazko, and Nei's (2002, p. 16140) simulations, which indicated the bootstrap to be ''slightly conservative,'' whereas Bayesian support values are ''excessively high.'' Our results also accord with Cummings et al. (2003, p. 484), who also found Bayesian support to be ''excessively high.''

In summary, our results indicate that (1) Bayesian support values are high relative to more common resampling support measures, (2) these higher Bayesian support values are inappropriate in magnitude (*contra* Rannala and Yang 1996; Wilcox et al. 2002), and (3) Bayesian support values should *not* be interpreted as probabilities that clades are correctly resolved (*contra* Rannala and Yang 1996; Huelsenbeck et al. 2001, 2002; Wilcox et al. 2002). The first conclusion has been found in all studies (including the present study) that have examined this issue. The second conclusion is one of statistical quantification, which only Douady et al. (2003), Cummings et al. (2003), and the present study have investigated, all of which agree with regards to inflation of Bayesian support values. The third conclusion is a logical deduction from the union of the first and second conclusions. By extension, our results suggest that methods that produce support values similar to Bayesian support values (e.g., metaGA [Lemmon and Milinkovitch 2002]) are also overestimating support and should not be interpreted as probabilities that clades are correctly resolved.

Larget and Simon (1999, p. 756) correctly noted that ''The validity of the [Bayesian] inferences depends on the validity of the likelihood model, prior distributions, and data'' (see also Wilcox et al. 2002, p. 369). However, given that the likelihood model, prior distributions, and subsets of data that we used for each Bayesian analysis were the same as those used to construct the reference tree, these qualifications cannot be used to explain away our results. We expect our results to apply equally well to traditional heuristic-search maximum-likelihood–based bootstrap analyses when compared with Bayesian support values. In conclusion, we advocate the continued use of the relatively conservative bootstrap and jackknife approaches to estimating branch support rather than the more extreme overestimates provided by the Markov Chain Monte Carlo–based Bayesian methods.

Acknowledgments

We thank Keith Crandall and two anonymous reviewers for detailed suggestions that significantly improved the manuscript, Mike Antolin, Donovan Bailey, Ryan Carr, John Freudenstein, Kevin Nixon, Helga Ochoterena, Chris Randle, the CSU Evolution Discussion Group, and the OSU Phylogenetics Discussion Group for helpful discussions and suggestions. Tom Waite was particularly helpful with the statistical analysis. This work was supported by a CSU Career Enhancement Grant (M.P.S.) and an OSU Presidential Fellowship (K.M.P.).

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. **19**:716–723.

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.

Baker, R. H., G. S. Wilkinson, and R. DeSalle. 2001. Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Syst. Biol. **50**:87–105.

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.

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

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.

Devore, J., and R. Peck. 1993. Statistics: the exploration and analysis of data. 2nd edition. Duxbury Press, Belmont, Calif.

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

Faith, D. P., and J. W. O. Ballard. 1994. Length differences and topology-dependent tests: a response to Källersjö et al. Cladistics **10**:57–64.

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

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22**:240–249.

———. 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.

Gatesy, J. 2000. Linked branch support and tree stability. Syst. Biol. **49**:800–807.

Goloboff, P. A. 1993. NONA (NO NAME). Published by the author, Tucumán, Argentina.

Harshman, J. 1994. The effect of irrelevant characters on bootstrap values. Syst. Biol. **43**:419–424.

Holder, M., and P. O. Lewis. 1993. Phylogenetic estimation: traditional and Bayesian approaches. Nat. Rev. Genet. **4**:275–284.

Hillis, D. M. 1999. Phylogenetics and the study of HIV. Pp. 105–121 *in* K. A. Crandall, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore.

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

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science **255**:589–592.

Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Ann. Rev. Ecol. Syst. **28**:437–466.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. **51**:673–688.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of 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.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2001. A mitogenomic perspective on the basal teleostean phylogeny: resolving higher-level relationships with longer DNA sequences. Mol. Phylogenet. Evol. **20**:275–282.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis for relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. **38**:7–25.

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.

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

Lemmon, A. R., and M. C. Milinkovitch. 2002. The metapopulations genetic algorithm: an efficient solution for the problem of large phylogeny estimation. Proc. Natl. Acad. Sci USA **99**:10516–10521.

Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

Miller, J. A. 2003. Assessing progress in systematics with continuous jackknife function analysis. Syst. Biol. **52**:55–65.

Miller, R. E., T. R. Buckley, and P. S. Manos. 2002. An examination of the monophyly of morning glory taxa using Bayesian phylogenetic inference. Syst. Biol. **51**:740–753.

Miya, M., A. Kawaguchi, and M. Nishida. 2001. Mitogenomic exploration of higher teleostean phylogenies: a case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. Mol. Biol. Evol. **18**:1993–2009.

Miya, M., and M. Nishida. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. **17**:437–455.

Miya, M., H. Takeshima, H. Endo et al. (12 co-authors). 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogenet. Evol. **26**:121–138.

Mort, M. E., P. S. Soltis, D. E. Soltis, and M. L. Mabry. 2000. Comparison of three methods of estimating internal support on phylogenetic trees. Syst. Biol. **49**:160–171.

Nixon, K. C. 2002. WinClada. Published by the author, Ithaca, New York.

Nixon, K. C., and J. M. Carpenter. 1996. On simultaneous analysis. Cladistics **12**:221–242.

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

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. **17**:1854–1858.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.

———. 2001. Selecting models of nucleotide substitution: an

application to human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **18**:897–906.
Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. **43**:304–311.
Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.
Rodrigo, A. G. 1993. Calibrating the bootstrap test of monophyly. Int. J. Parasitol. **23**:507–514.
Salamin, N., M. W. Chase, T. R. Hodkinson, and V. Savolainen. 2003. Assessing internal support with large phylogenetic DNA matrices. Mol. Phylogenet. Evol. **27**:528–539.
Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. **44**:299–320.
Sanderson, M. J., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. **49**:817–829.
Sanderson, M. J., and M. F. Wojciechowski. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). Syst. Biol. **49**:671–685.
Shevchuk, N. A., and M. W. Allard. 2001. Sources of incongruence among mammalian mitochondrial sequences: COII, COIII and ND6 genes are main contributors. Mol. Phylogenet. Evol. **21**:43–54.
Simmons, M. P., and J. V. Freudenstein. 2003. The effects of increasing genetic distance on alignment of, and tree construction from, rDNA internal transcribed spacer sequences. Mol. Phylogenet. Evol. **26**:444–451.
Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA **99**:16138–16143.
Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parssimony (*and other methods). Sinauer Associates, Sunderland, Mass.
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**:1396–1401.
Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.
Zharkikh, A., and W.-H. Li. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. **9**:1119–1147.
———. 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. **4**:44–63.

Keith Crandall, Associate Editor

Accepted September 9, 2003