Syst. Biol. 42(2):182–192, 1993

AN EMPIRICAL TEST OF BOOTSTRAPPING AS A METHOD FOR ASSESSING CONFIDENCE IN PHYLOGENETIC ANALYSIS

DAVID M. HILLIS AND JAMES J. BULL

Department of Zoology, The University of Texas, Austin, Texas 78712, USA

Abstract.—Bootstrapping is a common method for assessing confidence in phylogenetic analyses. Although bootstrapping was first applied in phylogenetics to assess the repeatability of a given result, bootstrap results are commonly interpreted as a measure of the probability that a phylogenetic estimate represents the true phylogeny. Here we use computer simulations and a laboratory-generated phylogeny to test bootstrapping results of parsimony analyses, both as measures of repeatability (i.e., the probability of repeating a result given a new sample of characters) and accuracy (i.e., the probability that a result represents the true phylogeny). Our results indicate that any given bootstrap proportion provides an unbiased but highly imprecise measure of repeatability, unless the actual probability of replicating the relevant result is nearly one. The imprecision of the estimate is great enough to render the estimate virtually useless as a measure of repeatability. Under conditions thought to be typical of most phylogenetic analyses, however, bootstrap proportions in majority-rule consensus trees provide biased but highly conservative estimates of the probability of correctly inferring the corresponding clades. Specifically, under conditions of equal rates of change, symmetric phylogenies, and internodal change of $\leq 20\%$ of the characters, bootstrap proportions of $\geq 70\%$ usually correspond to a probability of $\geq 95\%$ that the corresponding clade is real. However, under conditions of very high rates of internodal change (approaching randomization of the characters among taxa) or highly unequal rates of change among taxa, bootstrap proportions $>50\%$ are overestimates of accuracy. [Bootstrapping; accuracy; repeatability; phylogeny; parsimony; precision; statistical analyses; simulations.]

Felsenstein (1985) proposed using the statistical test of bootstrapping (Efron, 1979, 1982, 1987) to estimate confidence limits of internal branches in phylogenetic analyses. He suggested that characters in a matrix of taxa × characters can be sampled with replacement (bootstrapped) to create many new matrices of the same size as the original, each of which can be analyzed to find the best-fit tree (e.g., the shortest tree in the case of parsimony analysis). Felsenstein suggested that the results from numerous bootstrap replicates can then be combined in a majority-rule consensus tree to assess confidence in particular internal branches of the tree. An investigator can then evaluate whether a particular technique provides significant support of a particular a priori phylogenetic hypothesis.

The use of bootstrapping is increasing in systematic studies. Although Felsenstein (1985) suggested that "bootstrapping provides us with a confidence interval within which is contained not the true phylogeny, but the phylogeny that would be estimated on repeated sampling of many characters from the underlying pool of characters," many workers treat bootstrap results as statements about the probability that a particular clade is a real historical group. However, at best bootstrap estimates should provide an indication only of the degree of support of a particular technique for a particular clade. In cases in which the given technique is positively misleading about phylogeny (e.g., Felsenstein, 1978), we may be quite confident that the technique supports a clade, even though the probability that the clade is real may be quite low.

The first goal of this study was to examine the effectiveness of bootstrapping in assessing the probability that a given phylogeny would be obtained upon repeated sampling of different sets of characters from the underlying distribution of characters. The second goal was to examine the relationship, if any, between bootstrap

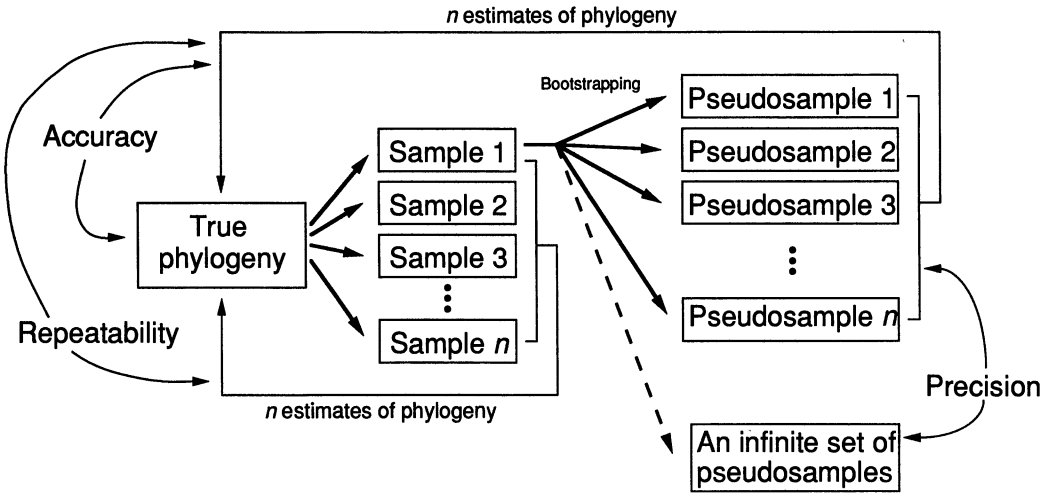

FIGURE 1. The relationships among a true phylogeny, samples of characters drawn from the taxa, bootstrap pseudosamples drawn from an initial sample, and the concepts of precision, accuracy, and repeatability.

proportions and the probability of obtaining a true clade in parsimony analyses under a variety of tree topologies and rates of change.

REPEATABILITY, ACCURACY, AND PRECISION

Before evaluating the performance of bootstrapping, it is necessary to distinguish among the terms "repeatability," "accuracy," and "precision." The meaning of these terms as they relate to bootstrapping is illustrated in Figure 1. Given an actual phylogeny of organisms, one could sample a set of characters among the taxa and then estimate the phylogeny from this sample. Bootstrapping is then used to create a set of pseudosamples by drawing characters from this initial sample with replacement. Phylogenetic analyses of these pseudosamples support a set of relationships for various clades; potential clades may appear in all, some, or none of the analyses based on the pseudosamples. We will refer to the proportions at which each clade is represented in these analyses as the *bootstrap proportions* (sometimes misleadingly called "bootstrap *P* values"). The *precision* of bootstrapping is the degree to which bootstrap proportions based on a finite set of pseudosamples are expected to match the values that would be obtained from an infinite set of pseudosamples. The sampling variance of bootstrap proportions follows the binomial distribution, such that $\sigma^2 = P(1 - P)/n$, where P is the bootstrap proportion and n is the number of replications (Hedges, 1992). Hedges (1992) called this relationship "accuracy" with respect to phylogenetic analyses, but this is not within the usual meaning of this term (and he no longer follows this terminology, e.g., Hedges and Maxon, 1993). Instead, we use *accuracy* to refer to the probability that a specified group is contained in the true phylogeny. Although the use of bootstrap proportions as a measure of accuracy has not been explicitly defended in print, they are commonly treated as such in the framework of hypothesis testing. In his original paper, Felsenstein (1985) suggested that the bootstrap proportions could be used as a measure of what we will call *repeatability*: the probability that a specified group will be found in an analysis of an independent sample of characters. If a group of interest is found in a phylogenetic analysis, an estimate of the repeatability of this result would tell us how likely we would be to find the same result in an identical analysis of a new sample of characters. Thus, as used in this paper, precision relates to the

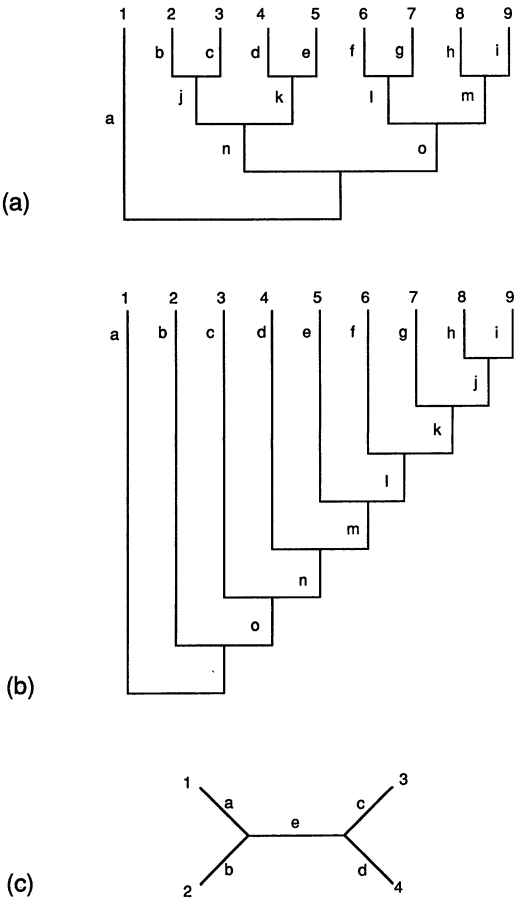

FIGURE 2. Simulated phylogenies. The letters correspond to branch lengths shown in Tables 1 and 2. (a) Phylogeny for simulations 1–9 for nine taxa (Table 1). (b) Phylogeny for simulation 10 for nine taxa (Table 1). (c) Phylogeny for simulations 11–21 for four taxa (Table 2).

correspondence between multiple sets of bootstrap pseudosamples taken from the same initial sample, accuracy is the probability that a given result represents the true phylogeny, and repeatability is the probability that a given result will be found again using a subsequent sample of characters.

Of these concepts, only the precision of bootstrapping has been examined to date with respect to phylogenetic analysis; the precision of any given bootstrap analysis has a simple relationship to the number of pseudosamples (see above and Hedges, 1992). Hedges (1992) recommended performing 400–2,000 bootstrap replications "if one wishes the expectation to be that the 95% confidence range is ±1% of the BP" (where BP = bootstrap *P* value, or bootstrap proportion of our terminology). However, to what does the "95% confidence range" apply? The findings of Hedges relate not to accuracy or repeatability but to the precision of bootstrapping: the number of pseudosamples needed to obtain an estimate (of specified precision) of the bootstrap proportions that would be obtained after an infinite number of pseudosamples. However, if the bootstrap proportions are not good estimates of accuracy, repeatability, or some other useful measure, it makes little sense to worry too much about precision. A highly precise estimate of an inaccurate measure is of little importance. Therefore, we will focus on the interpretation of the bootstrap proportions: how do they perform as measures of either accuracy or repeatability?

METHODS

To evaluate bootstrap proportions as measures of accuracy and repeatability, knowledge of the true phylogeny and ability to take true replicate samples are necessary. We have used two systems to meet these requirements: simulations and a laboratory-generated phylogeny of viruses.

Simulations

Ten nine-taxon phylogenies were simulated 100 times each (Figs. 2a, 2b; Table 1), and 11 four-taxon phylogenies were simulated 1,000 or 10,000 times each (Fig. 2c; Table 2). In each simulation, the ancestral node was assigned a DNA sequence with equal proportions of all four nucleotides, from 50 to 1,000 nucleotides in length. Branch lengths of the trees were assigned values between 1 and 10 units inclusive, with a unit equal to an assigned rate of change. The rate of change along a unit branch length (R) was varied from 4% to 75%; therefore, the instantaneous rates of change (r) varied from 5.5% to infinity, because $R = 0.75(1 - e^{-r})$. The ratio of transitions : transversions was varied from

TABLE 1. Relative branch lengths and rates of mutation for simulated phylogenies of nine taxa. These simulations were each replicated 100 times, and 100 bootstrap pseudoreplicates were generated for each replicate.

Simulation number	Topology[a]	Branch lengths															Mutation rate	
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	Transitions	Transversions
1	a	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.020	0.020
2	a	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.050	0.050
3	a	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.075	0.075
4	a	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.100	0.100
5	a	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.050	0.005
6	a	3	1	3	1	3	1	3	1	3	1	1	1	1	1	1	0.050	0.005
7	a	3	1	5	1	5	1	5	1	5	1	1	1	1	1	1	0.050	0.005
8	a	3	1	7	1	7	1	7	1	7	1	1	1	1	1	1	0.050	0.005
9	a	3	1	10	1	10	1	10	1	10	1	1	1	1	1	1	0.050	0.005
10	b	8	7	6	5	4	3	2	1	1	1	1	1	1	1	1	0.050	0.005

[a] See Figure 2.

1:2 (equal probability of all base changes) to 10:1 (strong bias in favor of transitions). In three sets of simulations (numbers 15–17, Table 2), the rate of change and the transition : transversion ratio were set to produce random DNA sequences for each taxon (i.e., no phylogenetic signal). The 21,000 simulated phylogenies were generated with a program written in C (TREES, available upon request) and compiled on a Sun SPARC station 1+.

Bacteriophage Phylogeny

Generation of the phylogeny of T7 bacteriophage was described by Hillis et al. (1992). The design of the experiment was identical to the topology in Figure 2a, with relative branch lengths equal to those in simulations 1–4 (Table 1). The ancestral taxon was wild-type bacteriophage T7 (with a total genome of 39,937 base pairs [bp] of DNA). The descendant lineages were grown in the presence of a mutagen (nitrosoguanidine), which produces an excess of GC → AT changes, although all combinations of mutations are possible. The phage at each node was obtained from a single plaque to ensure that all descendants were derived from the same ancestor (see Hillis et al., 1992, for further details).

The entire genomes of the phages at each node of the phylogeny were mapped for 34 restriction enzymes (with 4-, 5-, and 6-bp recognition sites). Three deletions and 290 restriction sites were mapped, of which all of the deletions and 199 restriction sites were variable among the lineages. Previous study (Hillis et al., 1992) has shown that standard methods of tree construction using this matrix of restriction sites produce the correct (known) branching relationships.

Bootstrapping

Each of the 21,000 matrices from the simulations was analyzed by bootstrapping with 100 replicates using the software

TABLE 2. Simulations based on four-taxon trees with equal branch lengths (Fig. 2c). These simulations were each replicated 1,000 times, except for simulation 21, which was replicated 10,000 times, and 100 bootstrap pseudoreplicates were drawn from each replicate.

Simulation number	Mutation rate		No. characters
	Transitions	Transversions	
11	0.05	0.05	50
12	0.10	0.10	50
13	0.15	0.15	50
14	0.20	0.20	50
15	0.25	0.50	50
16	0.25	0.50	200
17	0.25	0.50	1,000
18	0.05	0.05	100
19	0.05	0.05	200
20	0.05	0.05	1,000
21	0.15	0.30	50

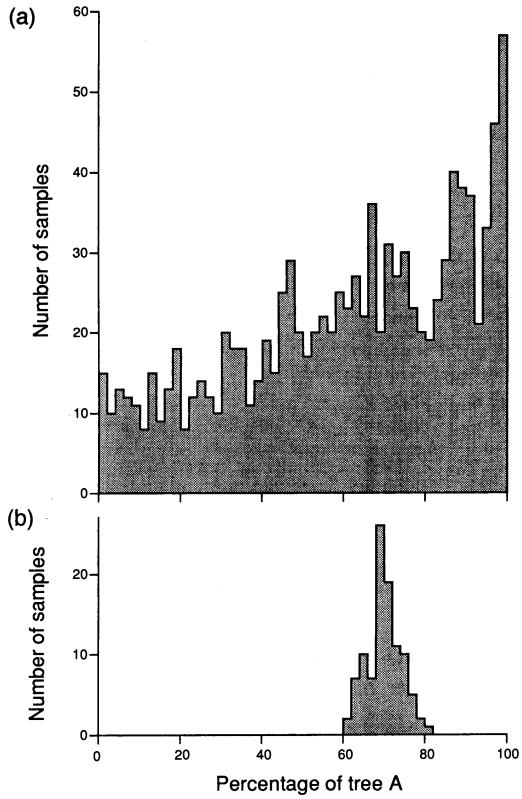

FIGURE 3. Bootstrap proportions as estimates of repeatability in phylogenetic analysis. (a) Results of 1,089 bootstrap analyses (100 pseudoreplicates each) on 1,089 actual replicates from simulation 21 (Table 2). Results shown are the proportions of the initial tree (tree A, which is the correct tree) found in each of the 1,089 analyses. (b) Proportions of solution A (the correct tree) in 100 samples of 100 actual replicates of simulation 21. The probability of estimating tree A from any given replicate is approximately 70%; samples of 100 actual replicates produce estimates of this value of 60–80%. In contrast, estimates of repeatability based on 100 bootstrap pseudoreplicates range from 0% to 100%, depending on the initial sample examined.

package PAUP 3.0q (Swofford, 1990). With 100 replicates, the sample variance of the bootstrap proportions (BP) ranges from a maximum of 0.0025 (at 50% BP) to a minimum of 0 (at 0 and 100% BP) (Hedges, 1992). Because the mean bootstrap proportions reported in this study are based on >100 replicates each, the standard error of the means is no greater than 0.005 for any value. The branch-and-bound algorithm was used in the bootstrap runs to ensure that all most-parsimonious trees were found. We also found the most-parsimonious trees for each matrix, so that the bootstrap proportions could be compared with the probability of finding the same result through repeated sampling of characters from the underlying distribution (repeatability). In addition, the collection of bootstrap estimates was compared with the known phylogenies to study the relationship between bootstrap proportions and the probability of correctly estimating phylogeny (accuracy).

To produce data matrices of phage restriction-site characters that were comparable to the simulations, we jackknifed (sampled characters without replacement) the complete matrix from the T7 phage study (Hillis et al., 1992) to produce 500 subsamples of 50 characters each. Each of these matrices was then analyzed by bootstrapping as above so that the bootstrap proportions could be compared with the probability of obtaining a true clade.

BOOTSTRAP ESTIMATES AS MEASURES OF REPEATABILITY

How well do bootstrap proportions represent the proportions that would be obtained from the independent samples that the bootstrap aspires to represent? This question was evaluated directly. Bootstrap proportions were calculated for a single model phylogeny (simulation 21, Table 2). Each bootstrap proportion was based on 100 pseudoreplicates of a data matrix. The distribution of this bootstrap proportion statistic is shown in Figure 3a. For comparison, the distribution of the analogous *sample proportion* was also calculated. A sample proportion is the fraction of correct reconstructions from 100 independent data matrices (without any pseudoreplication). The distribution of the sample proportion is shown in Figure 3b.

Comparison of these two distributions reveals that the process of bootstrap resampling is not the same as repeated, independent sampling of data. The means of both distributions are similar, but the variance of the bootstrap distribution is much

greater. The larger variance of the bootstrap proportion arises because of the bias inherent in the sample used for the 100 pseudoreplicates in each bootstrap proportion. Bootstrapping as a general statistical procedure provides a biased estimate of the mean, and consequently, bootstrapping is typically used to estimate the variance. Yet in the present case, which is one of binomial sampling, the variance is not independent of the mean, further complicating the use of bootstrapping.

What do these distributions indicate about repeatability? From our definition, the repeatability of recovering the correct tree is approximately 70% (the mean), and both distributions give similar values. However, the variance of the bootstrap proportion is so large that any single value is unreliable: we could easily obtain bootstrap proportions near 0 or 100. Use of a single bootstrap proportion as a measure of repeatability is therefore unreliable in this case.

We repeated similar analyses for each of the simulations in Tables 1 and 2, and the bootstrap proportions are unreliable estimates of repeatability unless repeatability is near 100% (in which case both distributions cluster near the right-hand side of the scale). Unfortunately, a value of 100% in a bootstrap analysis is not informative about repeatability because such an estimate is likely to arise when the true repeatability value is considerably lower (e.g., see Fig. 3). When repeatability is appreciably less than 100%, the size of the initial sample of characters has little effect on the quality of the estimates. However, increasing sample size does tend to force the true repeatability toward 100%.

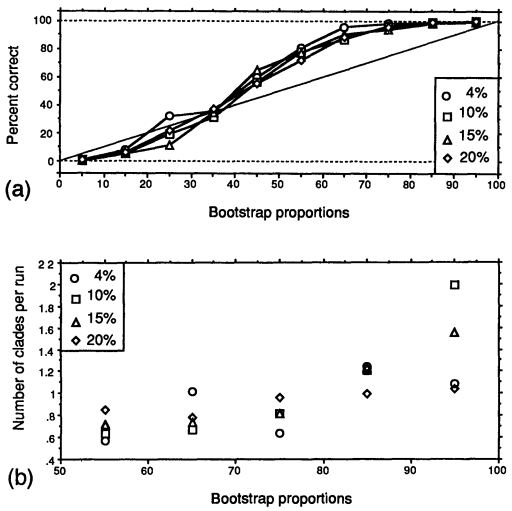

FIGURE 4. (a) Relationship between bootstrap proportions and the probability of the corresponding clade being correct at various rates of internodal change (shown in inset) in nine-taxon simulations (1–4). The diagonal line indicates direct correspondence between x and y axes. (b) The average number of clades found within given bootstrap proportions in simulations 1–4 (Table 1).

Repeatability versus Accuracy

Figure 3 indicates that bootstrap proportions are not good estimates of the repeatability of a given phylogenetic analysis. However, under most circumstances systematists are interested in accuracy more than repeatability. A phylogenetic analysis may be highly repeatable but always produce the wrong answer (e.g., the conditions described by Felsenstein, 1978). However, if a technique only provides an unambiguous answer when it is likely to be correct, then the technique may be accurate even though the result is obtained in only a small percentage of trials. Although bootstrapping was introduced as a measure of repeatability, bootstrap results commonly are interpreted as a measure of accuracy (e.g., in a framework of hypothesis testing). Therefore, we examined the possibility of a relationship between bootstrap proportions and the probability of obtaining a correct clade.

The relationship between bootstrap proportions and probability of the corresponding clade being correct is shown for simulations 1–4 in Figure 4a. This figure shows that estimated internal branches with bootstrap proportions above 70% represent true clades over 95% of the time (for the conditions tested in these simulations). In fact, the bootstrap proportions are lower than the probability of being correct for all estimates above 50% in these simulations. Although the probability of character change (over the tested range of 4–

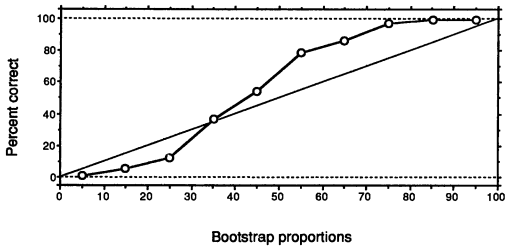

FIGURE 5. Relationship between bootstrap proportions and the probability of the corresponding clade being correct for the laboratory-generated phylogeny of nine taxa derived from bacteriophage T7. The diagonal line indicates direct correspondence between x and y axes.

20% change between nodes) has little effect on the relationship between these two variables, it does affect the likelihood of obtaining a bootstrap proportion within a given range (Fig. 4b). For instance, one is twice as likely to obtain a clade with a bootstrap proportion of at least 90% if the probability of change between nodes is 10% than if it is 4% or 20% (Fig. 4b). Nonetheless, almost every internal branch with a bootstrap proportion of >80% defined a true clade in all these simulations, and >95% of the estimated clades with bootstrap confidence limits above 70% were correct (Fig. 4a). Therefore, under these conditions, bootstrapping provides a biased but highly conservative estimate of accuracy.

Simulations are sometimes criticized because of the numerous simplifying assumptions they must incorporate (Hillis et al., 1993). Experimental phylogenies of actual organisms, generated in the laboratory so that the actual phylogeny is known, are a step closer to reality. We have generated such a phylogeny (using bacteriophage T7) for conditions similar to simulations 1–4 (Hillis et al., 1992). We jackknifed the complete data matrix for this actual phylogeny to produce sample matrices of 50 characters each; these matrices provide a comparison for the simulations. Figure 5 shows that the relationship between bootstrap proportions (bootstrapping done on the sample matrices) and the probability of an estimated branch being correct is virtually the same for the T7 phylogeny as it is in

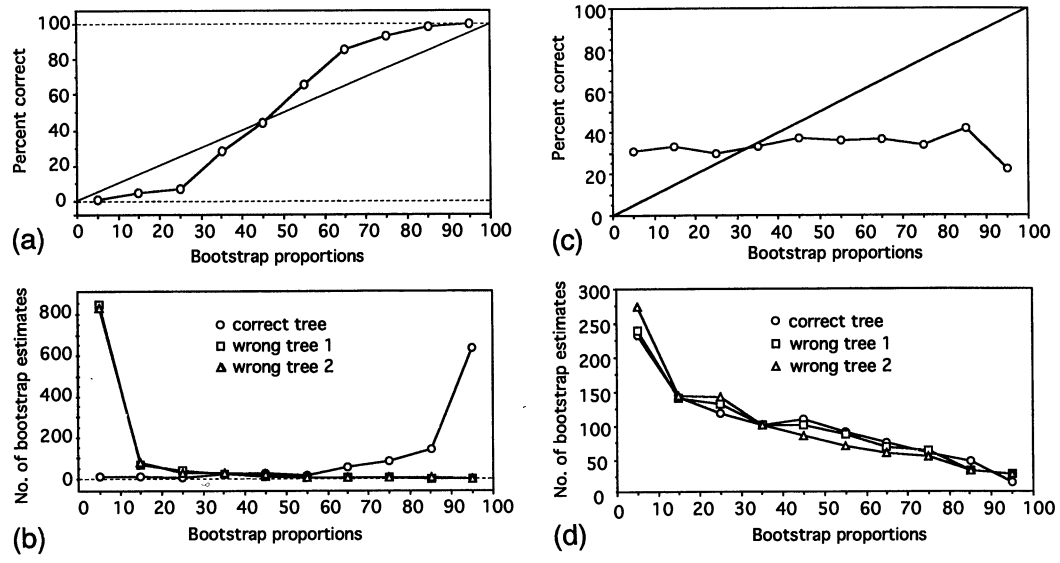

FIGURE 6. Results from simulations 11 (a, b) and 15 (c, d) with four taxa. The relationship between bootstrap proportions and accuracy (for simulation 11) is shown in (a); the shape of this curve can be understood by comparing the bootstrap results for each of the three possible topologies in (b). In contrast, simulations in which the characters are randomized among taxa by high rates of change produce bootstrap proportions in which any tree is equally likely to be supported (c, d).

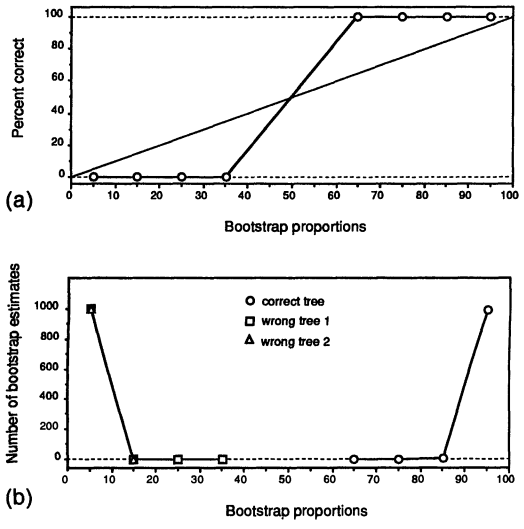

FIGURE 7. Bootstrap results from simulation 19 (Table 2) with four taxa. With a relatively large number of characters and appropriate rates of internodal change, all clades with bootstrap proportions above 50% are correct (a). This occurs because of the nonoverlapping ranges of bootstrap proportions for correct and incorrect trees (b).

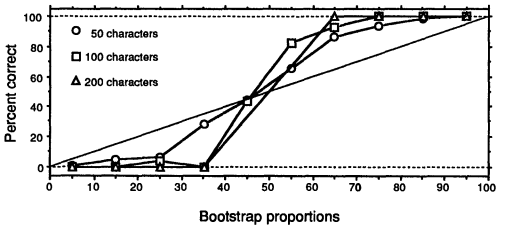

FIGURE 8. The effect of increasing size of the data matrix on the relationship between bootstrap proportions and the probability of the corresponding clade being correct (from simulations 11, 18, and 19, all with four taxa).

the simulations. Almost every internal branch with a bootstrap proportion above 70% defines a true clade, whereas fewer than 10% of the estimated branches with bootstrap proportions below 30% are correct.

The shape of the curves in Figures 4 and 5 can be understood by examining the frequency distributions of bootstrap proportions for correct and incorrect internal branches. In the simulations with four taxa (Table 2), there are only three possible unrooted trees (each with a unique internal branch). In Figure 6, the results of two four-taxon simulations are contrasted: simulation 11, in which the rate of change is appropriate for phylogenetic analysis, and simulation 15, in which the sequences are randomized among taxa by a high rate of change. In Figure 6a, the plot of bootstrap proportions against probability of a branch being correct takes on a shape similar to that for nine-taxon trees (Figs. 4, 5). In Figure 6b, the bootstrap frequencies for each of the three possible trees (one correct and two incorrect) are shown. These distributions are highly skewed in opposite directions, so that almost all the observations at the high end of the scale (>70%) correspond to the correct tree, and almost all the observations at the low end of the scale (<30%) correspond to one of the wrong trees. Compare this situation with that in Figure 6c: if characters are randomized among taxa by high rates of change, all bootstrap frequencies are equally likely to represent any of the three topologies (Fig. 6c), although the number of trees with high bootstrap proportions will be relatively low (Fig. 6d). At appropriate rates of change (simulations 18–20, 100–1,000 characters), an increase in the number of informative characters decreases the chances of a correct tree having a low bootstrap proportion or of a wrong tree having a high bootstrap proportion (Figs. 7, 8). In simulation 20 (1,000 characters, 10% change between nodes), the correct tree was found in every bootstrap replicate on all 1,000 matrices.

The results presented in Figures 4–8 indicate that when conditions are favorable for phylogenetic analysis (i.e., equal and appropriate rates of change and symmetrical phylogenies), bootstrap proportions are highly conservative measures of the probability that the corresponding clade is true. To investigate the conditions under which bootstrapping might positively reinforce misleading results, we further investigated high rates of change, highly unequal rates of change, and asymmetrical tree topologies.

Although high rates of change do reduce the gap between the probability of cor-

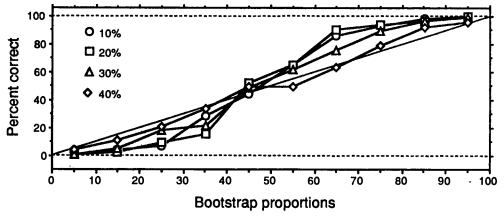

FIGURE 9. Relationship between bootstrap proportions and the probability of the corresponding clade being correct, as affected by different rates (10–40%) of internodal change. As internodal change approaches 40% of the characters, the bootstrap proportions approach the probability that the corresponding clade is correct.

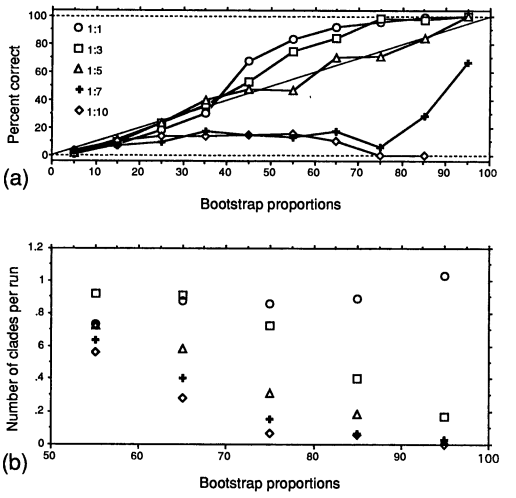

FIGURE 10. Relationship between bootstrap proportions and the probability of the corresponding clade being correct, as affected by unequal rates of change among lineages (from simulations with nine taxa). (a) When the ratio of short to long branches is less than 1:5 (simulations 5 and 6), bootstrap proportions are conservative estimates of accuracy. At a ratio of 1:5 (simulation 7), the bootstrap proportions are a fairly accurate indicator of accuracy; above a ratio of 1:5 (simulations 8 and 9), bootstrap results can be positively misleading. However, under the misleading conditions, the likelihood of finding any clade with a high bootstrap proportion is very low (b).

rectly resolving a node and bootstrap proportions (Fig. 9), the conservative nature of the bootstrap proportions is substantial throughout the range of rates typical of most phylogenetic studies. Only as internodal change approaches 40% are bootstrap proportions close to the probability of correctly resolving the corresponding branch, at least in the four-taxon simulations (Fig. 9). At such high rates of change, it is difficult to detect a significant amount of phylogenetic signal (Hillis and Huelsenbeck, 1992), and phylogenetic analyses are rarely attempted with such rapidly evolving sequences.

Trees that contain long, undivided branches interspersed with short branches are particularly difficult to reconstruct (Felsenstein, 1978). Such trees may exist if rates of evolution vary greatly among lineages or because of the timing of cladogenic events. In simulations 5–9, we examined the effects of various rates of change among terminal lineages. The results of the bootstrap analyses of these simulations are shown in Figure 10. For the nine-taxon topology tested (Fig. 2a; Table 1), bootstrapping proportions above 50% are consistently conservative measures of accuracy when the ratio of the short : long terminal branches is less extreme than 1:5 (Fig. 10). At more extreme ratios for the short : long branches, parsimony analysis becomes misleading, and branches with high bootstrap proportions are highly unlikely to be correct. However, under these extreme conditions, no clade (either correct or not) is likely to be represented in a high proportion of bootstrap replicates (Fig. 10b). For instance, among the 11,099 bootstrap proportions generated in simulation 9, none were >85% and only 17 were >65%.

Asymmetric phylogenies (Fig. 2b, simulation 10) also reduce the gap between bootstrap proportions and the probability of correctly resolving the corresponding branch (Fig. 11). Nonetheless, the bootstrap proportions (above 50%) for even a completely asymmetrical phylogeny are below the probability of correctly inferring the existence of a clade when rates of change are equal.

CONCLUSIONS AND RECOMMENDATIONS

The factors that affect the performance of bootstrapping depend on the use to which it is applied. If one wants only a precise measure of a bootstrap proportion, with no connection between that value and

any measure of repeatability or accuracy, then only the number of bootstrap iterations is of concern (Hedges, 1992). However, the statistical meaning of such a measure is obscure: the measure may be precise, but what is it measuring? If one is interested in the repeatability of a given result (i.e., what would happen if someone were to draw a new set of characters that are evolving in the same way as the first set), then bootstrap proportions are rarely useful. With an infinite number of bootstrap iterations on a given data set, one would obtain a perfectly precise but highly inaccurate estimate of repeatability. The proportions from bootstrap pseudoreplications are likely to differ dramatically from proportions calculated by actual replications. Finally, if one is interested in the probability that a recovered group represents a true clade, then at least the following variables should be taken into account: (1) number of characters, (2) number of taxa, (3) number of bootstrap iterations performed, (4) rate of change, (5) tree topology, (6) position of the group of interest within the tree, (7) variance of rates of change among lineages, (8) independence of characters, and (9) method of phylogenetic inference. Under a wide variety of conditions for the first seven of these variables using parsimony analysis of independently evolving characters, bootstrapping proportions above 50% are consistently much lower than the probability that the corresponding branch is correct. Because bootstrap results are typically presented in the form of a 50% majority-rule consensus tree, it is safe to assume that the bootstrap values are underestimates of phylogenetic accuracy unless (1) rates of change are highly unequal, (2) rates of change are high enough to randomize characters with respect to history, or (3) a systematic bias exists in the data set (such as lack of independence among the characters). The first two of these three conditions can be detected using existing methods (e.g., Allard and Miyamoto, 1992; Hillis and Huelsenbeck, 1992). Systematic biases may be harder to address for some data sets, but all phylogenetic analyses rest on the assumption that the characters examined are evolving independently, so some effort is usually expended to ensure that this assumption has been met (e.g., Wheeler and Honeycutt, 1988).

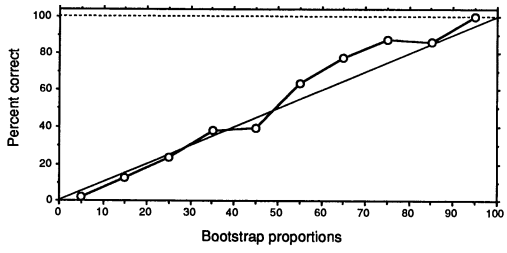

FIGURE 11. Relationship between bootstrap results and the probability of the corresponding clade being correct for a completely asymmetrical topology (simulation 10, with nine taxa). For such a topology, the bootstrap proportions are still conservative measures of reliability but less so than for symmetrical topologies (contrast with Fig. 4).

Zharkikh and Li (1992a, 1992b) have recently studied the four-taxon case (with and without a molecular clock) using analytical and simulation approaches. They studied the relationship between accuracy and bootstrap proportions in parsimony and neighbor-joining analyses and also found the bias that we report here for both types of analyses. For the four-taxon case, they recommended that bootstrap proportions could be used in a conservative assessment of accuracy as long as rates of evolution are not so variable that the phylogenetic method is inconsistent. Other previous studies of bootstrapping are also consistent with our finding that bootstrap proportions provide conservative estimates of accuracy under many conditions (e.g., Penny and Hendy, 1986). In responding to our paper, Felsenstein and Kishino (1993) addressed the issue of bias in bootstrap analyses and concluded that this bias may be a more general phenomenon, associated with placing a probability value on a prespecified hypothesis, rather than a characteristic that is limited to bootstrapping.

If bootstrap analyses provide biased estimates of accuracy, is it worth undertaking these analyses at all? Bootstrapping may provide a relative ranking of the degree of

support in a particular analysis for the various recovered clades. Sanderson (1989) concluded that these rankings are superior to other measures, such as number of characters supporting a given branch. The strong positive relationship between high bootstrap proportions and phylogenetic accuracy does indicate a use for bootstrapping. However, bootstrap results should not be interpreted directly as estimates of either repeatability or accuracy under most conditions. They are poor estimates of repeatability and are usually very conservative estimates of accuracy. Under many conditions, bootstrapping can be used as a highly conservative measure of accuracy, but the magnitude of bias will differ from branch to branch and study to study. The values cannot be directly compared among studies.

It may be possible to use retrospective simulation studies to calibrate bootstrap proportions so that these proportions can be converted into acceptable estimates of accuracy. Such studies would use a model phylogeny that is estimated from the initial sample to conduct simulations, which would be used in turn to calibrate the bootstrap proportions. It remains to be seen if biases in the initial phylogenetic estimate would adversely affect such retrospective simulations (=parametric bootstrapping; see Bull et al., in press).

ACKNOWLEDGMENTS

We thank Michael Donoghue, Joe Felsenstein, Argye Hillis, Wen-Hsiung Li, David Maddison, Wayne Maddison, David Penny, Michael Sanderson, and Andrey Zharkikh for commenting on various versions of this manuscript and Wen-Hsiung Li and Andrey Zharkikh for sending us advance copies of their work on bootstrapping. This work was supported by NSF grants DEB 9221052 and BSR 9106746.

REFERENCES

ALLARD, M. W., AND M. M. MIYAMOTO. 1992. Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. Mol. Biol. Evol. 9:778–786.

BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEUX, M. R. BADGETT, AND D. M. HILLIS. In press. Experimental molecular evolution of bacteriophage T7. Evolution.

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

EFRON, B. 1982. The jackknife, the bootstrap, and other resampling plans. Conf. Board Math. Sci. Soc. Ind. Appl. Math. 38:1–92.

EFRON, B. 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82:171–185.

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

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

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

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.

HEDGES, S. B., AND L. R. MAXON. 1993. A molecular perspective on lissamphibian phylogeny. Herpetol. Monogr. 7 (in press).

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.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT, AND I. J. MOLINEUX. 1993. Experimental approaches to phylogenetic analysis. Syst. Biol. 42:90–92.

HILLIS, D. M., AND J. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83:189–195.

PENNY, D., AND M. HENDY. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3:403–417.

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

SWOFFORD, D. L. 1990. PAUP: Phylogenetic analysis using parsimony, version 3.0. Illinois Natural History Survey, Champaign.

WHEELER, W. C., AND R. L. HONEYCUTT. 1988. Paired sequence difference in ribosomal RNAs: Evolution and phylogenetic implications. Mol. Biol. Evol. 5:90–96.

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

ZHARKIKH, A., AND W.-H. LI. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. J. Mol. Evol. 35:356–366.

Received 10 July 1992; accepted 7 January 1993