Letter to the Editor

The Number of Replications Needed for Accurate Estimation of the Bootstrap *P* Value in Phylogenetic Studies[1]

S. Blair Hedges
Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University

The bootstrap is a statistical method for obtaining a nonparametric estimate of error (Efron 1979, 1982). Felsenstein (1985) was the first to apply this method to phylogeny estimation, and his approach is now widely used. Taxa are held constant, and the characters (for sequence data, nucleotide or amino acid sites) are resampled randomly with replacement. A phylogeny is constructed from each replication of the data, and the frequency of appearance of particular phylogenetic groups (groups of alleles or taxa) among all of the trees constructed by this resampling is the bootstrap confidence limit, or bootstrap *P* value (BP). The BPs of different nodes within a tree can be used to assess the relative stability of those phylogenetic groups or, if applied strictly, to test their statistical significance (e.g., at the 95% or 99% level). The application of bootstrapping to phylogeny estimation is a tradeoff between the maximum number of replications that can be performed by the researcher in a reasonable amount of time and the minimum number of replications needed for accurate estimation of the BP. The purpose of the present report is to explore the variance (and hence the accuracy) of the phylogenetic BP and to establish guidelines for efficient bootstrap sampling.

BP is the proportion of trees containing a particular phylogenetic group. It therefore follows the binomial distribution, which has a variance of $\sigma^2 = [P(1 - P)/n]$, where P is the BP and n is the number of replications. Although Li and Gouy (1990) recently suggested that more replications are needed for larger numbers of taxa, the accuracy of the BP is a function only of P and n. If the interval containing 95% of the samples (± 1.96 standard deviations) is used as a measure of accuracy, then the application of the above formula shows that 1,825 replications [$=0.95 \times 0.05(1.96/0.01)^2$] are needed to attain $\pm 1\%$ accuracy at a BP of 0.95 (fig. 1). This is more than an order of magnitude higher than the number of replications (50–100) normally used in phylogenetic analyses.

Based on this, a practical guideline for efficient and accurate bootstrap sampling can be made: If one wishes the expectation to be that the 95% confidence range is $\pm 1\%$ of the BP, then one must perform *2,000 bootstrap replications* (if *BP = 0.95*) *or 400 replications* (if *BP = 0.99*) in phylogenetic analyses, unless the computational time is prohibitive; additional replications are unnecessary, and fewer replications may sacrifice statistical accuracy. Moreover, statistical testing at the 95% level cannot be made using <73 replications, even if the group is supported by a BP of 1.00. This is because the inaccuracy at a mean BP of 0.95 is greater than $\pm 5\%$ (fig. 1) when <73 replications are used. In other words, a BP ≥ 1.00 could be obtained when the actual (mean) BP is not significant (<0.95). Thus, the 50 replications used by Felsenstein (1985) in his original example and the 20–100 replications used in many subsequent studies (e.g., see Ovenden et al. 1987; Thomas et al. 1989; Jansen et al. 1990; Meyer and Wilson 1990; Douglas et al. 1991; Irwin et al. 1991) would appear to be far too few for the intended purpose of statistical testing at the 95% level.

1. Key words: phylogeny, statistics, DNA sequence, systematics, evolution.

Address for correspondence and reprints: S. Blair Hedges, Department of Biology, 208 Mueller Lab, Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 9(2):366–369. 1992.

0737-4038/92/0902-0013$02.00

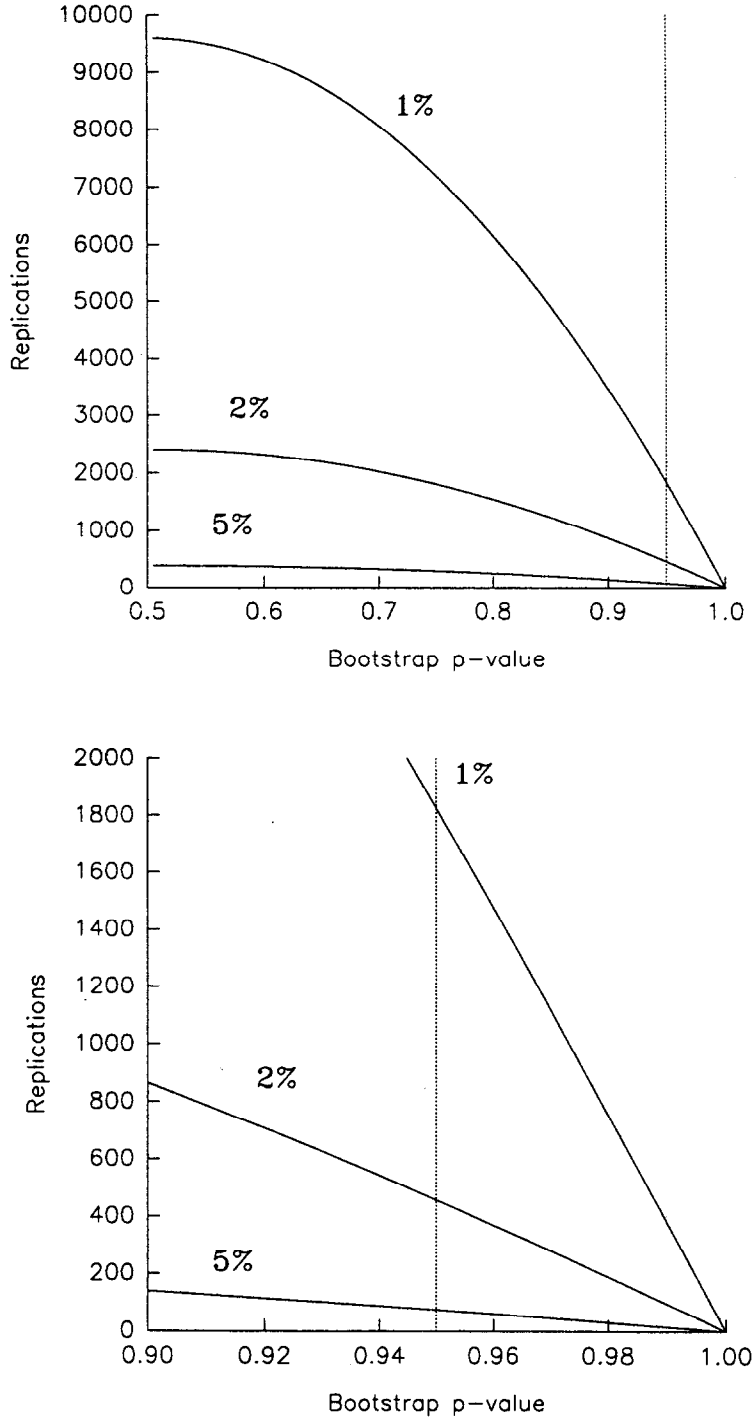

FIG. 1.—Accuracy of bootstrap P value (±1%, ±2%, and ±5%) vs. number of bootstrap replications, based on binomial variance. The dashed line is the 0.95 P value (95% confidence limit). *Top,* Region spanning P values 0.5–1.0. Bottom, More detailed plot spanning P values 0.90–1.00.

The number of bootstrap replications needed was addressed recently by Efron (1987). He showed that, in the case of bootstrap confidence intervals, the coefficient of variation is substantial (9%) with only 200 replications, whereas it decreases to 4% with 1,000 replications (the number of replications needed for calculating the bootstrap standard error is considerably fewer, ~100, but this statistic is of limited value in phylogeny estimation). Although Efron (1987) recommended that "on the order of 1000" replications are needed, it has been shown here that the actual phylogenetic BP may be over- or underestimated by 1%–2% in the region of the 0.95 BP with 1,000 replications. Some would consider this an acceptable error, but it would mean that the researcher would be unable to state (validly) that a group supported by a 0.96–0.97 BP is statistically significant. With 2,000 replications, such a statement can be made.

The typical size of data sets used in phylogeny estimation will almost certainly increase as more sequences become available. This may place a computational constraint on the number of bootstrap replications possible in large data sets. However, the bootstrap method still can be used even when only a small number of replications is feasible, as long as the variance of the BP is taken into consideration when one is drawing conclusions. BPs with ±6% accuracy can be obtained with only 50 replications (in the 0.95 region), and, although this error is too high for statistical testing, it can provide a reasonable indication of relative stability of groups within a phylogenetic tree, especially if no other statistical methods are available.

Only one aspect of the bootstrap method has been considered here: the number of replications necessary for accuracy. Other limitations of this method must be considered in any application. As noted by Felsenstein (1985), a substantial lack of independence of characters within the data set may require an adjustment in the sampling method, such as sampling fewer than the total number of characters randomly, with replacement, from the total number of characters. Also, the BP associated with a node reflects only that particular data set and clustering method. For example, either high levels (e.g., >50%) of sequence divergence or considerable rate variability among lineages may lead to statistical inconsistency with most methods of tree construction (Felsenstein 1983, 1988; Li and Gouy 1990). In these cases, bootstrapping could show statistically significant support for an incorrect topology (Nei 1991). If these limitations are kept in mind, the bootstrap method can be a simple and effective means of evaluating the results of phylogenetic analysis.

Acknowledgments

I thank Linda R. Maxson (L.R.M.) and Masatoshi Nei for use of research facilities; Sudhir Kumar, Tatsuya Ota, and an anonymous reviewer for many helpful suggestions; and especially Masatoshi Nei for invaluable statistical advice. This work was supported by National Science Foundation grant BSR 8918926 to L.R.M. and S.B.H.

LITERATURE CITED

DOUGLAS, S. E., C. A. MURPHY, D. F. SPENCER, and M. W. GRAY. 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature **350**:148–151.

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

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

———. 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. **82**:171–185.

FELSENSTEIN, J. 1983. Statistical inference of phylogenies. J. R. Stat. Soc. **A146**:246–272.

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

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

IRWIN, D. M., T. D. KOCHER, and A. C. WILSON. 1991. Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. **32**:128–144.

JANSEN, R. K., K. E. HOLSINGER, H. J. MICHAELS, and J. D. PALMER. 1990. Phylogenetic analysis of chloroplast DNA restriction site data at higher taxonomic levels: an example from the Asteraceae. Evolution **44**:2089–2105.

LI, W.-H., and M. GOUY. 1990. Statistical tests of molecular phylogenies. Methods Enzymol. **183**:645–659.

MEYER, A., and A. C. WILSON. 1990. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J. Mol. Evol. **31**:359–364.

NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. Pp. 90–128 *in* M. M. MIYAMOTO and J. L. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

OVENDEN, J. R., A. G. MACKINLAY, and R. H. CROZIER. 1987. Systematics and mitochondrial genome evolution of Australian Rosellas (Aves: Platycercidae). Mol. Biol. Evol. **4**:526–543.

THOMAS, R. H., W. SCHAFFNER, A. C. WILSON, and S. PAABO. 1989. DNA phylogeny of the extinct marsupial wolf. Nature **340**:465–467.

BRIAN CHARLESWORTH, reviewing editor

Received May 3, 1991; revision received July 12, 1991

Accepted September 9, 1991