Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12293–12297, December 1994
Evolution

Bootstrap hypothesis tests for evolutionary trees and other dendrograms

(phylogeny/population structure/plant evolution/5S rRNA/eyespot)

JAMES K. M. BROWN

Cereals Research Department, John Innes Centre, Colney Lane, Norwich, NR4 7UH, England

Communicated by John Maynard Smith, August 22, 1994

ABSTRACT **The bootstrap computer-intensive statistical technique is frequently applied to statistical analyses of phylogenetic trees. The widely used rule that a group is supported significantly if it appears in at least 95% of bootstrap trees is conservative in most situations. This paper describes three ways of using the bootstrap to carry out statistical inference on phylogenies. The first method tests whether there is nonrandom support for a single group or tree. The second method compares the support for two groups or trees. The third method tests whether a single group or tree has better support than the set of all possible alternatives; this may be a replacement for the "95% rule." These tests generally require fewer bootstrap trees to be estimated than do other methods of bootstrapping phylogenies. A simple, sequential statistical method can be used to increase the efficiency further. These methods can be applied to tests of multiple hypotheses about a single phylogeny. Parsimony analyses of 5S rRNA sequences of plants and cluster analyses of randomly amplified polymorphic DNA bands in three pathotypes of the cereal eyespot fungus are used as illustrative examples. The tests can be used to analyze dendrograms in subjects other than taxonomy.**

Phylogenetic trees classify organisms into groups of related species. If such taxonomic classifications are to reflect the course of evolution, each group should consist of species that have a common ancestor not shared by any species outside that group.

A phylogeny is estimated from only part of the total data that might possibly be sampled. How much confidence should be placed in the appearance of a set of species as a monophyletic group is therefore a statistical question (1, 2). Many statistical tests for phylogenies have been proposed (reviewed in refs. 2 and 3), but most of these methods do not actually test the hypothesis that a group of species is monophyletic. Of the few methods that do construct a confidence set of phylogenies, most are only applicable or computationally feasible for just a few taxa. Furthermore, most tests can only be applied to trees estimated by a particular algorithm from a particular kind of data (DNA or protein sequences, isozyme variation, etc.).

A useful approach to statistical inference for phylogenies employs the bootstrap, which is a computer-intensive statistical technique with many applications (ref. 4; introductory expositions in refs. 5 and 6). An application of the bootstrap to phylogenies was described by Felsenstein (7). This test can be used in conjunction with many kinds of data and many algorithms and is applicable to any number of species. Furthermore, it tests directly the hypothesis that a set of species is a monophyletic group (G). These features have made this test much the most widely used of all those available at present. Felsenstein suggested that the evidence for the existence of G should be considered to be significant only if G appeared in 95% or more of a sample of bootstrap trees (the "95% rule").

Simulation studies (8) and theory for small trees (9) indicate that the 95% rule is conservative in that the null hypothesis that G is not monophyletic is accepted more often than it should be. One reason for this is because the use of the 95% rule as a significance test confuses two different interpretations of the proportion of bootstrap trees (θ) in which G appears. θ estimates the probability that G would appear in a second tree, estimated from a data set similar to that actually studied. A significance test calculates the probability that a second set of data would deviate by at least as much as the observed data do from the values expected from a particular null hypothesis. These two probabilities are not the same (10).

The 95% rule is therefore inappropriate for two reasons. First, it measures the support for G by the observed value of θ but does not consider the distribution of θ under a specified null hypothesis. To carry out statistical inference, observed and expected values of θ should be compared. Second, the bootstrap is generally applied to phylogenies to test if G is better supported than the set of all possible alternatives to G. This is equivalent to asking whether or not there is significant evidence that G appears in more than 50% of all bootstrap trees. However, the 95% confidence interval for θ need not include the value 0.95, even if the estimate of θ is greater than 0.5 (8, 10). Note that the 95% rule is least conservative when the data are least informative (figure 9 in ref. 8). This is presumably because the variance of θ increases as the information in the data decreases, so that the confidence interval of θ becomes broader.

Despite these problems, the broad applicability of the bootstrap makes it a particularly valuable technique in taxonomy. Bootstrap tests for phylogenies, based on conventional statistical hypothesis tests, would therefore be desirable. This paper proposes some such tests, which are more efficient and less conservative than existing methods.

The problem of defining distinct groups arises in other natural and social sciences as well as taxonomy. Typically, some form of cluster analysis is used to define nonoverlapping categories of objects. For instance, a population geneticist may find that a species is subdivided into several races and wish to test whether these subdivisions are truly distinct entities or overlap to some extent. Statistical tests for the existence of groups in a phylogeny may therefore have a much wider range of applications.

DATA

Two sets of data are used as examples. One, an example from taxonomy, is a set of 5S rRNA sequences from plants (11). A subset of 26 species was analyzed (12); 5 green algae were used as an outgroup, while relationships were studied between 21 species. These were 1 charophyte, 4 bryophytes, 4

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: RAPD, randomly amplified polymorphic DNA bands.

pteridophytes, and 12 spermatophytes (seed plants), including 3 gymnosperms and 9 angiosperms. The aligned 5S rRNA sequence had 121 nucleotide positions. Maximum parsimony phylogenies were estimated by the DNAPARS program in the C version 3.52 of the PHYLIP package (7, 13), run on a VAX 4000-300 computer. The 5 algae were forced to be an outgroup by adding an additional, heavily weighted character. Parsimony was used to estimate phylogenies, in preference to maximum likelihood, to minimize the use of computer resources while the methods described in this paper were being developed. Species were added to the phylogeny in three random orders, using the "Jumble" option of DNAPARS. When more than one tree were equally the most parsimonious, the first tree printed was used in further analysis.

The second data set relates to the distinction of subspecific races. It includes 23 isolates of the fungus *Pseudocercosporella herpotrichoides*, which causes the eyespot disease of grasses. Of the 23 isolates, 11 were pathogenic on wheat only (W-type), while 8 were pathogenic on rye (R-type) and 4 were pathogenic on couch grass (C-type) as well as wheat (14). The data are sizes of 74 randomly amplified polymorphic DNA bands (RAPDs) generated by 18 primers (15). The similarity of each pair of isolates was calculated by scoring 1 for each band that they both possessed or both lacked and 0 for each band that only one of the two isolates had. Dendrograms were formed by average linkage cluster analysis using the GENSTAT 5 statistical package (16).

METHODS AND RESULTS

Three methods are described in this paper. As in Felsenstein's bootstrap test (7), these tests can be applied to the question of whether a single group of species is monophyletic. (Note that the tips of the tree are described as species, but they could equally be other taxonomic units or even individual organisms.) However, they are equally applicable to testing the support for other features of a phylogeny, such as the existence of several groups or of a particular sequence of branching. G is therefore redefined as either a monophyletic group or as a set of trees that share a common characteristic, as appropriate to the question in hand.

Test 1: Support for a Single Group

The first method tests whether or not the data provide significant support for the existence of a group or topology, G. The method tests a null hypothesis, H_0, that G does not exist, against H_1, that it does. This is done by examining whether the observed data give better support to G than random data would.

These hypotheses can be tested by comparing the frequencies of G in trees estimated from bootstrap and randomized data sets. In summary (Fig. 1): (*i*) the data, D, are arranged as a matrix of c characters × s species; (*ii*) a bootstrap data set, D^*, is obtained by sampling c characters, with replacement from the c characters in D; each species retains its original value of each character in D^* (Fig. 1); (*iii*) a phylogeny, T^*, is estimated, by applying an algorithm, A, to D^*; and (*iv*) steps *ii* and *iii* are repeated as many times as required.

For the bootstrap to be valid, the characters being resampled should be independent of one another and identically distributed (iid). The implications of this assumption for molecular sequences were described by Felsenstein (7). A statistical distribution based on bootstrap sampling is an approximation to the distribution that would be obtained by drawing repeated samples of characters from the population itself (4–6). T^* phylogenies are therefore estimates of the trees that would be obtained by sampling a second set of c characters from the same population as the c characters in D. D^* can therefore be used to represent repeated samples of data.

```
Original data
   Base abcde fghij klmno pqrst uvwxy
Species
   1    ATACC AGCAC TAGAG CACCG GATCT
   2    ATACC AACAC TAGGG CACCG GATCT
   3    ATACC GGGAT CGGGG CTTTG AGTCC
   4    ATACC CGAAA CGGGA TTTGA GCTCC

Bootstrap data
        abbdg jjjlm mqstt uuvvw wxyyy
   1    ATTCG CCCAG GACGG GGAAT TCTTT
   2    ATTCA CCCAG GACGG GGAAT TCTTT
   3    ATTCG TTTGG GTTGG AAGGT TCCCC
   4    ATTCG AAAGG GTGAA GGCCT TCCCC

Randomised data
        abbdg jjjlm mqstt uuvvw wxyyy
   1    ATTCG ACCGG GTCGA AGAAT TCCTC
   2    ATTCG CTTAG GACGG GGCAT TCTCC
   3    ATTCA TCAAG GAGAG GAGGT TCTTT
   4    ATTCG CACGG GTTGG GGACT TCCCT
```

FIG. 1. Sampling of bootstrap and randomized data from an original data matrix.

Randomized data sets are constructed from bootstrap data sets by the method of Archie (17): (*i*) the values of each of the c characters in D^* are permuted among the s species, to make a new data set, $D^\dagger$ (Fig. 1), such that any residual taxonomic information in $D^\dagger$ is randomized among the species; and (*ii*) the same algorithm, A, as was applied to D^*, is then used to estimate a phylogeny, $T^\dagger$, from $D^\dagger$.

This method of randomization is used because it compares each T^* with a corresponding $T^\dagger$. Any tendency of A to estimate a tree with a particular topology, given certain frequencies of characters in the data set, is thus consistent for D^* and $D^\dagger$. For instance, the frequencies of nucleotides in a sequence may influence the topology of an estimated phylogeny (18).

The frequencies of G in the sets of bootstrap and random trees ($\{T^*\}$ and $\{T^\dagger\}$, respectively) are θ^* and $\theta^\dagger$, respectively. When a pair of trees, (T^* and $T^\dagger$) is examined for the presence of G, there are four possible outcomes (Table 1). p is the proportion of untied pairs that are $Y^*N^\dagger$:

$$p = \frac{\theta^*(1 - \theta^\dagger)}{\theta^*(1 - \theta^\dagger) + (1 - \theta^*)\theta^\dagger}. \quad [1]$$

The difference between θ^* and $\theta^\dagger$ can be analyzed by a sign test (19) of the numbers of the two types of untied pair. Tied pairs are ignored because they do not indicate whether D^* or $D^\dagger$ provides more support for G.

The appropriate sign test is one-sided because the question is whether or not θ^* is greater than $\theta^\dagger$. As explained in the next paragraph, the number of D^* to be sampled, n, must be chosen by the investigator. When n is fixed, the sign test compares the support for G being equally frequent in $\{T^*\}$ and $\{T^\dagger\}$ so that $\theta^* = \theta^\dagger$, with the support for θ^* being greater than that for $\theta^\dagger$ by a certain amount. Formally, the hypotheses are

Table 1. Appearance of a group of species or tree topology, G, in bootstrap and randomized phylogenies (T^* and $T^\dagger$, respectively)

G in phylogenies				
T^*	$T^\dagger$	Result	Probability	Type of pair
Yes	Yes	$Y^*Y^\dagger$	$\theta^*\theta^\dagger$	Tied
Yes	No	$Y^*N^\dagger$	$\theta^*(1 - \theta^\dagger)$	Untied
No	Yes	$N^*Y^\dagger$	$(1 - \theta^*)\theta^\dagger$	Untied
No	No	$N^*N^\dagger$	$(1 - \theta^*)(1 - \theta^\dagger)$	Tied

$$H_0: p \leq 0.5,$$

$$H_1: p \geq p_1, \quad [2]$$

where p_1 is a value of p chosen by the investigator, as described below.

n must be chosen because there is no upper limit on the number of D^* data sets that can be sampled. As n increases, the power, $1 - \beta$ (the probability of rejecting a false H_0), tends to 1. However, the probability of rejecting a true H_0 also rises. Sampling more trees to increase $1 - \beta$ also reduces the test's efficiency. Furthermore, as n rises, smaller differences between θ^* and $\theta^\dagger$ can be detected. For any value of n, H_0 is more likely to be rejected at larger true values of p. p_1 is therefore the smallest value of p that the investigator wishes to declare as indicating a significant difference between θ^* and $\theta^\dagger$; p_1 must also be chosen.

The choice of n therefore depends on the significance level, α (i.e., the probability of rejecting a true H_0), $1 - \beta$, and p_1. From the formula for the power of the sign test (chapter 32 in ref. 20),

$$n = \left(\frac{w_\beta p_1^{0.5}(1 - p_1)^{0.5} + 0.5 w_\alpha}{p_1^{-0.5}} \right)^2, \quad [3]$$

where w_α and w_β are the standardized normal deviates for upper-tail probabilities of α and β. In practice, n is the next integer above the value calculated by Eq. 3. This expression is slightly conservative for small n, but only slightly improved efficiency is gained by calculating n exactly. n is tabulated for some values of α, β, and p_1 (Table 2).

When n samples are drawn, then for given α, the rejection number, r, is the number of Y*N† pairs that must be sampled for H_0 to be rejected. When at least $(n - r) + 1$ N*Y† pairs are sampled, H_0 is accepted. Some values of r are listed in Table 2, and more extensive tables are available (19).

The efficiency of the test can be improved by a simple sequential testing procedure. Once r Y*N† pairs have been sampled, the test can be stopped and H_0 rejected, since H_0 cannot now possibly be accepted. Similarly, H_0 can be accepted as soon as $(n - r) + 1$ N*Y† pairs have been drawn. This simple curtailment of the test does not affect its power or significance level (21) and may produce savings in the use of computing resources and the investigator's time.

To set p_1, a value of θ^* must be chosen by the investigator. $\theta^\dagger$ must be estimated, either from theoretical models of evolution (22, 23) or by examining the frequency of G and similar groups or trees in $\{T^\dagger\}$. The later method is generally preferable and is used in this paper, since phylogeny estimation programs may not produce trees in the frequencies expected from theory.

If both $\theta^\dagger$ and θ^* are very small, many D^* data sets might be sampled, but insufficient untied pairs might be obtained for H_0 to be accepted or rejected. In this case, it would be reasonable to decide before starting the test that, should a certain number, t, of D^* be sampled but neither r Y*N† pairs nor $(n - r) + 1$ N*Y† pairs be obtained, sampling should be terminated. This inevitably results in a loss of power, but it might reasonably be felt that, if θ^* were so small that no decision could be made before t trees were drawn, any support for G would be too low to be of genuine interest.

Table 2. Number of samples (n) to achieve power $1 - \beta$ in sign tests and rejection numbers (r) for several p_1

p_1	Tails of test	$\beta = 0.05$		$\beta = 0.01$		$\beta = 0.001$	
		n	(r)	n	(r)	n	(r)
0.75	1	38	(25)	76	(49)	134	(85)
0.75	2	46	(30)	85	(55)	143	(92)
0.90	1	11	(9)	22	(17)	39	(30)
0.90	2	14	(12)	25	(19)	42	(32)
0.95	1	7	(7)	14	(12)	25	(21)
0.95	2	9	(8)	16	(14)	27	(23)
0.99	1	5	(5)	9	(9)	15	(14)
0.99	2	6	(6)	10	(10)	16	(15)

Significance level $\alpha = \beta$ for one-tailed tests and $\alpha = \beta/2$ for two-tailed tests.

Examples. For the plant 5S rRNA data, D^* data sets were drawn from the 121 positions in the aligned sequence. Random numbers required for bootstrap sampling were generated by subroutine G05FAF in the NAG Fortran Library, version 15 (Numerical Algorithms Group, Oxford). One $D^\dagger$ was drawn from each D^* by permuting the bases in each position in D^* among the 21 species studied, using the NAG G05EHF subroutine. (Randomization and permutation can also be done by the SEQBOOT programme in PHYLIP.) The same random number seed for the Jumble option in DNAPARS was used for corresponding D^* and $D^\dagger$, so that the same input order of species was used in both cases.

Four tests are used to illustrate features of the method. In each case, a group or topology of interest, G, is defined, and the null hypothesis, H_0, that G does not exist, is tested by examining whether $\theta^* \leq \theta^\dagger$—i.e., whether the observed data give no better support to G than random data do.

(*i*) *Spermatophytes as a group.* Of the 21 plants in the data set, 12 are spermatophytes (seed plants). G was any tree in which these 12 were a monophyletic group. $\theta^\dagger$ was extremely small; the probability that, if a $T^\dagger$ had a clade of 12 species (which was not always the case), that clade would consist of the 12 spermatophytes is $\binom{21}{12}^{-1} = 3.4 \times 10^{-6}$. Even if p_1 were as high as 0.99, θ^* could be at least as low as 3.4×10^{-4} [Eq. 1]. Therefore, $r = 5$ for $\alpha = \beta = 0.05$, 9 for $\alpha = \beta = 0.01$, and 14 for $\alpha = \beta = 0.001$ (Table 2). These r values were reached after a total of 8, 13, and 20 samples respectively, with no N*Y† pair being observed; the remaining samples were N*Y† tied pairs. Therefore, H_0 was rejected for any reasonable value of θ^*. The data are consistent therefore with the uncontroversial hypothesis that spermatophytes are a natural group.

(*ii*) *Charophytes as a sister group of land plants.* This example is a situation in which $\theta^\dagger$ must be chosen more judiciously than in example *i*. The Charophyta, a division of freshwater algae, are thought to be the sister group of land plants (11, 24). To test this hypothesis, G was defined as any tree in which *Nitella flexilis*, a charophyte, is an outgroup of the 20 land plants in the data set. Fifteen of 20 $T^\dagger$ phylogenies had one species as an outgroup of the other 20, so $\theta^\dagger$ was estimated to be $0.75/21 = 0.036$. For the sake of this example, let H_0 be rejected when $\theta^* \geq 0.25$. With 0.04 as a conservative estimate of $\theta^\dagger$, $p_1 = 0.89$ (Eq. 1). By setting $p_1 = 0.9$, then $r = 9$ for $\alpha = \beta = 0.05$, 17 for $\alpha = \beta = 0.01$, and 30 for $\alpha = \beta = 0.001$ (Table 2). These numbers were reached after 13, 22, and 45 samples, respectively. The hypothesis that *N. flexilis* is a sister group of land plants was therefore supported.

(*iii*) *Pteridophytes and spermatophytes as a group.* In addition to 12 spermatophytes, the data include 4 pteridophytes. Of 100 samples, these 16 species were never observed as a group in either $\{T^*\}$ or $\{T^\dagger\}$. The test might reasonably be terminated at this point because any evidence that pteridophytes and spermatophytes together form a group might be too weak to be of interest. (Note that the choice of 100 samples is arbitrary; a higher or lower number can be used.)

(*iv*) *Wheat as an outgroup.* To illustrate the acceptance of an H_0, the unreasonable hypothesis that wheat is an outgroup of the other 20 plants was tested. θ^* and $\theta^\dagger$ had the same values as in example *ii*. For $\alpha = \beta = 0.05$, therefore, $n = 11$ and $r = 9$ (Table 2). Wheat was an outgroup in $\{T^\dagger\}$ in samples 5, 30, and 45 but in none of the first 45 T^* phylogenies. H_0 was therefore accepted.

For the eyespot RAPD data, D^* data sets were sampled from the set of 74 bands, and one $D^\dagger$ was drawn from each D^* by permuting the presence or absence of each band in D^* among the 23 isolates. Random numbers were generated by the URAND function in the GENSTAT 5 package.

The purpose of this study was to test whether or not the C-, R-, and W-type isolates were three distinct groups (15). G was therefore any tree in which they were indeed distinct. $\theta^\dagger$ was tiny; if $T^\dagger$ were to contain groups of 4, 8, and 11 isolates (which is unlikely in itself), the probability that these would consist of all C-types, all R-types, and all W-types, respectively, is $4!8!11!/23! = 1.5 \times 10^{-9}$. When using $p_1 = 0.99$, G appeared in 14 of the first 20 T^* bootstrap phylogenies but in none of the corresponding $T^\dagger$ random phylogenies. H_0 (null hypothesis) that the C, R, and W pathotypes are not all distinct groups was therefore rejected, with $\alpha = \beta = 0.001$ for any reasonable value of θ^*.

Test 2: Relative Support for Two Groups

If more than a few species are studied, the data may support many different groups significantly. Usually, some of these groups will be contradictory in that they cannot both be possible. Therefore, a test of the relative support for two alternative groups, G_1 and G_{-1}, would be useful.

A sign test can also be used here, to compare the frequencies of two different groups in bootstrap trees. Although a χ^2 test of the goodness of fit of the frequencies to a 1:1 distribution would be more efficient, the calculation of power is considerably simpler for a sign test.

The frequencies of G_1 and G_{-1} in $\{T^*\}$ are θ_1 and θ_{-1} respectively, such that $\theta_1 + \theta_{-1} = 1$. Only trees containing G_1 or G_{-1} are included in the analysis, the remainder being ignored. There are therefore no tied pairs, so $p = \theta_1$. n is approximately

$$n = \left(\frac{w_\beta p_1^{0.5}(1 - p_1)^{0.5} + 0.5 w_{\alpha/2}}{p_1^{-0.5}} \right)^2 \quad [4]$$

(20). The hypotheses for a two-tailed sign test are

$$H_0: p = 0.5,$$
$$H_1: p > p_1, \quad [5]$$
$$H_{-1}: p < p_{-1},$$

with the assumption that G_1 is more frequent than G_{-1} (i.e., $\theta_1 > \theta_{-1}$). Therefore, H_0 is that neither tree is supported significantly over its alternative, H_1 is that G_1 is supported over G_{-1}, and H_{-1} is that G_{-1} is supported over G_1. n and r are given in Table 2 for some values of α, β, and p_1. H_0 is rejected in favor of H_1 and of H_{-1} when at least r G_1 or r G_{-1} are observed, respectively. H_0 is accepted otherwise. As in Test 1, curtailment can be used to improve efficiency without reducing the power or significance level.

Examples. In the examples below, $\theta_1 = 0.9$ and $\theta_{-1} = 0.1$. The tests are therefore of whether or not the data support G_1 at least 9 times as well as G_{-1} or vice versa, with $\alpha = \beta = 0.05$, $n = 14$ (Eq. 4), and $r = 12$ (Table 2).

(*i*) *Test of two hypotheses about the relationships between angiosperms and gymnosperms.* G_1 is any tree in which angiosperms are a monophyletic group descended from gymnosperms in such a way that gymnosperms are not monophyletic. In G_{-1}, angiosperms and gymnosperms are sister groups. Trees in which spermatophytes are not monophyletic or which do not fit the criteria for G_1 or G_{-1} are ignored. Of the first 21 T^* bootstrap phylogenies sampled, 8 contained G_1 while 3 had G_{-1}. H_0 was accepted, since neither alternative hypothesis could be accepted. Therefore, there is insufficient information in the 5S rRNA sequence for at least 9 times more support to be given to one tree than to the other.

(*ii*) *Comparison of two hypotheses about the C-, R-, and W-types of eyespot.* Considering only trees in which all three are distinct groups, G_1 is any tree in which R-types are a sister group of C- and W-types, while in G_{-1}, W-types are a sister group of C- and R-types. Hypothesis H_1, that G_1 is at least 9 times better supported than G_{-1}, was accepted after 12 T^* bootstrap phylogenies with G_1 but none with G_{-1} had been drawn. A total of 21 samples were drawn, 9 of which had neither G_1 nor G_{-1}. This result is consistent with the reproductive biology of *P. herpotrichoides*, since C- and W-type isolates can be crossed with one another, but attempts to cross either with R-types have not been successful (25, 26).

Test 3: Support for a Group over All Possible Alternatives

The aim of most taxomonists who use the 95% rule is to test whether or not a group of species is supported over all possible alternatives. A possible alternative to this rule uses a similar method to Test 2. In this case, G_1 is the group of interest, while G_{-1} is any group that includes G_1 and other species so that G_1 is not monophyletic. Alternatively, G_1 might be a tree that has certain characteristics, and G_{-1} be any other tree. The only difference between this test and Test 2 is that a one-tailed test should be used, since the question is whether or not G_1 is better supported than G_{-1}. All T^* bootstrap phylogenies are used in the analysis, since all contain either G_1 or G_{-1}.

As in Test 2, $p = \theta_1$. n is given by Eq. 3, and the hypotheses are given by Eq. 2, since the test is one-sided. Any value of θ_1 between 0.5 and 1 can be chosen. A smaller θ_1 allows H_0 to be rejected when the true frequency of G_1 is closer to 0.5, but lower values of θ_1 require more samples. The two tests shown here use $\alpha = \beta = 0.05$ and $\theta_1 = 0.75$, so $n = 38$ (Eq. 3) and $r = 25$ (Table 2).

In the first example, G_1 was seed plants. The hypothesis that G_1, being monophyletic, was supported over all alternatives was accepted after 32 samples had been drawn. At that point, 25 T^* phylogenies had G_1 and 7 did not. In the second example, G_1 was any tree in which C-, R-, and W-type eyespot isolates were distinct groups. The hypothesis that G_1 was supported over all alternatives was rejected after 35 samples because at that point 13 T^* phylogenies did not contain G_1. This does not mean that there is no evidence for this tree (see Test 1) but merely that the data are insufficiently informative for the support for this tree to be significantly greater than the total support for all other trees, given the chosen value of θ_1.

DISCUSSION

This paper describes ways in which the bootstrap can be used to construct statistical tests for phylogenies. By comparing specified null and alternative hypotheses, they permit inferences about whether or not groups of species are significantly supported. The methods allow tests for tree topologies as well as groups of species and can be applied to other dendrograms as well as phylogenies.

Hypothesis Tests. Test 1 examines the question posed by Felsenstein (7), whether the data give significant support to a group of species being monophyletic. This is done by testing whether the observations provide more support to a group or topology (G) than random data would. This test, of whether or not the observations are consistent with a particular distribution of the data, is akin to a conventional hypothesis test, such as whether or not a sample could have been drawn from a population with a particular mean value of a variable.

Since this method only tests whether or not the data support a single G, it may indicate better-than-random support for a number of trees that differ somewhat from the one

that is best supported. In itself, therefore, Test 1 is not sufficient to indicate that a group or tree is supported to the exclusion of all other groups or trees.

Test 3 is more ambitious in testing the support for G over all possible alternatives. Summarizing the evidence for G by its observed frequency in $\{T^*\}$ alone and declaring support to be significant when this frequency is at least 95% are inappropriate (refs. 8–10; also see the Introduction). Test 3 takes a different approach to analyzing the evidence for G—namely, by testing statistically whether or not its frequency in $\{T^*\}$ exceeds 0.5. This hypothesis-testing method is considerably more efficient than attempting to calibrate the bootstrap to relate the frequency of G in $\{T^*\}$ to the probability that G is true (27).

The hypothesis examined by Test 3 is more restrictive than that of Test 1. The eyespot example shows that Test 3 may not reject all possible alternatives to G, even if Test 1 indicates better-than-random support for G. This may occur when the data are insufficiently informative to give significant support to one single group over all alternatives.

Test 2 examines whether one of two different, possibly contradictory groups or topologies is significantly better supported than the other. This method may be particularly useful when Test 1 shows that both groups have significant support, but Test 3 supports neither over all of the possible alternatives.

A sequential form of the sign test, using curtailment, may improve the tests' efficiency. The examples given above show that a curtailed test is little more efficient than the noncurtailed version when H_1 is true and α and β are high, but the saving in the number of trees estimated can be high when H_0 is true or α or β is low. More sophisticated sequential tests are available (28), but their implementation is generally rather more complex than that of simple curtailment.

In many papers on evolution, a phylogeny is shown, and the frequency with which every group in that tree appeared in $\{T^*\}$ is given. This procedure is equivalent to simultaneously carrying out tests of numerous different hypotheses. Furthermore, since bootstrap frequencies are shown only for groups that appear in the phylogeny, the hypotheses tested are, in effect, only chosen after the data have been inspected. The statistical problems of multiple hypothesis tests and of testing hypotheses suggested by the data are well known. To avoid the latter problem, it would be more satisfactory if the hypotheses to be tested were chosen before the data were analyzed.

The former problem, of multiple hypothesis tests, is a long-standing difficulty in the analysis of phylogenies (7). The main problem is that, given a finite set of species, no two groups are completely independent of one another. Although the hypotheses cannot be independent, they can still be tested independently by drawing a different set of bootstrap samples to test each hypothesis. In this paper, the six tests using the plant 5S rRNA data used six different bootstrap samples; a similar approach was taken with the eyespot RAPD data. When several hypotheses are tested, the true significance level—i.e., the probability of a Type I error—will be higher than the quoted value of α. As with other cases of multiple, independent tests, this can be corrected either by using a lower α or, more rigorously, by using sequential Bonferroni tests (e.g., ref. 29, which describes a less conservative test than that in ref. 30).

Validity of the Bootstrap. Although I have described hypothesis tests based on Felsenstein's method of bootstrapping phylogenies (7), the use of the bootstrap in phylogeny estimation lacks rigorous justification. In particular, three areas require further study if the validity of bootstrapping is to be established.

First, phylogeny estimation is only reliable when the algorithm used is consistent and unbiased (1, 2). The frequency of a group in $\{T^*\}$ is therefore only meaningful when such an algorithm is used. The base composition of the sequence may also affect the bootstrap frequency of a group produced by some algorithms (31).

Secondly, the bootstrap relies on the assumption that the resampled units are independent and identically distributed. This assumption is unlikely to be wholly valid for bases in the sequence of a functional gene. Clearly, some pairs of bases in 5S rRNA cannot be independent of one another because compensatory double mutations can occur in stem-and-loop structures (12). However, there are similar problems with other genes, since the function of the protein product depends on its tertiary structure, which in turn depends on the primary sequence. Similarly, RAPD bands, isozymes, and other molecular markers are only independent if they are linked neither to each other nor to genes that are under selection. This assumption is rarely tested. The effects of correlation of resampled units on the outcome of bootstrap methods requires analysis (7).

More generally, the bootstrap assumes that the statistic to be analyzed is a smooth function of the distribution of the variable that is resampled (4, 5). This is true for real-valued statistics such as the mean and correlation coefficient, but in a phylogeny, the group or topology G can have only one of two states, being present or absent. Making inferences about phylogenies by bootstrapping, by methods described here or elsewhere (7, 8, 27, 32), requires the theory of the bootstrap to be extended to binary-valued statistics. This has yet to be done.

I thank Paul Nicholson for kindly providing the eyespot data and David Balding, Noel Ellis, Chris Howe, and Richard Nichols for helpful comments on drafts of this paper. This research was supported by the Ministry of Agriculture, Fisheries and Food.

1. Felsenstein, J. (1983) *J. R. Statist. Soc.* **146**, 246–272.
2. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22**, 521–565.
3. Li, W. H. & Gouy, M. (1990) *Methods Enzymol.* **183**, 645–659.
4. Efron, B. (1979) *Ann. Statist.* **7**, 1–26.
5. Diaconis, P. & Efron, B. (1983) *Sci. Am.* **248**(5), 96–108.
6. Efron, B. & Tibshirani, R. (1986) *Statist. Sci.* **1**, 54–75.
7. Felsenstein, J. (1985) *Evolution* **39**, 783–791.
8. Hillis, D. M. & Bull, J. J. (1993) *Syst. Biol.* **42**, 182–192.
9. Zharkikh, A. & Li, W. H. (1992) *Mol. Biol. Evol.* **9**, 1119–1147.
10. Felsenstein, J. & Kishino, H. (1993) *Syst. Biol.* **42**, 193–200.
11. Hori, H., Lim, B.-L. & Osawa, S. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 820–823.
12. Steele, K. P., Holsinger, K. E., Jansen, R. K. & Taylor, D. W. (1991) *Mol. Biol. Evol.* **8**, 240–248.
13. Felsenstein, J. (1989) *Cladistics* **5**, 164–166.
14. Scott, P. R. & Hollins, T. W. (1980) *Ann. Appl. Biol.* **94**, 297–300.
15. Nicholson, P. & Rezanoor, H. N. (1994) *Mycol. Res.* **98**, 13–21.
16. GENSTAT 5 Committee (1993) *GENSTAT 5 Release 3 Reference Manual* (Oxford Univ. Press, Oxford).
17. Archie, J. W. (1989) *Syst. Zool.* **38**, 239–252.
18. Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J. & Larkum, A. W. D. (1992) *J. Mol. Evol.* **34**, 153–162.
19. Conover, W. J. (1980) *Practical Nonparametric Statistics* (Wiley, New York), 2nd Ed., pp. 122–128.
20. Kendall, M. G. & Stuart, A. (1979) *The Advanced Theory of Statistics: Inference and Relationship* (Charles Griffin, London), Vol. 2, 4th Ed.
21. Fleming, T. R. (1982) *Biometrics* **38**, 143–151.
22. Rohlf, F. J. (1982) *Math. Biosci.* **59**, 131–144.
23. Brown, J. K. M. (1994) *Syst. Biol.* **43**, 78–91.
24. Graham, L. E. & Kaneko, Y. (1991) *Crit. Rev. Pl. Sci.* **10**, 323–342.
25. Dyer, P. S., Bateman, G. L., Lucas, J. A. & Peberdy, J. S. (1994) *Ann. Appl. Biol.* **125**, in press.
26. Nicholson, P., Dyer, P. S., Rezanoor, H. N., Lucas, J. A., Walkowiak-Cagara, I. & Hollins, T. W. (1994) *Mycol. Res.*, in press.
27. Rodrigo, A. G. (1993) *Intl. J. Parasitol.* **23**, 507–514.
28. Xu, D.-Z. (1990) *Computer Analysis of Sequential Medical Trials* (Horwood, Chichester, U.K.).
29. Rom, D. M. (1990) *Biometrika* **77**, 663–665.
30. Rice, W. R. (1989) *Evolution* **43**, 223–225.
31. Steel, M. A., Lockhart, P. J. & Penny, D. (1993) *Nature (London)* **364**, 440–442.
32. Sanderson, M. J. (1989) *Cladistics* **5**, 113–129.