Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7085–7090, July 1996
Evolution

Bootstrap confidence levels for phylogenetic trees

BRADLEY EFRON, ELIZABETH HALLORAN‡, AND SUSAN HOLMES†§

†Department of Statistics, Stanford University, Stanford, CA 94305; and ‡Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322

Contributed by Bradley Efron, January 26, 1996

ABSTRACT **Evolutionary trees are often estimated from DNA or RNA sequence data. How much confidence should we have in the estimated trees? In 1985, Felsenstein [Felsenstein, J. (1985) *Evolution* 39, 783–791] suggested the use of the bootstrap to answer this question. Felsenstein's method, which in concept is a straightforward application of the bootstrap, is widely used, but has been criticized as biased in the genetics literature. This paper concerns the use of the bootstrap in the tree problem. We show that Felsenstein's method is not biased, but that it can be corrected to better agree with standard ideas of confidence levels and hypothesis testing. These corrections can be made by using the more elaborate bootstrap method presented here, at the expense of considerably more computation.**

The bootstrap, as described in ref. 1, is a computer-based technique for assessing the accuracy of almost any statistical estimate. It is particularly useful in complicated nonparametric estimation problems, where analytic methods are impractical. Felsenstein (2) introduced the use of the bootstrap in the estimation of phylogenetic trees. His technique, which has been widely used, provides assessments of "confidence" for each clade of an observed tree, based on the proportion of bootstrap trees showing that same clade. However Felsenstein's method has been criticized as biased. Hillis and Bull's paper (3), for example, says that the bootstrap confidence values are consistently too conservative (i.e., biased downward) as an assessment of the tree's accuracy.

Is the bootstrap biased for the assessment of phylogenetic trees? We will show that the answer is no, at least to a first order of statistical accuracy. Felsenstein's method provides a reasonable first approximation to the actual confidence levels of the observed clades. More ambitious bootstrap methods can be fashioned to give still better assessments of confidence. We will describe one such method and apply it to the estimation of a phylogenetic tree for the malaria parasite *Plasmodium*.

Bootstrapping Trees

Fig. 1 shows part of a data set used to construct phylogenetic trees for malaria. The data are the aligned sequences of small subunit RNA genes from 11 malaria species of the genus *Plasmodium*. The 11 × 221 data matrix we will first consider is composed of the 221 polytypic sites. Fig. 1 shows the first 20 columns of $\mathbf{x}$. There are another 1399 monotypic sites, where the 11 species are identical.

Fig. 2 shows a phylogenetic tree constructed from $\mathbf{x}$. The tree-building algorithm proceeds in two main steps: (*i*) an 11 × 11 distance matrix $\hat{D}$ is constructed for the 11 species, measuring differences between the row vectors of $\mathbf{x}$; and (*ii*) $\hat{D}$ is converted into a tree by a connection algorithm that connects the closest two entries (species 9 and 10 here), reduces $\hat{D}$ to a 10 × 10 matrix according to some merging rule, connects the two closest entries of the new D matrix, etc.

We can indicate the tree-building process schematically as

$$\mathbf{x} \rightarrow \hat{D} \rightarrow \widehat{\text{TREE}},$$

the hats indicating that we are dealing with estimated quantities. A deliberately simple choice of algorithms was made in constructing Fig. 2: $\hat{D}$ was the matrix of the Euclidean distances between the rows of $\mathbf{x}$, with (A, G, C, T) interpreted numerically as (1, 2, 5, 6), while the connection algorithm merged nodes by maximization. Other, better, tree-building algorithms are available, as mentioned later in the paper. Some of these, such as the maximum parsimony method, do not involve a distance matrix, and some use all of the sites, including the monotypical ones. The discussion here applies just as well to all such tree-building algorithms.

Felsenstein's method proceeds as follows. A bootstrap data matrix $\mathbf{x}^*$ is formed by randomly selecting 221 columns from the original matrix $\mathbf{x}$ *with replacement*. For example the first column of $\mathbf{x}^*$ might be the 17th column of $\mathbf{x}$, the second might be the 209th column of $\mathbf{x}$, the third the 17th column of $\mathbf{x}$, etc. Then the original tree-building algorithm is applied to $\mathbf{x}^*$, giving a bootstrap tree $\widehat{\text{TREE}}^*$,

$$x \rightarrow \hat{D}^* \rightarrow \widehat{\text{TREE}}^*,$$

This whole process is independently repeated some large number B times, $B = 200$ in Fig. 2, and the proportions of bootstrap trees agreeing with the original tree are calculated. "Agreeing" here refers to the topology of the tree and not to the length of its arms.

These proportions are the bootstrap confidence values. For example the 9-10 clade seen in Fig. 2 appeared in 193 of the 200 bootstrap trees, for an estimated confidence value of 0.965. Species 7-8-9-10 occurred as a clade in 199 of the 200 bootstrap trees, giving 0.995 confidence. (Not all of these 199 trees had the configuration shown in Fig. 2; some instead first having 8 joined to 9-10 and then 7 joined to 8-9-10, as well as other variations.)[2]

Felsenstein's method is, nearly, a standard application of the *nonparametric bootstrap*. The basic assumption, further discussed in the next section, is that the columns of the data matrix $\mathbf{x}$ are independent of each other and drawn from the same probability distribution. Of course, if this assumption is a bad one, then Felsenstein's method goes wrong, but that is not the point of concern here nor in the references, and we will take the independence assumption as a given truth.

The bootstrap is more typically applied to statistics $\hat{\theta}$ that estimate a parameter of interest θ, both $\hat{\theta}$ and θ being single numbers. For example, $\hat{\theta}$ could be the sample correlation coefficient between the first two malaria species, Pre and Pme, at the 221 sites, with (A, G, C, T) interpreted as (1, 2, 5, 6): $\hat{\theta} = 0.616$. How accurate is $\hat{\theta}$ as an estimate of the true correlation θ? The nonparametric bootstrap answers such questions without making distributional assumptions.

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.

§Permanent address: Biometrie–Institut National de la Recherche Agronomique, Montpellier, France.

Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Species	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1 Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
2 Pme (Lizard)	T	C	T	A	A	A	G	A	T	T	A	T	A	T	A	G	A	T	A
3 Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	A	A	T	T
4 Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
5 Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A	T	A	A	A	T	A
6 Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A	C	A	A	A	T	C
7 Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T	T	A	G	G	A	A
8 Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T	T	A	G	A	T	A
9 Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
10 Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	C	G	G	A	T	A
11 Pga(Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	C

FIG. 1. Part of the data matrix of aligned nucleotide sequences for the malaria parasite *Plasmodium*. Shown are the first 20 columns of the 11 × 221 matrix **x** of polytypic sites used in most of the analyses below. The final analysis of the last section also uses the data from 1399 monotypic sites.

Each bootstrap data set $\mathbf{x}^*$ gives a bootstrap estimate $\hat{\theta}^*$, in this case the sample correlation between the first two rows of $\mathbf{x}^*$. The central idea of the bootstrap is to use the observed distribution of the differences $\hat{\theta}^* - \theta$ to infer the unobservable distribution of $\hat{\theta} - \theta$; in other words to learn about the accuracy of $\hat{\theta}$. In our example, the 200 bootstrap replications of $\hat{\theta}^* - \theta$ were observed to have expectation 0.622 and standard deviation 0.052. The inference is that $\hat{\theta}$ is nearly unbiased for estimating θ, with a standard error of about 0.052. We can also calculate bootstrap confidence intervals for θ. A well-developed theory supports the validity of these inferences [see Efron and Tibshirani (1)].

Felsenstein's application of the bootstrap is nonstandard in one important way: the statistic $\widehat{\text{TREE}}$, unlike the correlation coefficient, does not change smoothly as a function of the data set **x**. Rather, $\widehat{\text{TREE}}$ is constant within large regions of the **x**-space, and then changes discontinuously as certain boundaries are crossed. This behavior raises questions about the bootstrap inferences, questions that are investigated in the sections that follow.

A Model For The Bootstrap

The rationale underlying the bootstrap confidence values depends on a simple multinomial probability model. There are $K = 4^{11} - 4$ possible column vectors for **x**, the number of vectors of length 11 based on a 4-letter alphabet, not counting the 4 monotypic ones. Call these vectors $X_1, X_2, \ldots, X_K$, and suppose that each observed column of **x** is an independent selection from $X_1, X_2, \ldots, X_K$, equaling X_k with probability π_k. This is the *multinomial model* for the generation of **x**.

Denote $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, so the sum of $\underset{\sim}{\pi}$'s coordinates is 1. The data matrix **x** can be characterized by the proportion of its $n = 221$ columns equaling each possible X_k, say

$$\hat{\pi}_k = \#\{\text{columns of } \mathbf{x} \text{ equalling } X_k\}/n,$$

with $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K)$. This is a very inefficient way to represent the data, since $4^{11} - 4$ is so much bigger than 221, but it is useful for understanding the bootstrap. Later we will see that only the vectors X_k that actually occur in **x** need be considered, at most n of them.

Almost always the distance matrix $\hat{D}$ is a function of the observed proportions $\hat{\underset{\sim}{\pi}}$, so we can write the tree-building algorithm as

$$\hat{\pi} \rightarrow \hat{D} \rightarrow \widehat{\text{TREE}}.$$

In a similar way the vector of true probabilities $\underset{\sim}{\pi}$ gives a true distance matrix and a true tree,

$$\pi \rightarrow D \rightarrow \text{TREE}.$$

D would be the matrix with ijth element $\{\Sigma_k \pi_k (X_{ki} - X_{kj})^2\}^{1/2}$ in our example, and TREE the tree obtained by applying the maximizing connection algorithm to D.

Fig. 3 is a schematic picture of the space of possible $\underset{\sim}{\pi}$ vectors, divided into regions $\mathcal{R}_1, \mathcal{R}_2, \ldots$. The regions correspond to different possible trees, so if $\pi \in \mathcal{R}_j$ the jth possible tree results. We hope that $\widehat{\text{TREE}} = \text{TREE}$, which is to say that $\underset{\sim}{\pi}$ and $\hat{\underset{\sim}{\pi}}$ lie in the same region, or at least that $\widehat{\text{TREE}}$ and TREE agree in their most important aspects.

The bootstrap data matrix $\mathbf{x}^*$ has proportions of columns say

$$\hat{\pi}_k^* = \#\{\text{columns of } \mathbf{x}^* \text{ equalling } X_k\}/n,$$

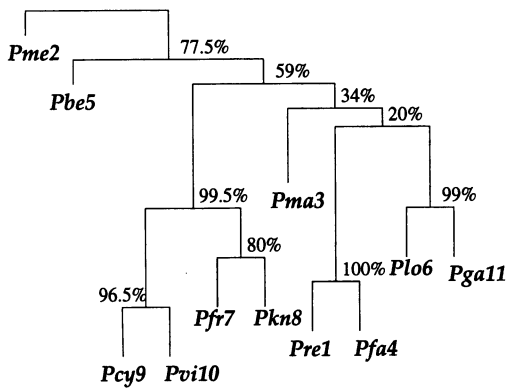


FIG. 2. Phylogenetic tree based on the malaria data matrix; species are numbered as in Fig. 1. The numbers at the branches are confidence values based on Felsenstein's bootstrap method. $B = 200$ bootstrap replications.

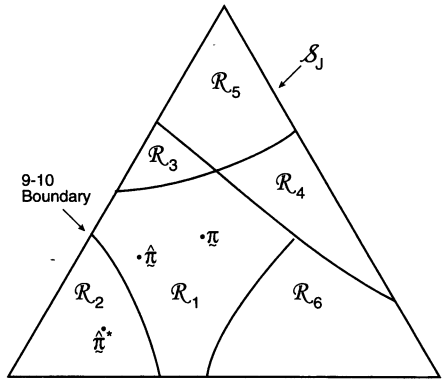


FIG. 3. Schematic diagram of tree estimation; triangle represents the space of all possible $\underset{\sim}{\pi}$ vectors in the multinomial probability model; regions $\mathcal{R}_1, \mathcal{R}_2 \ldots$ correspond to the different possible trees. In the case shown $\underset{\sim}{\pi}$ and $\hat{\underset{\sim}{\pi}}$ lie in the same region so TREE = $\widehat{\text{TREE}}$, but $\hat{\underset{\sim}{\pi}}^*$ lies in a region where $\widehat{\text{TREE}}^*$ does not have the 9-10 clade.

$\hat{\underline{\pi}}^* = (\hat{\pi}_1^*, \hat{\pi}_2^*, \ldots, \hat{\pi}_K^*)$. We can indicate the bootstrap tree-building

$$\hat{\underline{\pi}}^* \rightarrow \hat{D} \rightarrow \widehat{\text{TREE}},$$

The hypothetical example of Fig. 3 puts $\underline{\pi}$ and $\hat{\underline{\pi}}$ in the same region, so that the estimate $\widehat{\text{TREE}}$ exactly equals the true TREE. However $\hat{\underline{\pi}}^*$ lies in a different region, with $\widehat{\text{TREE}}^*$ not having the 9-10 clade. This actually happened in 7 out of the 200 bootstrap replications for Fig. 2.

What the critics of Felsenstein's method call its bias is the fact that the probability $\widehat{\text{TREE}}^* = \text{TREE}$ is usually less than the probability $\widehat{\text{TREE}} = \text{TREE}$. In terms of Fig. 3, this means that $\hat{\underline{\pi}}^*$ has less probability than $\hat{\underline{\pi}}$ of lying in the same region as $\underline{\pi}$. Hillis and Bull (3) give specific simulation examples. The discussion below is intended to show that this property is not a bias, and that to a first order of approximation the bootstrap confidence values provide a correct assessment of $\widehat{\text{TREE}}$'s accuracy. A more valid criticism of Felsenstein's method, discussed later, involves its relationship with the standard theory of statistical confidence levels based on hypothesis tests.

Returning to the corrrelation example of the previous section, it is *not* true that $\hat{\theta}^* - \theta$ (as opposed to $\hat{\theta}^* - \hat{\theta}$) has the same distribution as $\hat{\theta}^* - \theta$, even approximately. In fact $\hat{\theta}^* - \theta$ will have nearly twice the variance of $\hat{\theta} - \theta$, the sum of the variances of $\hat{\theta}$ around θ and of $\hat{\theta}^*$ around $\hat{\theta}$. Similarly in Fig. 3 the average distance from $\hat{\underline{\pi}}^*$ to $\underline{\pi}$ will be greater than the average distance from $\hat{\underline{\pi}}$ to $\underline{\pi}$. This is the underlying reason for results like those of Hillis and Bull, that $\hat{\underline{\pi}}^*$ has less probability than $\hat{\underline{\pi}}$ of lying in the same region as $\underline{\pi}$. However, to make valid bootstrap inferences we need to use the observed differences between $\widehat{\text{TREE}}^*$ and $\widehat{\text{TREE}}$ (not between $\widehat{\text{TREE}}^*$ and TREE) to infer the differences between $\widehat{\text{TREE}}$ and TREE. Just how this can be done is discussed using a simplified model in the next two sections.

A Simpler Model

The meaning of the bootstrap confidence values can be more easily explained using a simple normal model rather than the multinomial model. This same tactic is used in Felsenstein and Kishino (4). Now we assume that the data $\mathbf{x} = (x_1, x_2)$ is a two dimensional normal vector with expectation vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and identity covariance matrix, written

$$\mathbf{x} \sim N_2(\boldsymbol{\mu}, I).$$

In other words x_1 and x_2 are independent normal variates with expectations μ_1 and μ_2, and variances 1. The obvious estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{x}$, and we will use this notation in what follows. The $\boldsymbol{\mu}$-plane is partitioned into regions $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots$ similarly to Fig. 3. We observe that $\hat{\boldsymbol{\mu}}$ lies in one of these regions, say $\mathcal{R}_1$, and we wish to assign a confidence value to the event that $\boldsymbol{\mu}$ itself lies in $\mathcal{R}_1$.

Two examples are illustrated in Fig. 4. In both of them $\mathbf{x} = \hat{\boldsymbol{\mu}} = (4.5,0)$ lies in $\mathcal{R}_1$, one of two possible regions. Case I has $\mathcal{R}_2 = \{\boldsymbol{\mu} : \mu_1 \leq 3\}$, a half-plane, while case II has $\mathcal{R}_2 = \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq 3\}$, a disk of radius 3.

Bootstrap sampling in our simplified problem can be taken to be

$$\mathbf{x}^* \sim N_2(\hat{\boldsymbol{\mu}}, I).$$

This is a parametric version of the bootstrap, as in section 6.5 of Efron and Tibshirani (1), rather than the more familiar nonparametric bootstrap considered previously, but it provides the proper analogy with the multinomial model. The dashed circles in Fig. 4 indicate the bootstrap density of $\hat{\boldsymbol{\mu}}^* = \mathbf{x}^*$, centered at $\hat{\boldsymbol{\mu}}$. Felsenstein's confidence value is the bootstrap probability that $\hat{\boldsymbol{\mu}}^*$ lies in $\mathcal{R}_1$, say

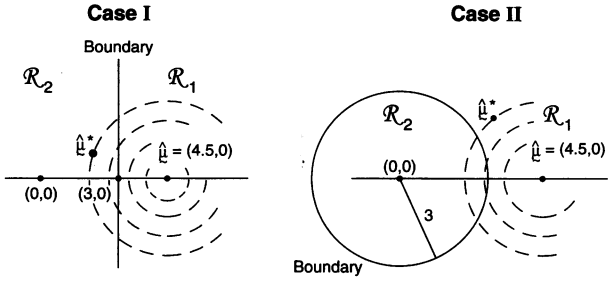


FIG. 4. Two cases of the simple normal model; in both we observe $\hat{\boldsymbol{\mu}} = (4.5, 0) \in \mathcal{R}_1$, and wish to assign a confidence value to $\boldsymbol{\mu} \in \mathcal{R}_1$. Case I, $\mathcal{R}_2$ is the region $\{\mu_1 \leq 3\}$. Case II, $\mathcal{R}_2$ is the region $\{\|\boldsymbol{\mu}\| < 3\}$. The dashed circles indicate bootstrap sampling $\hat{\boldsymbol{\mu}}^* \sim N_2(\hat{\boldsymbol{\mu}}, I)$.

$$\tilde{\alpha} = \text{Prob}_{\hat{\boldsymbol{\mu}}}\{\hat{\boldsymbol{\mu}}^* \in \mathcal{R}_1\}.$$

The notation $\text{Prob}_{\hat{\boldsymbol{\mu}}}$ emphasizes that the bootstrap probability is computed with $\hat{\boldsymbol{\mu}}$ fixed and only $\hat{\boldsymbol{\mu}}^*$ random. The bivariate normal model of this section is simple enough to allow the $\tilde{\alpha}$ values to be calculated theoretically, without doing simulations,

$$\tilde{\alpha}_{\text{I}} = 0.933 \quad \text{and} \quad \tilde{\alpha}_{\text{II}} = 0.949.$$

Notice that $\tilde{\alpha}_{\text{II}}$ is bigger than $\tilde{\alpha}_{\text{I}}$ because $\mathcal{R}_1$ is bigger in case II.

In our normal model, $\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}$ has the same distribution as $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$, both distributions being the standard bivariate normal $N_2(\mathbf{0}, \text{I})$. The general idea of the bootstrap is to use the observable bootstrap distribution of $\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}$ to say something about the unobservable distribution of the error $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$. Notice, however, that the marginal distribution of $\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}$ has *twice* as much variance,

$$\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu} \sim N_2(\mathbf{O}, 2I).$$

This generates the "bias" discussed previously, that $\hat{\boldsymbol{\mu}}^*$ has less probability than $\hat{\boldsymbol{\mu}}$ of being in the same region as $\boldsymbol{\mu}$. But this kind of interpretation of bootstrap results cannot give correct inferences. Newton (5) makes a similar point, as do Zharkikh and Li (6) and Felsenstein and Kishino (4).

We can use a Bayesian model to show that $\tilde{\alpha}$ is a reasonable assessment of the probability that $\mathcal{R}_1$ contains $\boldsymbol{\mu}$. Suppose we believe *apriori* that $\boldsymbol{\mu}$ could lie anywhere in the plane with equal probability. Then having observed $\hat{\boldsymbol{\mu}}$, the *aposteriori* distribution of $\boldsymbol{\mu}$ given $\hat{\boldsymbol{\mu}}$ is $N_2(\hat{\boldsymbol{\mu}}, I)$, exactly the same as the bootstrap distribution of $\hat{\boldsymbol{\mu}}^*$. In other words, $\tilde{\alpha}$ is the *aposteriori* probability of the event $\boldsymbol{\mu} \in \mathcal{R}_1$, if we begin with an "uninformative" prior density for $\boldsymbol{\mu}$.

Almost the same thing happens in the multinomial model. The bootstrap probability that $\widehat{\text{TREE}}^* = \widehat{\text{TREE}}$ is almost the same as the *aposteriori* probability that $\text{TREE} = \widehat{\text{TREE}}$ starting from an uninformative prior density on $\underline{\pi}$ [see section 10.6 of Efron (7)]. The same statement holds for any part of the tree, for example the existence of the 9-10 clade in Fig. 2. There are reasons for being skeptical about the Bayesian argument, as discussed in the next section. However, the argument shows that Felsenstein's bootstrap confidence values are at least reasonable and certainly cannot be universally biased downward.

Hypothesis-Testing Confidence Levels

Fig. 5 illustrates another more customary way of assigning a confidence level to the event $\boldsymbol{\mu} \in \mathcal{R}_1$. In both case I and case II, $\hat{\boldsymbol{\mu}}_0 = (3, 0)$ is the closest point to $\hat{\boldsymbol{\mu}}$ on the boundary

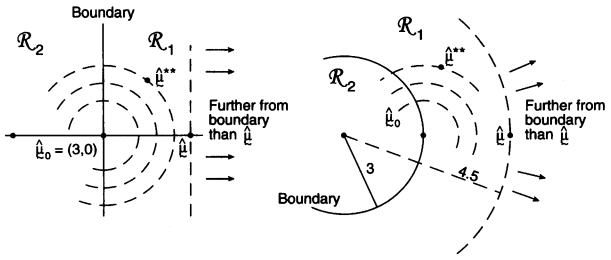


FIG. 5. Confidence levels of the two cases in Fig. 4; $\hat{\boldsymbol{\mu}}_0 = (3, 0)$ is the closest point to $\hat{\boldsymbol{\mu}} = (4.5, 0)$ on the boundary separating $\mathcal{R}_1$ from $\mathcal{R}_2$; bootstrap vector $\hat{\boldsymbol{\mu}}^{**} \sim N_2(\hat{\boldsymbol{\mu}}_0, I)$. The confidence level $\hat{\alpha}$ is the probability that $\hat{\boldsymbol{\mu}}^{**}$ is closer than $\hat{\boldsymbol{\mu}}$ to the boundary.

separating $\mathcal{R}_1$ from $\mathcal{R}_2$. We now bootstrap from $\hat{\boldsymbol{\mu}}_0$ rather than from $\hat{\boldsymbol{\mu}}$, obtaining bootstrap data vectors

$$\mathbf{x}^{**} \sim N_2(\hat{\boldsymbol{\mu}}_0, I).$$

[The double star notation is intended to avoid confusion with the previous bootstrap vectors $\mathbf{x}^* \sim N_2(\hat{\boldsymbol{\mu}}, I)$.] The *confidence level* $\hat{\alpha}$ for the event $\boldsymbol{\mu} \in \mathcal{R}_1$ is the probability that the bootstrap vector $\hat{\boldsymbol{\mu}}^{**} = \mathbf{x}^{**}$ lies closer than $\hat{\boldsymbol{\mu}}$ to the boundary. This has a familiar interpretation: $1 - \hat{\alpha}$ is the rejection level, one-sided, of the usual likelihood ratio test of the null hypothesis that $\boldsymbol{\mu}$ does *not* lie in $\mathcal{R}_1$. Here we are computing $\hat{\alpha}$ numerically, rather than relying on an asymptotic χ^2 approximation. In a one-dimensional testing problem, $1 - \hat{\alpha}$ would exactly equal the usual p value obtained from the test of the null hypothesis that the true parameter value lies in $\mathcal{R}_2$.

Once again it is simple to compute the confidence level for our two cases, at least numerically,

$$\hat{\alpha}_{\mathrm{I}} = 0.933 \quad \text{and} \quad \hat{\alpha}_{\mathrm{II}} = 0.914.$$

In the first case $\hat{\alpha}_{\mathrm{I}}$ equals $\tilde{\alpha}_{\mathrm{I}}$, the Felsenstein bootstrap confidence value. However $\hat{\alpha}_{\mathrm{II}} = 0.914$ is less than $\tilde{\alpha}_{\mathrm{II}} = 0.949$.

Why do the answers differ? Comparing Figs. 4 and 5, we see that, roughly speaking, the confidence value $\tilde{\alpha}$ is a probabilistic measure of the distance from $\hat{\boldsymbol{\mu}}$ to the boundary, while the confidence level $\hat{\alpha}$ measures distance from the boundary to $\hat{\boldsymbol{\mu}}$. The two ways of measuring distance agree for the straight boundary, but not for the curved boundary of case II. Because the boundary curves *away* from $\hat{\boldsymbol{\mu}}$, the confidence value $\tilde{\alpha}$ is increased from the straight-line case. However the set of vectors further than $\hat{\boldsymbol{\mu}}$ from the boundary curves *toward* $\hat{\boldsymbol{\mu}}_0$, which decreases $\hat{\alpha}$. We would get the opposite results if the boundary between $\mathcal{R}_1$ and $\mathcal{R}_2$ curved in the other direction.

The confidence level $\hat{\alpha}$, rather than $\tilde{\alpha}$, provides the more usual assessment of statistical belief. For example in case II let $\theta = \|\boldsymbol{\mu}\|$ be the length of the expectation vector $\boldsymbol{\mu}$. Then $\hat{\alpha} = 0.914$ is the usual confidence level attained for the event $\{\theta \geq 3\}$, based on observing $\hat{\theta} = \|\hat{\boldsymbol{\mu}}\| = 4.5$. And $\{\theta \geq 3\}$ is the same as the event $\{\boldsymbol{\mu} \in \mathcal{R}_1\}$.

Using the confidence value $\tilde{\alpha}$ as equivalent to assuming a flat Bayesian prior for $\boldsymbol{\mu}$. It can be shown that using $\hat{\alpha}$ amounts, approximately, to assuming a different prior density for $\boldsymbol{\mu}$, one that depends on the shape of the boundary. In case II this prior is uniform on polar coordinates for $\boldsymbol{\mu}$, rather than uniform on the original rectangular coordinates [see Tibshirani (8)].

The Relationship Between the Two Measures of Confidence

There is a simple approximation formula for converting a Felsenstein confidence value $\tilde{\alpha}$ to a hypothesis-testing confidence level $\hat{\alpha}$. This formula is conveniently expressed in terms of the cumulative distribution function $\boldsymbol{\Phi}(z)$ of a standard one-dimensional normal variate, and its inverse function $\boldsymbol{\Phi}^{-1}$: $\boldsymbol{\Phi}^{-1}(1.645) = 0.95$, $\boldsymbol{\Phi}^{-1}(0.95) = 1.645$, etc. We define the "$z$ values" corresponding to $\tilde{\alpha}$ and $\hat{\alpha}$,

$$\tilde{z} = \boldsymbol{\Phi}^{-1}(\tilde{\alpha}) \quad \text{and} \quad \hat{z} = \boldsymbol{\Phi}^{-1}(\hat{\alpha}).$$

In case II, $\tilde{z} = \boldsymbol{\Phi}^{-1}(0.949) = 1.64$ and $\hat{z} = \boldsymbol{\Phi}^{-1}(0.914) = 1.37$.

Now let $\hat{\boldsymbol{\mu}}^{**} \sim N_2(\hat{\boldsymbol{\mu}}_0, I)$ as in Fig. 5, and define

$$z_0 = \boldsymbol{\Phi}^{-1}(\mathrm{Prob}_{\hat{\boldsymbol{\mu}}_0}\{\hat{\boldsymbol{\mu}}^{**} \in R_1\}).$$

For case I it is easy to see that $z_0 = \boldsymbol{\Phi}^{-1}(0.50) = 0$. For case II, standard calculations show that $z_0 = \boldsymbol{\Phi}^{-1}(0.567) = 0.17$.

In normal problems of the sort shown in Figs. 4 and 5 we can approximate $\hat{z}$ in terms of $\tilde{z}$ and z_0:

$$\hat{z} \doteq \tilde{z} - 2z_0. \qquad \textbf{[1]}$$

Formula **1** is developed in Efron (9), where it is shown to have "second order accuracy." This means that in repeated sampling situations [where we observe independent data vectors $x_1, x_2, \ldots, x_n \sim N_2(\boldsymbol{\mu}, I)$ and estimate $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}} = \Sigma_i^n = 1\, x_i/n$] z_0 is of order $1/\sqrt{n}$, and formula **1** estimates $\hat{z}$ with an error of order only $1/n$.

Second-order accuracy is a large sample property, but it usually indicates good performance in actual problems. For case I, Eq. **1** correctly predicts $\hat{z} = \tilde{z}$, both equalling $\boldsymbol{\Phi}^{-1}(0.933) = 1.50$. For case II the prediction is $\hat{z} = 1.64 - 0.34 = 1.30$, compared with the actual value $\hat{z} = 1.37$.

Formula **1** allows us to compute the confidence level $\hat{\alpha}$ for the event $\{\boldsymbol{\mu} \in \mathcal{R}_1\}$ solely in terms of bootstrap calculations, no matter how complicated the boundary may be. A first level of bootstrap replications with $\hat{\boldsymbol{\mu}}^* \sim N_2(\hat{\boldsymbol{\mu}}, I)$ gives bootstrap data vectors $\hat{\boldsymbol{\mu}}^*(1), \hat{\boldsymbol{\mu}}^*(2), \ldots, \hat{\boldsymbol{\mu}}^*(B)$, from which we calculate

$$\tilde{z} = \boldsymbol{\Phi}^{-1}\left(\frac{\#\{\hat{\boldsymbol{\mu}}^* \text{ vectors in } \mathcal{R}_1\}}{B}\right).$$

A second level of bootstrap replications with $\hat{\boldsymbol{\mu}}^{**} \sim N_2(\hat{\boldsymbol{\mu}}_0, I)$, giving say $\hat{\boldsymbol{\mu}}^{**}(1), \hat{\boldsymbol{\mu}}^{**}(2), \ldots, \hat{\boldsymbol{\mu}}^{**}(B_2)$, allows us to calculate

$$\hat{z}_0 = \boldsymbol{\Phi}^{-1}\left(\frac{\#\{\hat{\boldsymbol{\mu}}^{**} \text{ vectors in } \mathcal{R}_1\}}{B_2}\right).$$

Then formula **1** gives $\hat{z} = \tilde{z} - 2z_0$.

As few as $B = 100$, or even 50, replications $\hat{\boldsymbol{\mu}}^*$ are enough to provide a rough but useful estimate of the confidence value $\tilde{\alpha}$. However, because the difference between $\tilde{z} = \boldsymbol{\Phi}^{-1}(\tilde{\alpha})$ and $\hat{z} = \boldsymbol{\Phi}^{-1}(\hat{\alpha})$ is relatively small, considerably larger bootstrap samples are necessary to make formula **1** worthwhile. The calculations in section 9 of Efron (9) suggest both B and B_2 must be on the order of at least 1000. This point did not arise in cases I and II where we were able to do the calculations by direct numerical integration, but it is important in the kind of complicated tree-construction problems we are actually considering.

We now return to the problem of trees, as seen in Fig. 2. The version of formula **1** that applies to the multinomial model of Fig. 3 is

$$\hat{z} = \frac{\tilde{z} - z_0}{1 + a(\tilde{z} - z_0)} - z_0. \qquad \textbf{[2]}$$

Here "a" is the *acceleration constant* introduced in ref. 9. It is quite a bit easier to calculate than z_0, as shown in the next section. Formula **2** is based on the bootstrap confidence intervals called "BC_a" in ref. 9.

If we tried to draw Fig. 3 accurately we would find that the multi-dimensional boundaries were hopelessly complicated. Nevertheless, formula **2** allows us to obtain a good approximation to the hypothesis-testing confidence level $\hat{\alpha} = \boldsymbol{\Phi}(\hat{z})$

solely in terms of bootstrap computations. How to do so is illustrated in the next section.

An Example Concerning the Malaria Data

Fig. 2 shows an estimated confidence value of

$$\tilde{\alpha} = 0.965$$

for the existence of the 9-10 clade on the malaria evolutionary tree. This value was based on $B = 200$ bootstrap replications, but (with some luck) it agrees very closely with the value $\tilde{\alpha} = 0.962$ obtained from $B = 2000$ replications. How does $\tilde{\alpha}$ compare with $\hat{\alpha}$, the hypothesis-testing confidence level for the 9-10 clade? We will show that

$$\hat{\alpha} = 0.942$$

(or $\hat{\alpha} = 0.938$ if we begin with $\tilde{\alpha} = 0.962$ instead of 0.965). To put it another way, our nonconfidence in the 9-10 clade goes from $1 - \tilde{\alpha} = 0.035$ to $1 - \hat{\alpha} = 0.058$, a substantial change.

We will describe, briefly, the computational steps necessary to compute $\hat{\alpha}$. To do so we need notation for multinomial sampling. Let $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ indicate a probability vector on $n = 221$ components, so the entries of the vector $\mathbf{P}$ are nonnegative numbers summing to 1. The notation

$$\mathbf{P}^* \sim \text{Mult}(\mathbf{P})$$

will indicate that $\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$ is the vector of proportions obtained in a multinomial sample of size n from $\mathbf{P}$. In other words we independently draw integers $I_1^*, I_2^*, \ldots, I_n^*$ from $\{1, 2, \ldots, n\}$ with probability P_k on k, and record the proportions $P_k^* = \#\{I_i^* = k\}/n$. This is the kind of multinomial sampling pictured in Fig. 3, expressed more efficiently in terms of $n = 221$ coordinates instead of $K = 4^{11} - 4$.

Each vector $\mathbf{P}^*$ is associated with a data matrix $\mathbf{x}^*$ that has proportion P_k^* of its columns equal to the kth column of the original data matrix $\mathbf{x}$. Then $\mathbf{P}^*$ determines a distance matrix and a tree according to the original tree-building algorithm,

$$\mathbf{P}^* \rightarrow \hat{D}^* \rightarrow \widehat{\text{TREE}}^*.$$

The "central" vector

$$\mathbf{P}^{(\text{cent})} = (1/n, 1/n, \ldots, 1/n)$$

corresponds to the original data matrix $\mathbf{x}$ and the original tree $\widehat{\text{TREE}}$. Notice that taking $\mathbf{P}^* \sim \text{Mult}(\mathbf{P}^{(\text{cent})})$ amounts to doing ordinary bootstrap sampling, since then $\mathbf{x}^*$ has its columns chosen independently and with equal probability from the columns of $\mathbf{x}$.

Resampling from $\mathbf{P}^{(\text{cent})}$ means that each of the 221 columns is equally likely, but this is not the same as all possible 11 vectors being equally likely. There were only 149 *distinct* 11 vectors among the columns of $\mathbf{x}$, and these are the only ones that can appear in $\mathbf{x}^*$. The vector *TTTTCTTTTTT* appeared seven times among the columns of $\mathbf{x}$, so it shows up seven times as frequently in the columns of $\mathbf{x}^*$, compared with *ATA-AAAAAAAA* which appeared only once in $\mathbf{x}$.

Here are the steps in the computation of $\hat{\alpha}$.

Step 1. $B = 2000$ first-level bootstrap vectors $\mathbf{P}^*(1), \mathbf{P}^*(2), \ldots, \mathbf{P}^*(B)$ were obtained as independent multinomials $\mathbf{P}^* \sim \text{Mult}(\mathbf{P}^{(\text{cent})})$. Some 1923 of the corresponding bootstrap trees had the 9-10 clade, giving the estimate $\tilde{\alpha} = 0.962 = 1923/2000$.

Step 2. The first 200 of these included seven cases without the 9-10 clade. Call the seven $\mathbf{P}^*$ vectors $\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(7)}$. For each of them, a value of w between 0 and 1 was found such that the vector

$$\mathbf{p}^{(j)} = w \cdot \mathbf{P}^{(j)} + (1 - w)\mathbf{P}^{(\text{cent})}$$

was right on the 9-10 boundary. The vectors $\mathbf{p}^{(j)}$ play the role of $\hat{\boldsymbol{\mu}}_0$ in Fig. 5.

Finding w is easy using a one-dimensional binary search program, as on page 90 of ref. 10. At each step of the search it is only necessary to check whether or not the current value of $w\mathbf{P}^{(j)} + (1 - w)\mathbf{P}^{(\text{cent})}$ gives a tree having the 9-10 clade. Twelve steps of the binary search, the number used here, locates the boundary value of w within $1/2^{12}$. The vectors $\mathbf{p}^{(j)}$ play the role of $\hat{\boldsymbol{\mu}}_0$ in Fig. 5.

Step 3. For each of the boundary vectors $\mathbf{p}^{(j)}$ we generated $B_2 = 400$ second-level bootstrap vectors

$$\mathbf{P}^{**} \sim \text{Mult}(\mathbf{p}^{(j)}),$$

computed the corresponding tree, and counted the number of trees having the 9-10 clade. The numbers were as follows for the seven cases:

Case	No.	B_2
1	218	400
2	204	400
3	223	400
4	214	400
5	213	400
6	216	400
7	223	400
Total	1151	2800

From the total we calculated an estimate of the correction term z_0 in formula **2**,

$$z_0 = \Phi^{-1}\left(\frac{1511}{2800}\right) = 0.0995.$$

Binomial calculations indicate that $z_0 = 0.0995$ has a standard error of about 0.02 due to the bootstrap sampling (that is, due to taking 2800 instead of all possible bootstrap replications), so 2800 is not lavishly excessive. Notice that we could have started with the 77 out of the 2000 $\mathbf{P}^*$ vectors not having the 9-10 clade, rather than the 7 out of the first 200, and taken $B_2 = 40$ for each $\mathbf{p}^{(j)}$, giving about the same total second-level sample.

Step 4. The acceleration constant "a" appearing in formula **2** depends on the direction from $\mathbf{P}^{(\text{cent})}$ to the boundary, as explained in section 8 of ref. 9. For a given direction vector $\mathbf{U}$,

$$a(\mathbf{U}) = \frac{1}{6}\sum_1^n U_k^3/(\sum_1^n U_k^2)^{3/2}.$$

Taking $\mathbf{U} = \mathbf{p}^{(j)} - \mathbf{P}^{(\text{cent})}$ for each of the seven cases gave

Case	a
1	0.014
2	0.009
3	0.014
4	0.012
5	0.014
6	0.012
7	0.014
Average	0.0129

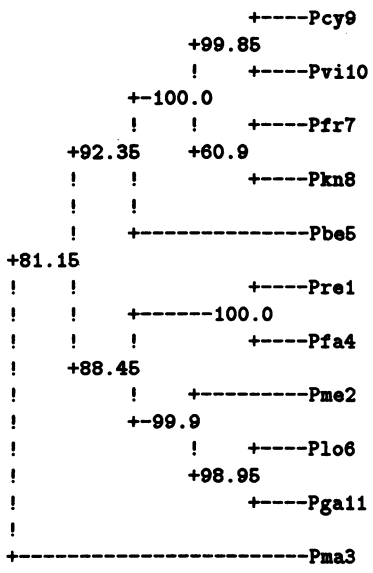
Step 5. Finally we applied formula **2** with $\tilde{z} = \Phi^{-1}(0.962) = 1.77$, $z_0 = 0.0995$, and $a = 0.0129$, to get $\hat{z} = 1.54$, or $\hat{\alpha} = \Phi(\hat{z}) = 0.938$. If we begin with $\tilde{z} = \Phi^{-1}(0.965)$ then $\hat{\alpha} = 0.942$.

Notice that in this example we could say that Felsenstein's bootstrap confidence value $\tilde{\alpha}$ was biased *upward*, not downward, at least compared with the hypothesis-testing level $\hat{\alpha}$. This happened because z_0 was positive, indicating that the 9-10

boundary was curving away from $\mathbf{P}^{(cent)}$, just as in case 2 of Fig. 5. The opposite can also occur, and in fact did for other clades. For example the clade at the top of Fig. 2 that includes all of the species except lizard (species 2) had $\tilde{\alpha} = 0.775$ compared with $\hat{\alpha} = 0.875$.

We carried out these same calculations using the more efficient tree-building algorithm employed in Escalante and Ayala (11); that is we used Felsenstein's PHYLIP package (12) on the complete RNA sequences, neighbor-joining trees based on Kimura's (13) two-parameter distances.

In order to vary our problem slightly, we looked at the clade 7-8 (Pfr-Pkn), which is more questionable than the 9-10 clade. The tree produced from the original set is:



Step 1. $B = 2000$ first-level bootstrap vectors. Some 1218 of the corresponding bootstrap trees had the 7-8 clade, giving the estimate $\tilde{\alpha} = 0.609 = 1218/2000$.

Step 2. We took, as before, seven cases without the 7-8 clade, and for each one found a multinomial vector near the 7-8 boundary.

Step 3. For each of the boundary vectors $\mathbf{p}^{(j)}$ we generated $B_2 = 400$ second-level bootstrap vectors

$$\mathbf{P}^{**} \sim \text{Mult}(\mathbf{p}^{(j)}),$$

computed the corresponding tree, and counted the number of trees having the 7-8 clade. The numbers were as follows for the seven cases:

Case	No.	B_2
1	120	400
2	184	400
3	145	400
4	187	400
5	176	400
6	197	400
7	240	400
Total	1249	2800

From the total we calculated an estimate of the correction term z_0 in formula **2**,

$$z_0 = \Phi^{-1}\left(\frac{1249}{2800}\right) = -0.136$$

Step 4. The acceleration constant "a" appearing in formula **2** was computed as before giving:

Case	a
1	−0.118
2	−0.0176
3	0.0172
4	−0.0256
5	0.00981
6	−0.0540
7	−0.0198
Average	−0.0296

Step 5. Finally we applied formula 2 with $\tilde{z} = \Phi^{-1}(0.609) = 0.277$ to get $\hat{z} = 0.417$, or $\hat{\alpha} = \Phi(\hat{z}) = 0.662$. In this case $\hat{\alpha}$ is bigger than $\tilde{\alpha}$, reflecting the fact that the 7-8 boundary curves toward the central point, at least in a global sense.

Computing $\hat{\alpha}$ is about 20 times as much work as $\tilde{\alpha}$, but it is work for the computer and not for the investigator. Once the tree-building algorithm is available, all of the computations require no more than applying this algorithm to resampled versions of the original data set.

Discussion and Summary

The discussion in this paper, which has gone lightly over many technical details of statistical inference, makes the following main points about the bootstrapping of phylogenetic trees.

(*i*) The confidence values $\tilde{\alpha}$ obtained by Felsenstein's bootstrap method are not biased systematically downward.

(*ii*) In a Bayesian sense, the $\tilde{\alpha}$ can be thought of as reasonable assessments of error for the estimated tree.

(*iii*) More familiar non-Bayesian confidence levels $\hat{\alpha}$ can also be defined. Typically $\hat{\alpha}$ and $\tilde{\alpha}$ will converge as the number n of independent sites grows large, at rate $1/\sqrt{n}$.

(*iv*) The $\hat{\alpha}$ can be estimated by a two-level bootstrap algorithm.

(*v*) As few as 100 or even 50 bootstrap replications can give useful estimates of $\tilde{\alpha}$, while $\hat{\alpha}$ estimates require at least 2000 total replications. None of the computations requires more than applying the original tree-building algorithm to resampled data sets.

We are grateful to A. Escalante and F. Ayala (14) for providing us with these data. B.E. is grateful for support from Public Health Service Grant 5 R01 CA59039-20 and National Science Foundation Grant DMS95-04379. E.H. is supported by National Science Foundation Grant DMS94-10138 and National Institutes of Health Grant NIAID R29-A131057.

1. Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap* (Chapman & Hall, London).
2. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
3. Hillis, D. & Bull, J. (1993) *Syst. Biol.* **42,** 182–192.
4. Felsenstein, J. & Kishino, H. (1993) *Syst. Biol.* **42,** 193–200.
5. Newton, M. (1995) *Bootstrap Phylogenies: Large Deviations and Dispersion Effect* (Univ. of Wisconsin, Madison), Tech. Rep. 923.
6. Zharkikh, A. & Li, W. H. (1992) *Mol. Biol. Evol.* **9,** 1119–1147.
7. Efron, B. (1982) *SIAM CBMS-NSF Monogr.* **38.**
8. Tibshirani, R. J. (1989) *Biometrika* **76,** 604–608.
9. Efron, B. (1987) *J. Am. Stat. Assoc.* **82,** 171–185.
10. Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (1987) *Numerical Recipes* (Cambridge Univ. Press, New York).
11. Escalante, A. & Ayala, F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5793–5797.
12. Felsenstein, J. (1993) PHYLIP (Univ. of Washington, Seattle).
13. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
14. Escalante, A. & Ayala, F. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11371–11377.