Interior-Branch and Bootstrap Tests of Phylogenetic Trees

Tatyana Sitnikova, Andrey Rzhetsky, and Masatoshi Nei

Institute of Molecular and Evolutionary Genetics and Department of Biology, Pennsylvania State University

We have compared statistical properties of the interior-branch and bootstrap tests of phylogenetic trees when the neighbor-joining tree-building method is used. For each interior branch of a predetermined topology, the interior-branch and bootstrap tests provide the confidence values, P_C and P_B, respectively, that indicate the extent of statistical support of the sequence cluster generated by the branch. In phylogenetic analysis these two values are often interpreted in the same way, and if P_C and P_B are high (say, ≥ 0.95), the sequence cluster is regarded as reliable. We have shown that P_C is in fact the complement of the P-value used in the standard statistical test, but P_B is not. Actually, the bootstrap test usually underestimates the extent of statistical support of species clusters. The relationship between the confidence values obtained by the two tests varies with both the topology and expected branch lengths of the true (model) tree. The most conspicuous difference between P_C and P_B is observed when the true tree is starlike, and there is a tendency for the difference to increase as the number of sequences in the tree increases. The reason for this is that the bootstrap test tends to become progressively more conservative as the number of sequences in the tree increases. Unlike the bootstrap, the interior-branch test has the same statistical properties irrespective of the number of sequences used when a predetermined tree is considered. Therefore, the interior-branch test appears to be preferable to the bootstrap test as long as unbiased estimators of evolutionary distances are used. However, when the interior-branch is applied to a tree estimated from a given data set, P_C may give an overestimate of statistical confidence. For this case, we developed a method for computing a modified version (P'_C) of the P_C value and showed that this P'_C tends to give a conservative estimate of statistical confidence, though it is not as conservative as P_B. In this paper we have introduced a model in which evolutionary distances between sequences follow a multivariate normal distribution. This model allowed us to study the relationships between the two tests analytically.

Introduction

There are several different methods that are currently in use for testing the statistical significance of a particular branching pattern of a phylogenetic tree (see, e.g., Felsenstein 1988; Li and Gouy 1991; Nei 1991 for review). Since the significance levels obtained by different methods for the same phylogenetic tree do not necessarily agree with each other, it is important to understand statistical properties of these methods. In this paper we compare the interior-branch test (Nei et al. 1985; Li 1989; Rzhetsky and Nei 1992*a*) and the bootstrap test (Efron 1982; Felsenstein 1985) using both computer simulation and analytical study. A similar study was carried out by Pamilo (1990), but his study was concerned with the UPGMA tree-building method in which a constant rate of evolution is assumed. We will consider a tree-building method which does not require this assumption.

In the present study we used two different types of computer simulations. In the first type the evolutionary distances corrected for multiple hits were estimated from observed nucleotide differences. In the second type of simulation the evolutionary distances were drawn from a multivariate normal distribution with the same mean vector and the variance-covariance matrix as those for distances calculated from nucleotide sequences. By comparing the results of the two types of simulation for the same model tree, we were able to study the effect of the different distributions of distances on the relationships of the interior-branch and bootstrap tests of statistical confidence. In addition, the second type of simulation led us to study the relationships of the two tests analytically.

We start our analysis with a simple model tree of four sequences. We examine the relationship between

Key words: interior-branch test, bootstrap, phylogenetic inference, neighbor joining.

Address for correspondence and reprints: Tatyana Sitnikova, Institute of Molecular Evolutionary Genetics, 328 Muller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 12(2):319–333. 1995.

0737-4038/95/1202-0013$02.00

the statistical confidences obtained by the two methods for various model trees. We will then compare the two tests for the case of six sequences. Before presenting our results, we briefly explain the essential aspects of the two statistical tests.

Interior-Branch Test

Consider a given tree for a set of m nucleotide sequences. The interior-branch test for this tree is conducted as follows. We first compute the unbiased estimates of the evolutionary distances (the numbers of nucleotide substitutions) for all pairs of sequences, that is, $\hat{d}_{ij}$'s, where i and j refer to the ith and jth sequences, respectively. The estimates ($\hat{b}_i$'s) of branch lengths for the tree are then given by

$$\hat{\mathbf{b}} = \mathbf{L}\hat{\mathbf{d}}, \tag{1}$$

where $\hat{\mathbf{b}}^t = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{2m-3})$ and $\hat{\mathbf{d}}^t = (\hat{d}_{12}, \hat{d}_{13}, \ldots, \hat{d}_{m-1,m})$ are the transposes of the column vectors of branch length and distance estimates, respectively. $\mathbf{L}$ is a $([2m-3] \times m[m-1]/2)$ matrix that is specified by the tree (Rzhetsky and Nei 1992*a*). Let $\hat{\mathbf{W}}$ be the estimate of the variance-covariance matrix ($\mathbf{W}$) of vector $\hat{\mathbf{d}}$ (see below). The vector of the variances of branch length estimates, $\mathbf{V}(\hat{\mathbf{b}}^t) = (V[\hat{b}_1], V[\hat{b}_2], \ldots, V[\hat{b}_{2m-3}])$, can then be obtained by

$$\mathbf{V}(\hat{\mathbf{b}}) = \mathbf{L}^t\hat{\mathbf{W}}\mathbf{L}. \tag{2}$$

In practice, the computation of $\hat{b}_i$'s and $V(\hat{b}_i)$'s for a large tree becomes simpler if we use Rzhetsky and Nei's (1993) formulas that require no matrix algebra.

The null hypothesis of the interior-branch test is that the interior branch under consideration has length 0. To test this hypothesis, we can use the following test statistic (normal deviate) if $\hat{b}_i$ can be assumed to be normally distributed:

$$Z = \hat{b}_i / s(\hat{b}_i), \tag{3}$$

where $s(\hat{b}_i) = V(\hat{b}_i)^{1/2}$. We have done computer simulation to examine the distribution of Z in equation (3) under the null hypothesis and have shown that the distribution is indeed approximately normal. Therefore, Z can be used to construct the two-sided normal deviate test. The null hypothesis is rejected at the significance level of α if $|Z| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $(\alpha/2)$-critical value for the standard normal distribution.

To compare the interior-branch test with the bootstrap test, it is convenient to consider the following probability:

$$P_C = 2\Phi(|Z|) - 1, \tag{4}$$

where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-x^2/2} dx. \tag{5}$$

P_C tends to be 0 when $\hat{b}_i$ approaches 0 and tends to b 1 when $|\hat{b}_i|$ increases. We call this confidence value. is the complement of the P-value used in the standar statistical test.

Note that the above statistical test applies when topology to be tested is predetermined. This situatio occurs when an investigator is interested in the reliabilit of a particular topology. However, the interior-branc test is often applied to a phylogenetic tree estimated fror actual data rather than to a predetermined tree. In th case the statistical properties of the P_C test become di ferent. Later we will consider a correction for P_C nec essary for this case.

Interior-Branch Test for Large Trees

In the case of four-sequence trees the expected valu of the estimate of the interior branch length can be eithe positive (for the true tree) or negative (for a wrong tree if the true tree is not starlike. Hence, the estimate of th interior branch length gives some idea about the validit of a tree. In the case of a large number of sequences w can test the null hypothesis $E(\hat{b}_i) = 0$ for the ith interio branch, but the interpretation of the test outcome somewhat more complicated. It can be shown that th expectation of the length estimate of an incorrect interio branch (which gives a sequence partition that is no present in the true tree) in a large tree can be positiv ($E[\hat{b}_i] > 0$). However, a negative $E(\hat{b}_i)$ always indicate that the corresponding partition is wrong. We illu strate this point with the following example of a six sequence tree.

Let tree A in figure 1 be the true tree for the si sequences 1, 2, 3, 4, 5, and 6, and let b_7, b_8, and b_9 b the expected lengths of interior branches of this tree. Le us compute the expectation of the least-square estimate of the interior branch lengths of wrong trees B and C i figure 1. For interior branches 7, 8, and 9 of tree B (fi 1B)), we obtain

$$E(\hat{b}_7) = b_7 + \frac{2}{3} b_8 + \frac{1}{3} b_9,$$

$$E(\hat{b}_8) = -\frac{5}{18}(2b_8 + b_9), \tag{6}$$

and $$E(\hat{b}_9) = \frac{1}{6}(4b_8 - b_9).$$

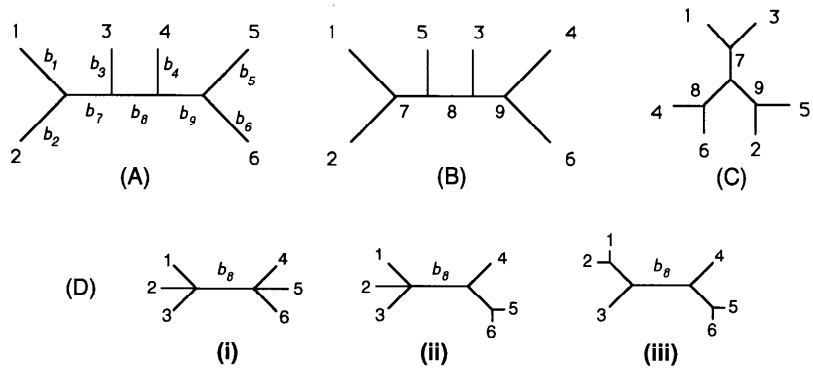

FIG. 1.—*A*, Hypothetical true tree for six sequences. *B* and *C*, Incorrect trees for the same sequences. *D*, Three model trees for six sequences used for generating the relationship between $\bar{P}_C$ and $\bar{P}_B$ in fig. 6*B*. When $b_8 = 0$, trees (i), (ii), and (iii) behave as six-, five-, and four-sequence star trees, respectively.

Similarly, for tree C, we have

$$E(\hat{b}_7) = \frac{1}{4}(-b_7+2b_8+b_9),$$

$$E(\hat{b}_8) = \frac{1}{4}(b_7+2b_8-b_9), \qquad (7)$$

$$\text{and} \quad E(\hat{b}_9) = -\frac{1}{4}(b_7+2b_8+b_9).$$

Therefore, (1) the value of $E(\hat{b}_7)$ for tree B is always positive and the corresponding interior branch gives the correct partition of sequences, (2) the values of $E(\hat{b}_8)$ for tree B and $E(\hat{b}_9)$ for tree C are always negative and both interior branches give an incorrect partition of sequences, and (3) the values of $E(\hat{b}_9)$ for tree B and $E(\hat{b}_7)$ and $E(\hat{b}_8)$ for tree C can be either positive or negative depending on the actual values of b_7, b_8, and b_9. We note that all these interior branches in trees B and C give an incorrect partition of sequences.

From a number of other similar examples, we have conjectured that any wrong bifurcating tree has at least one interior branch with a negative value of expected length estimate. (We do not attempt to prove this conjecture in this paper but assume that it is true.) Therefore, the interior-branch test indicates that the tree under consideration can not be rejected if there is no interior branch length estimate that is significantly smaller than 0. Note that due to sampling errors one may obtain some negative branch length estimates even for the true topology, but the probability that these estimates are significantly different from 0 should be very small.

Since the interpretation of P_C varies with the sign of the corresponding branch length estimate ($\hat{b}_i$), we shall compare P_C and its equivalent quantity for the bootstrap test for the positive and negative values of $\hat{b}_i$ separately.

Bootstrap Test

Unlike the interior-branch test, the bootstrap test is not independent of the tree-building method. In this paper we consider the neighbor-joining (NJ) method (Saitou and Nei 1987) using unbiased estimates of evolutionary distances. However, the results obtained here should apply to other methods as well if the methods are as efficient as the NJ method in obtaining the correct tree. A common way to apply the bootstrap test is first to construct an NJ tree from a given set of sequence data and then test this tree with a bootstrap test. However, as in the case of the interior-branch test, the bootstrap test can be applied to any predetermined tree (Zharkikh and Li 1992*a*, 1992*b*, 1995).

Once a tree is obtained, the original data set is used to generate B independent pseudosamples of sequences. Each pseudosample is obtained by drawing nucleotide sites randomly from the original set of sequences with replacement until a sample of the same size (number of nucleotides) as the original one is obtained. Some sites are sampled several times, and others are omitted. We then apply the NJ method to construct a phylogenetic tree from each pseudosample and compare each bootstrap tree with the original tree. This is repeated for all B pseudosamples, and for each cluster (or partition of sequences; see Penny and Hendy 1985; Rzhetsky and Nei 1992*a*) of the original tree, the proportion of bootstrap trees in which this cluster appears is computed. We call this valuc (P_B) the bootstrap confidence value or bootstrap value.

Note that the above procedure of the bootstrap test is somewhat different from that of Felsenstein (1985). In his method the topology of a bootstrap tree is not compared with that of the original tree. Instead, a bootstrap consensus tree is produced, and this tree is regarded as a representative tree obtained from the data set. The P_B for sequence cluster of this tree is the proportion of

bootstrap trees in which the cluster appears. The bootstrap tests for the maximum-parsimony and NJ methods in the PHYLIP program package (Felsenstein 1991) are both conducted in this way. By contrast, the bootstrap test for the NJ method in the MEGA package (Kumar et al. 1993) follows the method described above. Kumar et al.'s method is for testing the reliability of the tree obtained from the original set of data.

Literally speaking, P_B is an estimate of the proportion of cases in which the tree-building method recovers a specific sequence cluster from infinitely many data sets generated by the same evolutionary process. Recent studies (Zharkikh and Li 1992*a*, 1992*b*, 1995; Hillis and Bull 1993; Felsenstein and Kishino 1993) have shown that P_B is a biased estimate of this proportion. Bootstrap tests tend to give conservative estimates of statistical confidence when the true tree contains the cluster under investigation, and the length of the corresponding interior branch is positive and small. By contrast, bootstrap tests tend to be liberal if the cluster does not belong to the true tree. (The latter problem rarely affects the phylogenetic inference, since it occurs only for small values of P_B.) All the above statements about statistical properties of the bootstrap test are correct only when the tree-building method used in bootstrapping is statistically consistent.

The null hypothesis of bootstrap tests has not been clearly defined. Since P_B for a particular sequence cluster or an interior branch is required to be greater than a threshold value (say, 0.95) for accepting this cluster as statistically significant, P_B has been implicitly interpreted as though it is the same quantity as P_C. In practice, this is not true, as will be shown below.

Comparison of P_C and P_B by Computer Simulation and Analytical Study

In the following section we present details of the procedure of our computer simulation only for the case of four-sequence trees, since its extension for the case of larger trees is straightforward.

Nucleotide Sequence Data

We used the Jukes and Cantor (1969) substitution model to simulate the evolution of nucleotide sequences and compare P_C and P_B, though any model can be used as long as the distance estimates obtained are unbiased. The procedure of the simulation was as follows. First, specify the values of expected branch lengths b_1, b_2, b_3, b_4, and b_5 for the model tree in figure 2*A*. Second, generate four "extant" nucleotide sequences of length n according to the model tree as in Rzhetsky and Nei (1992*b*). Third, estimate the vector of evolutionary distances between sequences, $\hat{\mathbf{d}}^t = (\hat{d}_{12}, \hat{d}_{13}, \hat{d}_{14}, \hat{d}_{23}, \hat{d}_{24}, \hat{d}_{34})$, using the Jukes-Cantor method. The variance-covariance matrix ($\mathbf{W}$) of the vector $\hat{\mathbf{d}}$ is then estimated

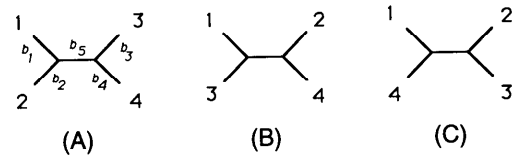

FIG. 2.—*A*, Hypothetical true tree for four sequences. *B* and *C*, Incorrect trees for the same sequences.

by using Kimura and Ohta's (1972) formula for the variances and Bulmer's (1991) formula for the covariances. Fourth, estimate the length of the interior branch of the true tree (tree A in fig. 2) by the equation

$$\hat{b}_5 = \mathbf{k}_0\hat{\mathbf{d}}, \quad \text{where } \mathbf{k}_0 = (-1/2,\ 1/4,\ 1/4,\ 1/4,\ 1/4,\ -1/2), \tag{8}$$

and the variance of $\hat{b}_5$ by

$$V(\hat{b}_5) = \mathbf{k}_0\hat{\mathbf{W}}\mathbf{k}_0^t, \tag{9}$$

where $\hat{\mathbf{W}}$ is the estimate of $\mathbf{W}$ (Rzhetsky and Nei 1992*b*). Finally, compute P_C using equation (4) to test the null hypothesis $E(\hat{b}_5) = 0$. Compute P_B (i.e., the proportion of bootstrap trees in which the interior branch under consideration appears). In all simulations discussed below we used $n = 100$ and $B = 500$ unless otherwise stated.

Normally Distributed Distance Data

The joint distribution of $\hat{d}_{ij}$'s computed above is slightly asymmetrical compared with a multivariate normal distribution. To evaluate the effect of this skewness on the performance of the interior-branch test, we conducted another type of simulation, in which the vector of the distance estimates, $\hat{\mathbf{d}}$, exactly followed a multivariate normal distribution. In this simulation all assumptions underlying the interior-branch test are satisfied, and the results obtained could be compared with those from the analytical formulas.

The procedure of this simulation was exactly the same as those for nucleotide sequences except for some details concerning the estimation of evolutionary distances (steps 2 and 3 in the previous section). We used the following method to generate a vector of estimates of "evolutionary distances" ($\hat{\mathbf{d}}$) following a multivariate normal distribution. The expected distances for this multivariate normal distribution were given by

$$d_{12} = b_1 + b_2,\ d_{13} = b_1 + b_5 + b_3,$$
$$d_{14} = b_1 + b_5 + b_4,\ d_{23} = b_2 + b_5 + b_3, \tag{10}$$
$$d_{24} = b_2 + b_5 + b_4, \text{ and } d_{34} = b_3 + b_4,$$

where b_i's are the expected branch lengths of the model tree (see fig. 2*A*), and the variance-covariance matrix, $\mathbf{W}$, was

$$\mathbf{W} = \begin{bmatrix} f(b_1+b_2) & f(b_1) & f(b_1) & f(b_2) & f(b_2) & 0 \\ f(b_1) & f(b_1+b_3+b_5) & f(b_1+b_5) & f(b_3+b_5) & f(b_5) & f(b_3) \\ f(b_1) & f(b_1+b_5) & f(b_1+b_4+b_5) & f(b_5) & f(b_4+b_5) & f(b_4) \\ f(b_2) & f(b_3+b_5) & f(b_5) & f(b_2+b_3+b_5) & f(b_2+b_5) & f(b_3) \\ f(b_2) & f(b_5) & f(b_4+b_5) & f(b_2+b_5) & f(b_2+b_4+b_5) & f(b_4) \\ 0 & f(b_3) & f(b_4) & f(b_3) & f(b_4) & f(b_3+b_4) \end{bmatrix}. \quad (11)$$

Here

$$f(\delta) = (-9+6e^{4\delta/3}+3e^{8\delta/3})/(16n), \quad (12)$$

where δ is given by an appropriate sum of b_i's, and n is the sample size that is analogous to the number of nucleotide sites. The matrix $\mathbf{W}$ is the same as the variance-covariance matrix of vector $\hat{\mathbf{d}}$ for nucleotide sequence data generated by the Jukes-Cantor model following the model tree in figure 2*A* (see Rzhetsky and Nei 1992*b*). To generate observed distances, $\hat{d}_{ij}$'s, we computed a triangular matrix $\mathbf{C}$ such that

$$\mathbf{W} = \mathbf{C}'\mathbf{C} \quad (13)$$

(for details, see Press et al. 1988, pp. 39–45). We then computed n random vectors, $\mathbf{x_k} = (x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5}, x_{k6})$, where $k = 1, 2, \ldots, n$, and x_{ki}'s are independent random variables drawn from a standard normal distribution. We generated x_{ki}'s using the Box-Muller method (see Press et al. 1988, pp. 216–217). We used vectors $\mathbf{x_k}$'s to obtain a set of random vectors $\hat{\mathbf{d}}_1$, $\hat{\mathbf{d}}_2$, . . . , and $\hat{\mathbf{d}}_n$ by

$$\hat{\mathbf{d}}_\mathbf{k} = \mathbf{d} + \mathbf{C}\mathbf{x_k}\sqrt{n}, \quad (14)$$

where $\mathbf{d}$ is the vector of d_{ij}'s. We then computed a vector of estimates of evolutionary distances, $\hat{\mathbf{d}}$, by

$$\hat{\mathbf{d}} = (\hat{\mathbf{d}}_1+\hat{\mathbf{d}}_2+\ldots+\hat{\mathbf{d}}_n)/n \quad (15)$$

and the estimated variance-covariance matrix ($\hat{\mathbf{W}}$) of vector $\hat{\mathbf{d}}$ using the standard formulas for sample variances and covariances for a vector of sample means by

$$\hat{w}_{rs} = \left[\sum_{i=1}^{n} (\hat{d}_{r,i}-\hat{d}_r)(\hat{d}_{s,i}-\hat{d}_s)\right] \Big/ n(n-1), \quad (16)$$

where $\hat{w}_{rs}$ is an estimate of the covariance between the rth and sth entries of the $\hat{\mathbf{d}}$ vector, $\hat{d}_{r,i}$ is the rth entry of the vector $\hat{\mathbf{d}}_i$, and $\hat{d}_r$ is the rth entry of the vector $\hat{\mathbf{d}}$. The rest of the simulation was done in the same way as that for nucleotide sequence data (see steps 4 and 5 in the previous section). The vectors $\hat{\mathbf{d}}_1$, $\hat{\mathbf{d}}_2$, . . . , and $\hat{\mathbf{d}}_n$ were treated as nucleotide sites in bootstrap resampling.

When the estimates of evolutionary distances follow a multivariate normal distribution, one can obtain rather simple formulas for computing the expected value of P_B and the probability to recover the true tree by the NJ method. We shall consider these formulas in the next section.

Analytical Formulation for the Case of Normally Distributed Evolutionary Distances

In the case of four sequence trees, the NJ method is known to select the unrooted tree with the smallest sum of branch length estimates (Saitou and Nei 1987). Therefore, if we denote the sums of branch length estimates for trees *A*, *B*, and *C* in figure 2 by $\hat{S}_A$, $\hat{S}_B$, and $\hat{S}_C$, respectively, we can introduce the following variables, $\hat{D}_1$ and $\hat{D}_2$:

$$\hat{D}_1 = \hat{S}_B - \hat{S}_A, \text{ and } \hat{D}_2 = \hat{S}_C - \hat{S}_A. \quad (17)$$

Therefore, the NJ method recovers the correct tree (tree A) whenever both $\hat{D}_1$ and $\hat{D}_2$ are positive. We can compute $\hat{D}_1$ and $\hat{D}_2$ by

$$\hat{D}_1 = \mathbf{k}_1\hat{\mathbf{d}}, \text{ and } \hat{D}_2 = \mathbf{k}_2\hat{\mathbf{d}}, \quad (18)$$

where

$$\mathbf{k}_1 = (-1/4, 1/4, 0, 0, 1/4, -1/4), \text{ and}$$
$$\mathbf{k}_2 = (-1/4, 0, 1/4, 1/4, 0, -1/4). \quad (19)$$

When $\hat{d}_{ij}$'s follow a multivariate normal distribution, the joint distribution of $\hat{D}_1$ and $\hat{D}_2$ is a bivariate normal distribution with the expected values $\mu_1 = \mathbf{k}_1\mathbf{d}$ and $\mu_2 = \mathbf{k}_2\mathbf{d}$, and the variance-covariance matrix

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad (20)$$

where σ_1^2 and σ_2^2 are the variances of $\hat{D}_1$ and $\hat{D}_2$, re-

spectively, $\rho\sigma_1\sigma_2$ is the covariance for $\hat{D}_1$ and $\hat{D}_2$, and ρ is the correlation coefficient between $\hat{D}_1$ and $\hat{D}_2$. The elements of matrix **V** are given by

$$\sigma_1^2 = \mathbf{k}_1\mathbf{W}\mathbf{k}_1^t, \quad \sigma_2^2 = \mathbf{k}_2\mathbf{W}\mathbf{k}_2^t, \quad \text{and} \quad \rho\sigma_1\sigma_2 = \mathbf{k}_1\mathbf{W}\mathbf{k}_2^t, \tag{21}$$

where matrix **W** is computed by equations (11) and (12). Since the correct tree is recovered when $\hat{D}_1 > 0$ and $\hat{D}_2 > 0$, the probability (P) of obtaining the true tree (tree A in fig. 2) by the NJ method can be computed by

$$\begin{aligned} P &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_0^\infty\int_0^\infty \exp\left\{-\frac{1}{2(1-\rho^2)}\right. \\ &\times\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right. \\ &\left.\left.+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}dxdy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-(\mu_1/\sigma_1)}^\infty\int_{-(\mu_2/\sigma_2)}^\infty \exp\left\{-\frac{1}{2(1-\rho^2)}\right. \\ &\left.\times(x^2-2\rho xy+y^2)\right\}dxdy \\ &= F\left(-\frac{\mu_1}{\sigma_1}, -\frac{\mu_2}{\sigma_2}; \rho\right). \end{aligned} \tag{22}$$

This integral can be evaluated numerically (see Drezner and Wesolowsky 1990). Computer simulations have shown that equation (22) gives a surprisingly good estimate of the probability of obtaining the true tree by the NJ method as long as $n \geq 100$ even if the distribution of evolutionary distances deviates from normality to some extent, as in the case of nucleotide sequences.

We are now in a position to derive the expression for P_B. For each bootstrap pseudosample we can compute $\hat{d}^*$, $\hat{D}_1^*$ and $\hat{D}_2^*$, where the asterisk indicates that the quantity is computed from a pseudosample. As in the case of the original data set, the correct tree is recovered from a pseudosample only when $\hat{D}_1^*$ and $\hat{D}_2^*$ are both positive. The joint distribution of $\hat{D}_1^*$ and $\hat{D}_2^*$ for the pseudosamples obtained from a particular data set can be approximated by a bivariate normal distribution with mean vector $(\hat{\mathbf{D}}_1, \hat{\mathbf{D}}_2)$ and variance-covariance matrix $\hat{\mathbf{V}}$. Here, $\hat{D}_1$ and $\hat{D}_2$ are computed by equations (18) and (19) from the original set of data, and $\hat{\mathbf{V}}$ is the estimate of matrix **V** in equation (20). Therefore, P_B is equal to the proportion of cases where both $\hat{D}_1^*$ and $\hat{D}_2^*$ are positive and can be computed by

$$P_B \simeq F(-\hat{D}_1/\hat{\sigma}_1, -\hat{D}_2/\hat{\sigma}_2; \hat{\rho}). \tag{23}$$

Our numerical study has shown that this approximation works very well even when vector $\hat{\mathbf{d}}$ is computed from nucleotide sequences.

Case of Four-Sequence Trees

In this section we first examine the difference between the *average* P_C and P_B values for various model trees and then consider statistical properties of the *individual* values of P_C and P_B for a specified model tree.

Relationship between Average P_C and P_B for Various Model Trees

Let us consider the relationship of the averages ($\bar{P}_C$ and $\bar{P}_B$) of 3,000 P_C's and P_B's for the cases of $\hat{b}_5 > 0$ for the model tree of four sequences in figure 2*A* when b_5 gradually increases from 0 (fig. 3). $\bar{P}_C$ and $\bar{P}_B$ are quite close to each other, though the individual values of P_C and P_B obtained for each data set are not necessarily close (data not shown). Figure 3*A* and *B* represent the cases of normally distributed distance data and simulated nucleotide sequence data, respectively. Although these two figures are slightly different for small values of $\bar{P}_C$ and $\bar{P}_B$ (small values of b_5), the relationships between the two estimates are very similar for the two types of data.

The relationship between $\bar{P}_C$ and $\bar{P}_B$ depends on the expected lengths of the exterior branches of the model tree. The points ($\bar{P}_C$, $\bar{P}_B$) tend to be "upward" ($\bar{P}_C < \bar{P}_B$) when the model tree has short exterior branches

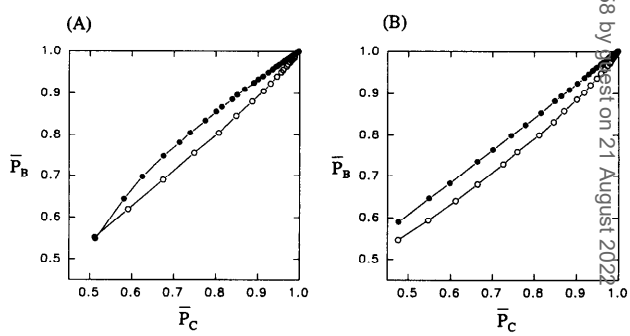

FIG. 3.—Comparison of the average values of P_C and P_B for four-sequence model trees for the cases of $\hat{b}_5 > 0$. We used the following two sets of the expected exterior branch lengths for the model tree in fig. 2*A*: $b_1 = 0.07$, $b_2 = 0.075$, $b_3 = 0.065$, $b_4 = 0.07$ (short branch trees, solid circles), and $b_1 = 0.3$, $b_2 = 0.3$, $b_3 = 0.3$, $b_4 = 0.3$ (long branch trees, open circles). The expected length of the interior branch, b_5, was changed from 0 to 0.7 by incrementing it by 0.005 for each model tree of short branches and by incrementing it by 0.03 for each model tree of long branches. The sample size (number of nucleotides) was equal to 100. Each point was obtained by averaging results of 3,000 replications. *A*, Normally distributed distance data; *B*, simulated nucleotide sequence data.

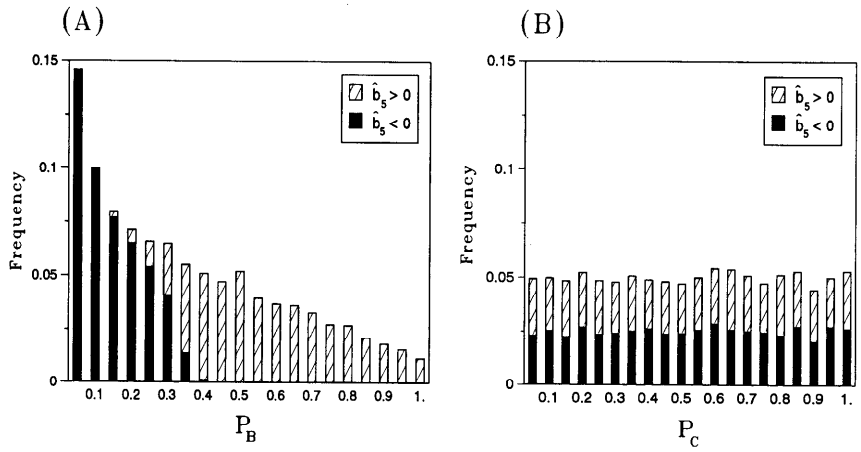

FIG. 4.—Frequency distributions of P_B (*A*) and P_C (*B*) for normally distributed distance data obtained in 10,000 replicate simulations. The frequencies of P_B and P_C for $\hat{b}_5 > 0$ and $\hat{b}_5 < 0$ are shown separately (corresponding bars are stacked one on top of the other). Here we used the model tree in fig. 2*A* with $b_1 = b_2 = b_3 = b_4 = 0.3$, and $b_5 = 0$. The number of nucleotides sampled (n) was equal to 100.

(fig. 3, solid circles) and "downward" when the exterior branches are long (fig. 3, open circles). We conducted an analytical study of these properties for the case of normally distributed distance data. In this case it is easy to derive an equation for $\bar{P}_C$ (Appendix B), where $\bar{P}_C$ is determined by $\zeta \equiv b_5/\sigma$ (where $\sigma = s[\hat{b}_5]$). To obtain an equivalent equation for $\bar{P}_B$, we use equation (23), the relationship $\hat{b}_5 = \hat{D}_1 + \hat{D}_2$ (see equations [8], [18], and [19]), and $V(\hat{b}_5) = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2$. To simplify the analysis, we replace $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\rho}$ in equation (23) by their expected values σ_1, σ_2, and ρ. Then, after some algebraic manipulations we obtain the following equation:

$$\bar{P}_B \simeq \frac{1}{\Phi(\zeta)2\pi(1-\rho^2)^{1/2}} \times \int_{-\infty}^{\infty}\int_{-\sigma_1 x_1/\sigma_2}^{\infty} \exp\{-(x_1^2 - 2\rho x_1 x_2 + x_2^2)/[2(1-\rho^2)]\} \times F[-(x_1+\zeta\alpha_1), -(x_2+\zeta\alpha_2); \rho]\, dx_1 dx_2, \quad (24)$$

where $\zeta = b_5/\sigma$, $\alpha_1 = (1+2\rho\sigma_2/\sigma_1+\sigma_2^2/\sigma_1^2)^{1/2}/2$, $\alpha_2 = (1+2\rho\sigma_1/\sigma_2+\sigma_1^2/\sigma_2^2)^{1/2}/2$, and $F(y_1, y_2; \rho)$ is a function given in (22). Both computer simulations and analytical studies have indicated that for $b_5 > 0$, the values of σ_1 and σ_2 are very close to each other for the majority of the model trees (see Appendix A). Therefore, we can assume that $\alpha_1 \simeq \alpha_2 \simeq \alpha \equiv ([1+\rho]/2)^{1/2}$. The value of $\bar{P}_B$ for a given ζ is then determined mainly by the correlation coefficient ρ. The higher the ρ value, the greater the $\bar{P}_B$ in equation (24). Therefore, to explain the difference in points ($\bar{P}_C$, $\bar{P}_B$) between the different model trees in figure 3, we only need to note that ρ tends to be higher for a model tree with short exterior branches than for a model tree with long exterior branches for the same value of ζ and, consequently, for the same value of $\bar{P}_C$ (see Appendices A and B).

Let us now consider the frequency distributions of the individual values of P_C and P_B for a starlike model tree (fig. 4). The P_C and P_B values are given separately for the cases of positive and negative values of $\hat{b}_5$. Figure 4 indicates that the bootstrap is conservative test under the null hypothesis $b_5 = 0$ since the distribution of P_B is skewed, and there are less than 5% of cases in which P_B is greater than or equal to 0.95. (A similar result was obtained by Zharkikh and Li [1992*a*, 1992*b*, 1995].) However, the distribution of P_C is uniform (fig. 4*B*). This indicates that the interior-branch test is unbiased. That is, there are about 5% of the cases in which P_C is greater than or equal to 0.95.

Despite the observed difference between the two estimates, we have found that for the majority of four-sequence model trees, the values of $\bar{P}_C$ and $\bar{P}_B$ for $b_5 > 0$ are quite close to each other. However, this is not the case when model trees with a large number of sequences are considered.

Case of Six-Sequence Trees

Let us now consider the relationship of $\bar{P}_C$ and $\bar{P}_B$ for the case of six-sequence trees. We consider P_C to test the null hypothesis $E(\hat{b}_8) = 0$, where $\hat{b}_8$ is the estimate of the branch length b_8 of the true tree (tree A in fig. 1). Here $\hat{b}_8$ and $V(\hat{b}_8)$ are computed by equations (8) and (9), redefining $\mathbf{k}_0 = \{0, -1/4, 5/36, 1/18, 1/18, -1/4, 5/36, 1/18, 1/18, 2/9, 5/36, 5/36, -1/4, -1/4, 0\}$ (see Rzhetsky and Nei 1992*a*). P_B is computed by the same procedure

as mentioned earlier to measure the statistical confidence of the partition of sequences at the branch corresponding to b_8.

The general pattern of the effect of the exterior branch lengths on the difference between $\bar{P}_C$ and $\bar{P}_B$ for the case of six sequences is somewhat similar to that for the case of four sequences (compare fig. 5*A* with fig. 3*A* and *B*). However, figure 5*A* does differ from figure 3 in one important aspect. That is, the values of $\bar{P}_B$ for six-sequence trees are considerably smaller than those for four-sequence trees. Before explaining this result, let us consider the relationships between $\bar{P}_C$ and $\bar{P}_B$ for other model trees (see fig. 5*B*). In this case we have considered three different sets of expected branch lengths of tree *A* in figure 1. The expected lengths of the exterior branches were the same in all cases (see figure legend to fig. 5*B*), but the expected lengths of interior branches were different. We used $b_7 = b_9 = 0$ (fig. 5*B*, solid circles) for the first model tree, $b_7 = 0$ and $b_9 > 0$ (fig. 5*B*, solid squares) for the second model tree, and $b_7 > 0$ and $b_9 > 0$ (fig. 5*B*, solid triangles) for the third model tree. These three types of model trees are schematically shown in figure 1*D*. In all three cases the value of b_8 was gradually increased from 0, and the values of P_C and P_B for the case of $\hat{b}_8 > 0$ were averaged for each value of b_8.

For the first tree (b_7=b_9=0; solid circles in fig. 5*B*), we obtain an already familiar relationship in figure 5*A* (solid circles). In the second case (b_7=0, b_9>0; solid squares in fig. 5*B*), $\bar{P}_B$'s increase. In the third case (b_7>0 and b_9>0; solid triangles in fig. 5*B*), we observe a relationship of $\bar{P}_C$ and $\bar{P}_B$ that is virtually identical with that for a four-sequence tree with short exterior branches (see fig. 3).

To find the reasons for the discrepancy between $\bar{P}_C$ and $\bar{P}_B$, we computed the frequency distributions of P_C and P_B for the case of the lowest point in figure 5*A* (open circle) where $b_7 = b_8 = b_9 = 0$. The distribution of P_B values is even more skewed than in the case of four sequences (compare fig. 6*A* with fig. 4*A*) (see also Zharkikh and Li, 1995). However, P_C is distributed uniformly as in the case of the four sequence model tree (compare fig. 6*B* with fig. 4*B*). Therefore, it is the bootstrap test that is biased (conservative), and the interior-branch test is an unbiased test.

Figure 5*B* indicates that the bias of P_B depends on both the topology and the expected branch lengths of the model tree. The larger the number of sequences (or sequence groups) that behave independently (or nearly independently) in bootstrap tests, the greater the bias of P_B. The lowest point in figure 5*B* (solid circle) corresponds to the six-sequence starlike model tree (fig. 1*D*[*i*]), but the lowest solid-square and the lowest solid-triangle points in figure 5*B* are obtained for the model trees that behave like a five-sequence star tree (see fig. 1*D*[*ii*]) and a four-sequence star tree (see fig. 1*D*[*iii*]), respectively. It can be shown that the bias of P_B tends to increase rapidly as the number of independent or semi-independent sequences in the model tree increases.

Statistical Tests of an Estimated Tree Topology

So far we have considered the interior-branch and bootstrap tests for a predetermined tree topology. How-

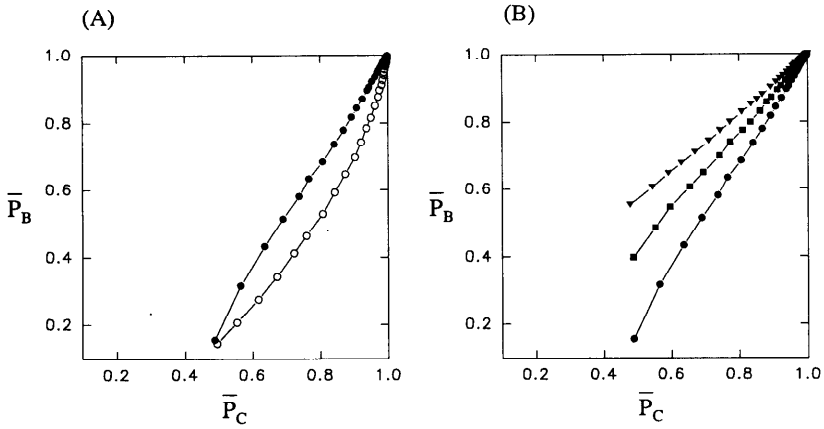

FIG. 5.—Comparison of $\bar{P}_C$ and $\bar{P}_B$ for interior branch b_8 for the case of six-sequence model trees (fig. 1*A*) in simulation for nucleotide sequence data. *A*, Simulations for model trees with the expected exterior branch lengths $b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = 0.07$ and $b_7 = b_9 = 0$ (short branch trees, solid circles) and $b_1 = 0.31$, $b_2 = 0.3$, $b_3 = 0.29$, $b_4 = 0.3$, $b_5 = 0.28$, $b_6 = 0.3$, and $b_7 = b_9 = 0$ (long branch trees, open circles). The value of b_8 was changed from 0 to 0.6 by incrementing it by 0.005 for each model tree of short branches and by incrementing it by 0.015 for each model tree of long branches. $\bar{P}_C$ and $\bar{P}_B$ were obtained by averaging all P_C's and P_B's (given $\hat{b}_8 > 0$) for each b_8 value. Each point was obtained by averaging the results for 3,000 replications. The number of nucleotides sampled (n) was equal to 100. *B*, Simulations for the model trees with the exterior branches $b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = 0.07$ and various values of the interior branch lengths (solid circles: $b_7 = b_9 = 0$; solid squares: $b_7 = 0$ and $b_9 = 0.09$; solid triangles: $b_7 = 0.08$, $b_9 = 0.09$). The value of b_8 in all three cases was increased from 0 to 0.25 by incrementing it by 0.005 for each model tree. Each point was obtained by averaging results of 3,000 replications; $n = 100$.

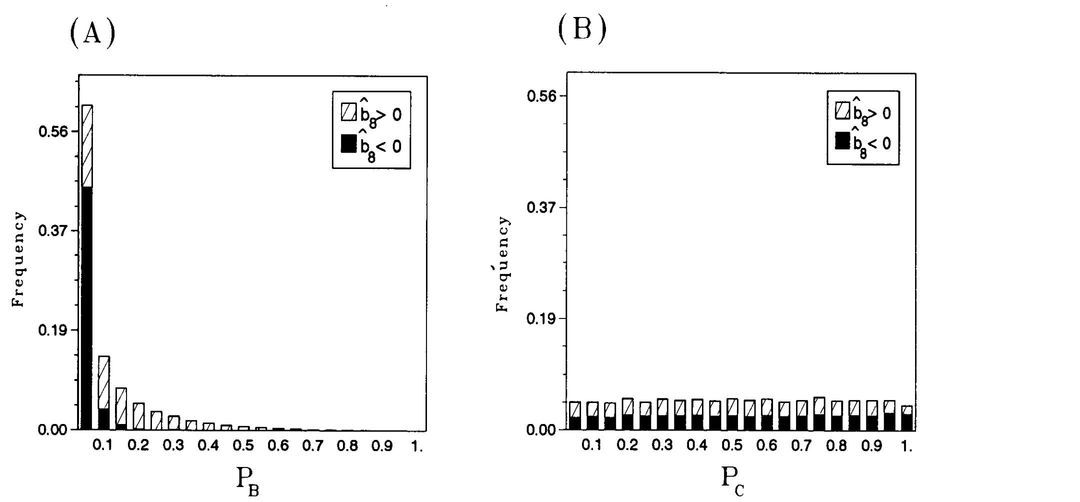

FIG. 6.—Frequency distributions of the individual values of $P_B(A)$ and $P_C(B)$ for the branch b_8 of the model tree in fig. 1*A* for nucleotide sequence data. The frequencies of P_C and P_B corresponding to $\hat{b}_8 > 0$ and $\hat{b}_8 < 0$ are shown separately (corresponding bars are stacked one on top of the other). The expected branch lengths of the tree were $b_1 = 0.31$, $b_2 = 0.3$, $b_3 = 0.29$, $b_4 = 0.3$, $b_5 = 0.28$, $b_6 = 0.3$, and $b_7 = b_8 = b_9 = 0$. The histogram was obtained for data from 30,000 replicate simulations; $n = 100$.

ever, when these tests are applied to a tree estimated from data rather than to a predetermined tree, their statistical properties change. This problem has been studied by Zharkikh and Li (1992*a*, 1992*b*, 1995) in the case of bootstrap tests for parsimony trees. In this paper we consider the same problem for the interior-branch and bootstrap tests for NJ trees.

When the interior-branch test is applied to an estimated tree, matrix **L** in equations (1) and (2) is no longer independent of the estimates of evolutionary distances (i.e., $\hat{d}_{ij}$'s). As a result, the estimates of branch lengths, $\hat{b}_i$'s, may not follow a normal distribution, and the interior-branch test, in which P_C is computed by equation (4), may become too liberal or too conservative depending on the parameter values of the true tree.

In the simplest case of a four-sequence tree (see fig. 2*A*, $b_5 = 0$), the conditional distribution of $\hat{b}_5/s(\hat{b}_5)$ for the case where the NJ tree agrees with topology *A* in figure 2 is close to a gamma distribution (see fig. 7*A*) rather than to the standard normal distribution. This is because the NJ method always chooses a tree with the longest interior branch, and trees that have negative $\hat{b}_5$'s are excluded from consideration. We have found that the conditional distribution of $Z \equiv \hat{b}_i/s(\hat{b}_i)$ varies considerably with the topology and expected branch lengths of the model tree (see fig. 7*B* and *C*). Although it is very difficult to obtain the exact distribution of Z under the null hypothesis of $b_i = 0$ for every interior branch of an estimated tree, one can use the worst-case distribution for conducting statistical tests. Let Z_α be the value of Z such that the probability of obtaining Z greater than Z_α is exactly α (say, $\alpha = 0.05$). Obviously, Z_α varies with model (true) tree. We call the distribution with the highest Z_α among various model trees the worst-case distribution. Our computer simulations have shown that the worst case for four-sequence starlike model trees occurs when all the exterior branches are of equal length (see fig. 7*B*). For five- or more sequence trees the worst-case distribution is observed when one interior branch of the model tree has length 0 and all other interior branches are long (see fig. 7*C*). This worst-case distribution is the same for any number of sequences and coincides with the four-sequence worst-case distribution. (The conditional distribution of Z tends to become narrower as the number of independent sequence groups increases; see fig. 7*C*. Therefore, the four-sequence tree provides the worst-case distribution.) The distribution of Z for the worst case is close to the gamma distribution

$$f(Z) = b^a \Gamma(a)^{-1} e^{-bZ} Z^{a-1}, \qquad (25)$$

where $a = 3.17$ and $b = 3.06$ (see Appendix C). Therefore, the P_C value for the worst case can be computed by

$$P'_C = \int_0^Z b^a \Gamma(a)^{-1} e^{-bx} x^{a-1} dx. \qquad (26)$$

This integral can be evaluated numerically (Press et al. 1988, pp. 171–174). Note that equation (26) tends to give smaller values than equation (4) for the same value of Z. However, the values of P_C and P'_C are usually quite close to each other when they are high. For example, $P_C = 0.95$ and $P'_C = 0.93$ for $Z = 1.96$, and $P_C = 0.99$ and $P'_C = 0.98$ for $Z = 2.58$.

Note that P'_C in equation (26) is defined only for

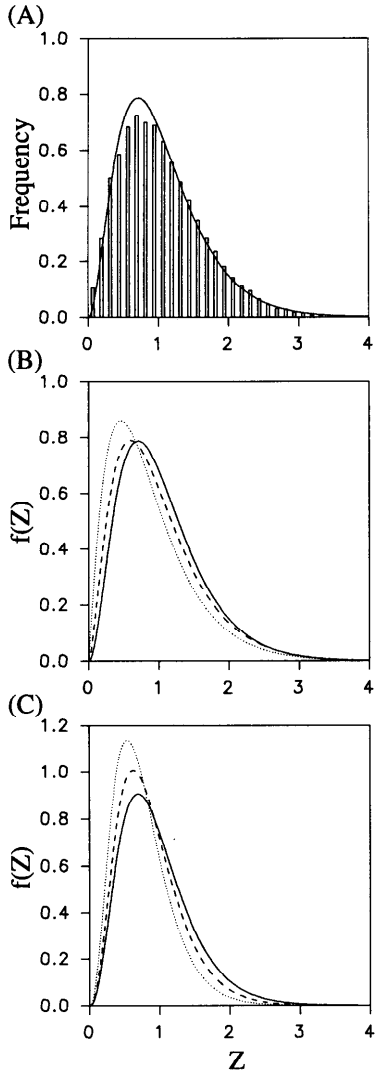

FIG. 7.—*A*, Distribution of $Z \equiv \hat{b}_5/s(\hat{b}_5)$ obtained for the model tree in fig. 2*A* by considering only data sets that generated the NJ tree identical with tree *A* in fig. 2. The histogram was obtained for data from 100,000 replications with $n = 100$. The distribution represented by a smooth curve is the gamma distribution (equation [25]) with parameters $a = 3.10$ and $b = 2.97$. The branch lengths of the model tree were $b_1 = b_2 = b_3 = b_4 = 0.3$ and $b_5 = 0$. "Estimates" of evolutionary distances were computed using a multivariate normal distribution. *B*, Various conditional distributions of Z for four-species model trees. Solid line, $b_1 = b_2 = b_3 = b_4 = 0.3$ and $b_5 = 0$. Dashed line, $b_1 = b_3 = 0.6$, $b_2 = b_4 = 0.02$, and $b_5 = 0$. Dotted line, $b_1 = b_2 = 0.6$, $b_3 = b_4 = 0.02$, and $b_5 = 0$. Here the solid line represents the worst-case distribution of Z. The number of replications was 100,000; $n = 100$. *C*, Conditional distributions of $Z \equiv \hat{b}_8/s(\hat{b}_8)$ for six-sequence model trees (see fig. 1*A*). The distributions of Z for a four-sequence-like tree (solid line, $b_7 = b_9 = 0.15$), a five-sequence-like tree (dashed line, $b_7 = 0.15$ and $b_9 = 0$), and a six-sequence-like tree (dotted line, $b_7 = b_9 = 0$). The expected lengths of exterior branches for six-sequence model trees were $b_1 = 0.31$, $b_2 = 0.3$, $b_3 = 0.29$, $b_4 = 0.3$, $b_5 = 0.28$, $b_6 = 0.3$, and the interior branch length $b_8 = 0$. The number of replications was 1,000,000; $n = 100$.

positive values of Z because in the worst-case distribution all values of Z are positive. In practice, some of branch length estimates of a NJ tree may become negative. However, since the expected values of Z for estimated trees were always positive in our computer simulations, we assume that negative estimates indicate that the expected values of the corresponding branch lengths are close to 0. We therefore suggest that $P'_C = 0$ be assigned to any interior branches with negative estimates.

Using computer simulation, we also studied the conditional distribution of the bootstrap values (P'_B) for the estimated tree that has the same topology as that of the model tree (fig. 8*A* and *B*). At the same time, the P'_C values were computed for the same set of simulated data. This simulation was done for the cases of (1) a four-sequence model tree (fig. 2*A*) with $b_1 = b_2 = b_3 = b_4 = 0.3$, and $b_5 = 0$ (see fig. 8*A*) and (2) a six-sequence model tree (fig. 1*A*) with $b_1 = 0.31$, $b_2 = 0.3$, $b_3 = 0.29$, $b_4 = 0.3$, $b_5 = 0.28$, $b_6 = 0.3$, and $b_7 = b_8 = b_9 = 0$ (see fig. 8*B*). In case (1), the distribution of P'_C is close to a uniform distribution, but that of P'_B is not. This case corresponds to the worst-case distribution of Z discussed above, and the uncorrected P_C test gives slight overestimates of statistical confidence as mentioned above. However, P'_B tends to give underestimates even in this case if a region of $P'_B > 0.9$ is considered. In case (2), in which a six-sequence tree is considered, both P'_C and P'_B becomes underestimates when a region of $P'_C > 0.9$ or $P'_B > 0.9$ is considered. However, the conditional interior-branch test is less conservative than the conditional bootstrap test, and thus the former test gives better test results than the latter.

Discussion

Both P_C and P_B (or P'_C and P'_B) undoubtedly measure the reliability of a branching pattern of a tree, but their properties are quite different. Our computer simulations showed that P_C is a more appropriate measure of statistical confidence than P_B when the tree under analysis is predetermined and that P_B is biased (see also Zharkikh and Li 1992*a*, 1992*b*, 1995; Hillis and Bull 1993).

The first source of bias of P_B is the uncertainty about magnitude of phylogenetic signal in the bootstrap test. Considering information provided by data with regard to a particular branching pattern, we can distinguish among the following three situations. First, sequence data on the average provide no phylogenetic signal concerning the branching pattern if the underlying true tree is multifurcating and the multifurcating topology becomes a bifurcating topology by chance. Second, a "positive phylogenetic signal" is observed on the average when the sequence partition under consideration exists in the true tree ($b_i > 0$). Third, a "negative phylogenetic

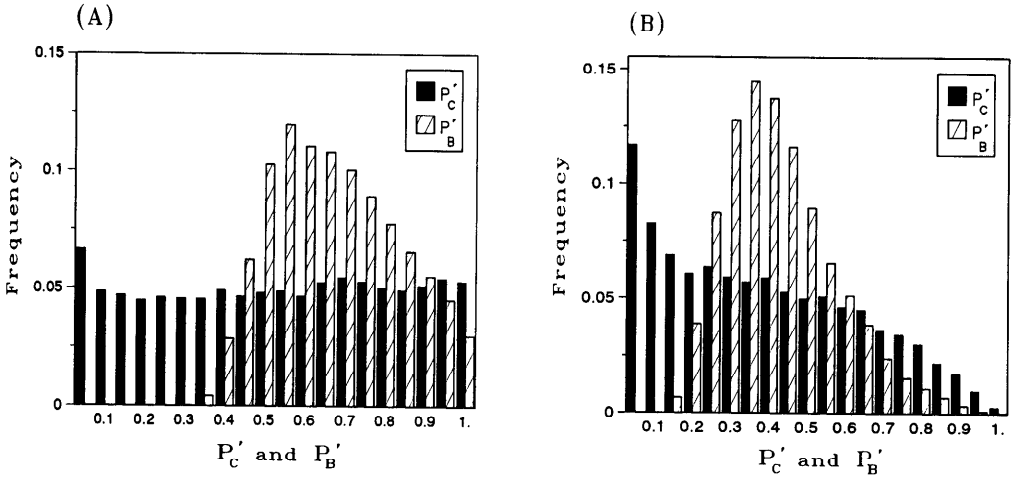

FIG. 8.—Frequency distributions of the P'_C and P'_B values for an estimated tree of which the topology is identical with that of the model tree. *A*, Distributions of P'_C and P'_B for normally distributed distance data obtained from 30,000 replicate simulations. We used the model tree shown in fig. 2*A* with $b_1 = b_2 = b_3 = b_4 = 0.3$, and $b_5 = 0$. *B*, Distributions of P'_C and P'_B for branch b_8 of the model tree in fig. 1*A*. These results were obtained from simulated sequence data with 1,000,000 replications. The expected branch lengths of the tree were $b_1 = 0.31$, $b_2 = 0.3$, $b_3 = 0.29$, $b_4 = 0.3$, $b_5 = 0.28$, $b_6 = 0.3$, and $b_7 = b_8 = b_9 = 0$, and $n = 100$.

signal" occurs when the partition under consideration does not exist in the true tree (incorrect partition).

In the case of star topologies with no phylogenetic signal it is possible to compute the average value ($\bar{P}_B^*$) of P_B. This $\bar{P}_B^*$ depends on the number of sequences in the two groups produced by a partition. If we consider the case in which m sequences are partitioned into two groups containing r and s sequences, respectively, and assume that a tree-building method chooses any tree with an equal probability, we obtain

$$\bar{P}_B^* = \frac{B_{r+1}B_{s+1}}{B_m}, \quad \text{where}$$
$$B_i = \prod_{j=1}^{i-2} (2j-1), \quad \text{and } i \geq 3. \tag{27}$$

Note that $\bar{P}_B^*$ is equal to 0.33 for $r = s = 2$ for the case of a four-sequence star tree and 0.2 for $r = 2$ and $s = 3$ in the case of a five-sequence star tree. Equation (27) indicates that the value of $\bar{P}_B^*$ tends to be 0 as the number of sequences increases.

However, the computation of $\bar{P}_B^*$ becomes quite complicated when some of the interior branches of the true tree have positive expected lengths (see fig. 5*B*). In this case some groups of sequences appear together in the majority of bootstrap replications, and thus the $\bar{P}_B^*$ for this grouping increases. To make things worse, the probabilities of recovering various trees containing a particular partition may not always be equal to one another (see Appendix A). Therefore, for every sequence partition of a large tree there is a unique $\bar{P}_B^*$ value that is determined by numerous factors, and there is no easy way to compute this value. So, it is difficult to say whether the individual P_B (say, P_B=0.2) indicates an absence of phylogenetic signal (which would be the case if $\bar{P}_B^* = 0.2$ for $r = 2$ and $s = 3$ in equation [27]), negative phylogenetic signal (if $\bar{P}_B^*$=0.33 for r=s=2), or positive phylogenetic signal (if $\bar{P}_B^*$=0.09 for r=s=3).

Clearly, the bias of the bootstrap estimate becomes more extreme as $\bar{P}_B^*$ decreases (see fig. 4*A* and fig. 6*A*). Curiously enough, the bootstrap value is unbiased only in the artificial case of two alternative trees as described by Felsenstein and Kishino (1993).

The interior-branch test for a predetermined topology is inherently free from this kind of bias. In the absence of phylogenetic signal (b_i=0) the probability of observing a P_C that is greater than or equal to 0.95 is exactly 0.05 regardless of the parameters of the true tree (see fig. 4*B* and fig. 6*B*).

There is another type of bias mentioned in the literature concerning the application of the bootstrap technique in phylogenetic analysis. One might think that P_B is an estimate of the probability of obtaining the true tree (Zharkikh and Li 1992*a*, 1995; Hillis and Bull 1993; Felsenstein and Kishino 1993). As we have seen earlier, P_B is a function of estimates of five parameters in the case of four-sequence trees ($\hat{D}_1$, $\hat{D}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\rho}$) (see equation [23]). In general, we can write $P_B = P_B(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is a vector of estimates of all parameters involved. We can then consider P_B as an estimate of $P_B[E(\hat{\boldsymbol{\theta}})]$, where $P_B[E(\hat{\boldsymbol{\theta}})]$ is equal to the probability of obtaining the true tree in the case of four sequences.

Clearly, P_B should be a biased estimate of this value just because P_B is a nonlinear function of $\hat{\boldsymbol{\theta}}$. It is this kind of bias that was studied by Felsenstein and Kishino (1993). As shown earlier, for a four-sequence model tree with short exterior branches, P_B is essentially a function of only two parameters, $\hat{D}$ and $\hat{\sigma}^2$ ($\hat{D}=\hat{D}_1 \simeq \hat{D}_2$, $\hat{\sigma}^2=\hat{\sigma}_1^2\simeq\hat{\sigma}_2^2$, and $\rho\simeq 1$). However, as the exterior branches of the model tree increase, the number of parameters that affect the value of P_B increases, and so does the bias of P_B. In the case of large trees, this effective number of parameters tends to increase drastically. Consequently, the larger the effective number of parameters, the greater the bias of this kind.

Some authors attempted to correct for the first type of bias for P_B. For this purpose Rodrigo (1993) proposed an application of the "iterated bootstrap" method (Hall and Martin 1988), whereas Zharkikh and Li (1995) developed the so-called "complete-and-partial" bootstrap technique. The efficiency of these techniques deserves additional study.

One should be aware that for each test there are certain assumptions that may not be satisfied in practice. In particular, the bootstrap test requires the assumption that the same evolutionary process operates independently at each nucleotide site. This is not true for many real sequence data. In principle, the interior-branch test can overcome this problem provided that there is a mathematical model that takes care of these biological peculiarities in the estimation of evolutionary distances. For example, the heterogeneity of substitution rate among different nucleotide sites may easily be taken care of by using Jin and Nei's (1990) gamma distances. Theoretically, similar problems may be taken care of in bootstrap tests as well by developing a new sampling procedure. In practice, however, the sampling procedure seems to be quite complicated.

The interior-branch test for a predetermined tree depends on the assumption of a normal distribution of branch length estimates. This assumption may appear to be violated when the nucleotide sequences under study are very short (<100 characters) so that the discreteness of nucleotide substitutions strongly affects the distribution of branch length estimates. However, our computer simulations showed that the assumption of normal approximation rapidly improves as the number of sequences and/or their length increases. Note also that the interior-branch test can be modified to relax the requirement of the normal distribution of branch length estimates (e.g., see Dopazo 1994). However, both the bootstrap and the interior-branch methods may lead to erroneous conclusions when inappropriate (biased) estimators of evolutionary distances are used. When the interior-branch test is applied to an estimated topology, we recommend that P'_C rather than P_C be used to avoid the possible overestimation of statistical confidence fo a sequence clusters by this test. In practice, however P'_C and P_C are close to each other when a region o $P'_C > 0.9$ or $P_C > 0.9$ is considered.

Acknowledgments

We are grateful to Andrey Zharkikh for stimulatin discussions. We also thank Susan Murphy, Bruce Lind say, and Spencer Muse for valuable comments. This worl was supported by grants from the National Institutes o Health and National Science Foundation to M.N.

APPENDIX A

Relationships between the Expected Branch Lengths of the Model Tree and the Values of σ_1, σ_2, and ρ

Since the values of σ_1^2, σ_2^2, and $\rho\sigma_1\sigma_2$ (the variance and covariance of $\hat{D}_1 = \hat{S}_B - \hat{S}_A$ and $\hat{D}_2 = \hat{S}_C - \hat{S}_A$ respectively) determine the probability of recovering th correct tree (see equation [22]), it is important to un derstand their dependence on the branch lengths of th model tree.

Using equations (11), (12), and (21), we find th following equation after some algebraic manipulation.

$$\sigma_1^2 = [g(b_1, b_2)+g(b_3, b_4)+g(b_1, b_5, b_3) + g(b_2, b_5, b_4)+4f(b_5)]/16, \quad \text{(A1a}$$

$$\sigma_2^2 = [g(b_1, b_2)+g(b_3, b_4)+g(b_1, b_5, b_4) +g[b_2, b_5, b_3)+4f(b_5)]/16, \quad \text{(A1b}$$

$$\rho\sigma_1\sigma_2 = [g(b_1, b_2)+g(b_3, b_4)+g(b_1, b_5) + g(b_2, b_5)+g(b_3, b_5)+g(b_4, b_5) +4f(b_5)]/16, \quad \text{(A1c}$$

where

$$g(b_i, b_j) = f(b_i+b_j) - [f(b_i)+f(b_j)], \quad \text{(A2a}$$

$$g(b_i, b_j, b_k) = f(b_i+b_j+b_k) - [f(b_i)+f(b_j)+f(b_k)], \quad \text{(A2b}$$

and function $f(\delta)$ is the same as that defined in equatio (12). An analysis of equations (A1a–b) shows that th values of σ_1 and σ_2 tend to be close to each other, an ρ approaches a value close to 1 as b_5 increases. Thi holds for a wide range of values of b_1, b_2, b_3, and b_4.

Using equation (A1) we can directly compar the values of ρ for a given expected value of $\hat{} \equiv \hat{b}_5/s(\hat{b}_5)$ [$E(Z) = b_5/\sigma$] for four-sequence model tree with "short" and "long" exterior branches (see fig. 3) Figure 9 shows that ρ increases faster for trees with shor exterior branches (solid circles) than for trees with lon exterior branches (open circles). This explains the dif

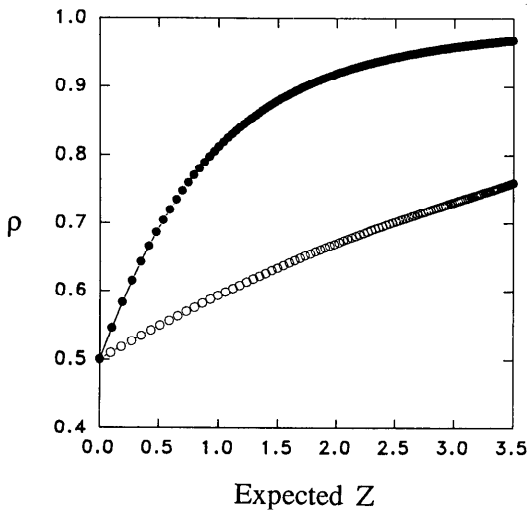

FIG. 9.—Relationships between ρ and the expected value of Z for four-sequence model trees with short (solid circles) and long (open circles) exterior branches. The actual values of the expected exterior branch lengths are the same as in fig. 3. The values of ρ and Z increase as the length of interior branch (b_5) of the model tree in fig. 2*A* increases.

ferences in the relationships of $\bar{P}_C$ and $\bar{P}_B$ between these two types of model trees in figure 5.

In the case of a star tree ($b_5 = 0$) equations (A1a–c) reduce to

$$\sigma_1^2 = [g(b_1, b_2)+g(b_3, b_4) +g(b_1, b_3)+g(b_2, b_4)]/16, \quad \text{(A3a)}$$

$$\sigma_2^2 = [g(b_1, b_2)+g(b_3, b_4) +g(b_1, b_4)+g(b_2, b_3)]/16, \quad \text{(A3b)}$$

$$\rho\sigma_1\sigma_2 = [g(b_1, b_2)+g(b_3, b_4)]/16. \quad \text{(A3c)}$$

From the above equations, we obtain

$$\frac{\sigma_1}{\sigma_2} = \left[\frac{g(b_1, b_2)+g(b_3, b_4)+g(b_1, b_3)+g(b_2, b_4)}{g(b_1, b_2)+g(b_3, b_4)+g(b_1, b_4)+g(b_2, b_3)}\right]^{1/2}, \quad \text{(A4)}$$

and

$$\rho = \left\{\left[1 + \frac{g(b_1, b_3) + g(b_2, b_4)}{g(b_1, b_2) + g(b_3, b_4)}\right] \times \left[1 + \frac{g(b_1, b_4) + g(b_2, b_3)}{g(b_1, b_2) + g(b_3, b_4)}\right]\right\}^{-1/2}. \quad \text{(A5)}$$

To study the values of σ_1/σ_2 and ρ, let us rewrite the function $g(b_i, b_j)$ as

$$g(b_i, b_j) = \frac{3}{16n}\{(e^{4b_i/3}-1)(e^{4b_j/3}-1) \times [2+(e^{4b_i/3}+1)(e^{4b_j/3}+1)]\}, \quad \text{(A6)}$$

combining equations (A2) and (12). Considering the first three terms of Taylor's series expansion and assuming that b_i is much smaller than ¾, we can approximate $e^{4b_i/3}$ by $1 + 4b_i/3 + 8b_i^2/9$. After some algebraic computation we can express equation (A6) as

$$g(b_i, b_j) = 2b_ib_j/n + O(b_i^r b_j^s), \quad \text{(A7)}$$

where $r + s = 3$. Next, we replace b_i by $b + \delta_i$ (i=1, 2, 3, and 4), where b is the average value of b_1, b_2, b_3, and b_4 and $\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0$. We then obtain the following simplified equations for $g(b_i, b_j)$, σ_1/σ_2 and ρ.

$$g(b_i, b_j) \simeq 2[b^2+b(\delta_i+\delta_j)+\delta_i\delta_j]/n, \quad \text{(A8)}$$

$$\frac{\sigma_1}{\sigma_2} \simeq \left[\frac{1-(\delta_2+\delta_3)^2/(4b^2)}{1-(\delta_1+\delta_3)^2/(4b^2)}\right]^{1/2}, \quad \text{(A9)}$$

and

$$\rho \simeq \frac{1 + [(\delta_3-\delta_1)(\delta_3-\delta_2)-2\delta_3^2]/(2b^2)}{2[1-(\delta_2+\delta_3)^2/(4b^2)]^{1/2} \times [1-(\delta_1+\delta_3)^2/(4b^2)]^{1/2}}. \quad \text{(A10)}$$

Equation (A9) indicates that the ratio σ_1/σ_2 is very close to 1 whenever δ_1 is close to δ_2, or δ_3 is close to δ_4, or all δ_i's are much smaller than b. It follows from equation (A10) that $\rho = 1/2$ when all δ_i's are equal to 0 and that ρ is close to ½ whenever δ_i's are much smaller than b.

We also see from equation (22) that the probability of recovering the tree in figure 2*A* when $b_5 = 0$ is determined solely by the value of ρ ($\mu_1=\mu_2=b_5/2=0$),

$$P = F(0, 0; \rho) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho. \quad \text{(A11)}$$

(Here we used the formula that was first derived by Sheppard [1899].) Note that for $\rho = 1/2$ we have $P = 1/3$. However, as ρ deviates from ½, three possible unrooted trees for four sequences are no longer equiprobable. On the one hand, when the values of b_1 and b_2 are much larger than the values of b_3 and b_4, the value of ρ tends to be 1 and P tends to be ½. On the other hand, when the values of b_1 and b_3 are very large and the values of b_2 and b_4 are very small, ρ tends to be 0 and P tends to be ¼.

This implies that one may encounter a situation in which the NJ method tends to choose an incorrect topology. In other words, we can find a four-sequence model tree with extremely short interior branch and two

extremely long exterior branches such that the probability of obtaining the correct tree by the NJ method from a small data set tends to be smaller than 1/3. However, this effect disappears as the number of nucleotides increases.

APPENDIX B
Expected Value of P_C when Z Is Positive

Let $\hat{b}_i$ and $V(\hat{b}_i)$ be the estimates of the length of the ith interior branch and its variance with the expected values b_i and σ^2, respectively. The P_C for this branch is computed by equation (4) in the text. Assuming that the test statistic $Z = \hat{b}_i / s(\hat{b}_i)$ follows the standard normal distribution and noting that conditions $Z > 0$ and $\hat{b}_i > 0$ are equivalent, we can obtain a formula for the mean value of P_C for the case of $\hat{b}_i > 0$:

$$E(P_C|\hat{b}_i>0) = E(P_C|Z>0) = \frac{\int_0^\infty P_C(x)\phi(x;\zeta,1)dx}{\text{Prob}(Z>0)}, \quad \text{(B1)}$$

where $\phi(x;\zeta,1)$ is a normal density function with mean $\zeta = b_i/\sigma$ and variance 1,

$$\text{Prob}(Z>0) = \Phi(\zeta), \quad \text{(B2)}$$

and

$$\begin{aligned}\int_0^\infty P_C(x)\phi(x;\tau,1)dx &= \frac{1}{\sqrt{2\pi}}\int_0^\infty [2\Phi(x)-1]\exp(-(x-\zeta)^2/2)dx \\ &= \pi^{-1}\int_0^\infty \exp\{-(x-\zeta)^2/2\} \\ &\quad \times \left(\int_{-x}^\infty \exp(-y^2/2)dy\right)dx - \Phi(\zeta).\end{aligned} \quad \text{(B3)}$$

Replacing variables x and y by $r = x - \zeta$ and $s = (x-\zeta+y)/\sqrt{2}$, we can simplify the above equation to

$$\pi^{-1}\sqrt{2}\int_{-\zeta}^\infty \int_{-\zeta/\sqrt{2}}^\infty \exp\{-(r^2-\sqrt{2}rs+s^2)\}drds - \Phi(\zeta) = 2F(-\zeta, -\zeta/\sqrt{2}; 1/\sqrt{2}) - \Phi(\zeta), \quad \text{(B4)}$$

where $F(-\zeta, -\zeta/\sqrt{2}; 1/\sqrt{2})$ is a function given in equation (22). Finally, we obtain

$$E(P_C|Z>0) = \frac{2F(-\zeta, -\zeta/\sqrt{2}; 1/\sqrt{2})}{\Phi(\zeta)} - 1. \quad \text{(B5}$$

For a starlike model tree we have $\zeta = 0$, and equation (B5) reduces to

$$E(P_C|Z>0, \zeta=0) = 2\sin^{-1}(1/\sqrt{2})/\pi = 0.5. \quad \text{(B6}$$

(Here we used equation [A11].) Equation (B5) shows that P_C is an unbiased estimate of $\text{Prob}(\hat{b}_i>0)$ for a starlike model tree (i.e., for $b_i = 0$). P_C is biased when $b_i > 0$. For example, equation (B5) shows that we have $E(P_C|Z>0) = 0.81$ when $\text{Prob}(\hat{b}_i>0) = 0.95$.

APPENDIX C
The Worst-Case Distribution of Z for an Estimated Topology

The worst-case distribution of $Z \equiv \hat{b}_i / s(\hat{b}_i)$ is observed for the case of a four-sequence star tree with exterior branches of equal length. In order to approximate this distribution by a gamma distribution we need to know the mean and variance of Z. The shape and scale parameters of the gamma distribution are then given by

$$a = \frac{E^2(Z)}{V(Z)}, \quad \text{and } b = \frac{E(Z)}{V(Z)}, \quad \text{(C1}$$

respectively, where $E(Z)$ and $V(Z)$ stand for the *conditional* mean and variance of Z, respectively, when all data sets that give an NJ tree different from tree A in figure 2 are discarded.

Using the same notations as in equations (17) and (22), we can write Z as $(\hat{D}_1 + \hat{D}_2)h$, where $h = 1/s(\hat{b}_5)$. Assuming that $\hat{D}_1$ and $\hat{D}_2$ follow a bivariate normal distribution with mean vector (0, 0) and the covariance matrix given in equation (20), we can compute the mean of Z by

$$\begin{aligned}E(Z) &= \frac{h}{2\pi\sigma_1\sigma_2 P\sqrt{1-\rho^2}} \\ &\quad \times \int_0^\infty \int_0^\infty (x+y)\exp\left\{-\frac{1}{2(1-\rho^2)}\right. \\ &\quad \left.\times\left(\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right\}dxdy \\ &= \frac{h(1+\rho)(\sigma_1+\sigma_2)}{2P\sqrt{2\pi}},\end{aligned} \quad \text{(C2}$$

where P is the probability computed by equation (22). Equation (C2) is derived by using the formula for incomplete moments of a bivariate normal distribution (Johnson and Kotz 1972, p. 92). Similarly, we can compute $E(Z^2)$ and $V(Z)$:

$$E(Z^2) = h^2\{(\sigma_1^2+\sigma_2^2)[1+\rho\sqrt{1-\rho^2}/(2\pi P)] + 2\sigma_1\sigma_2[\rho+\sqrt{1-\rho^2}/(2\pi P)]\}. \quad \text{(C3)}$$

$$V(Z) = E(Z^2) - E^2(Z) = h^2\sigma_1^2\{(1+\alpha^2)[1+\rho\sqrt{1-\rho^2}/(2\pi P)] + 2\alpha[\rho+\sqrt{1-\rho^2}/(2\pi P)] - (1+\alpha)^2(1+\rho)^2/(8\pi P^2)\}, \quad \text{(C4)}$$

where $\alpha = \sigma_2/\sigma_1$. In the present case $\rho = 1/2$, $\alpha = 1$, $P = 1/3$, and $h = 1/(\sigma_1\sqrt{3})$. Therefore, we have $E(Z) = 3\sqrt{3}/(2\sqrt{2\pi})$ and $V(Z) = 1 + 3(2\sqrt{3}-9)/(8\pi)$. Using equation (C1), we obtain $a = 3.17$ and $b = 3.06$.

LITERATURE CITED

BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

DOPAZO, J. 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J. Mol. Evol. **38**:300–304.

DREZNER, Z., and G. O. WESOLOWSKY. 1990. On the computation of the bivariate normal integral. J. Stat. Comput. Simul. **35**:101–107.

EFRON, B. 1982. The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia.

FELSENSTEIN, J. 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.

———. 1991. PHYLIP: phylogeny inference package, version 3.4. University of Washington, Seattle.

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.

HALL, P., and M. A. MARTIN. 1988. On bootstrap resampling and iteration. Biometrika **75**:661–671.

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

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JOHNSON, N. L., and S. KOTZ. 1972. Distributions in statistics: continuous multivariate distributions. Wiley, New York.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. M. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M., and T. OHTA. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2**:87–90.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetic analysis, version 1.0. Pennsylvania State University, University Park.

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. **6**:424–435.

LI, W.-H., and M. GOUY. 1991. Statistical methods for testing molecular phylogenies. Pp. 249–277 *in* M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

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

NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

PAMILO, P. 1990. Statistical tests of phenograms based on genetic distances. Evolution **44**:689–697.

PENNY, D., and M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. **34**:75–82.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, and W. T. VETTERLING. 1988. Numerical recipes in *C*. The art of scientific computing. Cambridge University Press, Cambridge.

RODRIGO, A. 1993. Calibrating the bootstrap test of monophyly. Int. J. Parasitol. **23**:507–514.

RZHETSKY, A., and M. NEI. 1992*a*. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**:945–967.

———. 1992*b*. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. **35**:367–375.

———. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073–1095.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SHEPPARD, W. F. 1899. On the application of the theory of error to cases of normal distribution and normal correlation. Philos. Trans. R. Soc. Lond., Series A **192**:101–167.

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

———. 1992*b*. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. J. Mol. Evol. **35**:356–366.

———. 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phyl. Evol. (accepted).

TAKASHI GOJOBORI, reviewing editor

Received May 13, 1994

Accepted September 30, 1994