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, $\geq 0.95$), 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 ($P_{Ck}$) of the $P_C$-value and showed that this $P_{Ck}$ 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 1992a) 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
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, \( d_{ij} \)'s, where \( i \) and \( j \) refer to the \( i \)th and \( j \)th sequences, respectively. The estimates \( \hat{b}_i \)'s) of branch lengths for the tree are then given by

\[
\hat{b} = L \hat{d},
\]

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

\[
\hat{V}(\hat{b}) = L' \hat{W} L.
\]

In practice, the computation of \( \hat{b}_i \)'s and \( \hat{V}(\hat{b}_i) \)'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:

\[
Z = \frac{\hat{b}_i}{s(\hat{b}_i)},
\]

where \( s(\hat{b}_i) = \sqrt{\hat{V}(\hat{b}_i)} \).

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 \( \alpha \) 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:

\[
P_C = 2\Phi(|Z|) - 1,
\]

where

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

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

Note that the above statistical test applies when topology to be tested is predetermined. This situation occurs when an investigator is interested in the reliability of a particular topology. However, the interior-branch test is often applied to a phylogenetic tree estimated for actual data rather than to a predetermined tree. In this case the statistical properties of the \( P_C \) test become different. Later we will consider a correction for \( P_C \) necessary for this case.

**Interior-Branch Test for Large Trees**

In the case of four-sequence trees the expected values of the estimate of the interior branch length can be either positive (for the true tree) or negative (for a wrong tree if the true tree is not starlike). Hence, the estimate of the interior branch length gives some idea about the validity of a tree. In the case of a large number of sequences we can test the null hypothesis \( E(\hat{b}_i) = 0 \) for the \( i \)th interior branch, but the interpretation of the test outcome somewhat more complicated. It can be shown that if the expectation of the length estimate of an incorrect interior branch (which gives a sequence partition that is not present in the true tree) in a large tree can be positive \((E[\hat{b}_i] > 0)\). However, a negative \( E(\hat{b}_i) \) always indicates that the corresponding partition is wrong. We illustrate this point with the following example of a six-sequence tree.

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

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

\[
E(\hat{b}_8) = -\frac{5}{18} (2b_8 + b_9),
\]

and

\[
E(\hat{b}_9) = \frac{1}{6} (4b_8 - b_9).
\]
Similarly, for tree C, we have

\[
E(b_7) = \frac{1}{4} (-b_7 + 2b_8 + b_9),
\]

\[
E(b_8) = \frac{1}{4} (b_7 + 2b_8 - b_9),
\]

and \( E(b_9) = -\frac{1}{4} (b_7 + 2b_8 + b_9) \).

Therefore, (1) the value of \( E(b_7) \) for tree B is always positive and the corresponding interior branch gives the correct partition of sequences, (2) the values of \( E(b_8) \) for tree B and \( E(b_9) \) for tree C are always negative and both interior branches give an incorrect partition of sequences, and (3) the values of \( E(b_8) \) for tree B and \( E(b_9) \) 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 \( (b_i) \), we shall compare \( P_C \) and its equivalent quantity for the bootstrap test for the positive and negative values of \( b_i \) 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 1992a, 1992b, 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 1992a) of the original tree, the proportion of bootstrap trees in which this cluster appears is computed. We call this value \( (P_B) \) 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 boot-
strap 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 propor-
tion 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 1992a, 1992b, 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 in-
terior 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 phy-
logenetic inference, since it occurs only for small values
of $P_B$.) All the above statements about statistical prop-
erties 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 inter-
preted 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 2A. Second, gen-
erate four "extant" nucleotide sequences of length $n$ ac-
ccording to the model tree as in Rzhetsky and Nei
(1992b). Third, estimate the vector of evolutionary dis-
tances between sequences, $\hat{d} = (d_{12}, d_{13}, d_{14}, d_{23},
d_{24}, d_{34})$, using the Jukes-Cantor method. The variance-co-
variance matrix ($W$) of the vector $\hat{d}$ is then estimated
by using Kimura and Ohta’s (1972) formula for the
variances and Bulmer’s (1971) formula for the covari-
ances. Fourth, estimate the length of the interior branch
of the true tree (tree A in fig. 2) by the equation

$$\hat{b}_5 = k_0 \hat{d}, \quad \text{where } k_0 = (-1/2, 1/4, 1/4, 1/4, -1/2),$$

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

$$V(\hat{b}_5) = k_0 W k_0^T,$$

where $W$ is the estimate of $W$ (Rzhetsky and Nei 1992b).
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 con-
sideration appears). In all simulations discussed below we
used $n = 100$ and $B = 500$ unless otherwise stated.

**Normally Distributed Distance Data**

The joint distribution of $d_{ij}$’s computed above is
slightly asymmetrical compared with a multivariate
normal distribution. To evaluate the effect of this skew-
ness on the performance of the interior-branch test, we
conducted another type of simulation, in which the vec-
tor of the distance estimates, $\hat{d}$, exactly followed a mul-
tivariate normal distribution. In this simulation all ass-
sumptions underlying the interior-branch test are satis-
fied, 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 dis-
tances (steps 2 and 3 in the previous section). We used
the following method to generate a vector of estimates of
"evolutionary distances" ($\hat{d}$) following a multivariate
normal distribution. The expected distances for this
multivariate normal distribution were given by

$$d_{12} = b_1 + b_2, d_{13} = b_1 + b_3 + b_5,$$

$$d_{14} = b_1 + b_3 + b_4, d_{23} = b_2 + b_5 + b_3,$$

$$d_{24} = b_2 + b_3 + b_4, \text{ and } d_{34} = b_3 + b_4,$$

where $b_i$’s are the expected branch lengths of the model tree
(see fig. 2A), and the variance-covariance matrix, $W$, was

![Fig. 2.—A, Hypothetical true tree for four sequences. B and C, Incorrect trees for the same sequences.](https://example.com/fig2.png)
$$W = \begin{bmatrix} f(b_1 + b_2) & f(b_1) & f(b_1) \\ f(b_1) & f(b_1 + b_3) & f(b_1 + b_3) \\ f(b_1) & f(b_1 + b_3) & f(b_1 + b_3) \\ f(b_2) & f(b_2 + b_5) & f(b_2 + b_5) \\ f(b_2) & f(b_2 + b_5) & f(b_2 + b_5) \\ 0 & f(b_3) & f(b_3) \end{bmatrix}.$$  

(11)

Here

$$f(\delta) = (-9 + 6e^{45/3} + 3e^{88/3})/(16n), \quad (12)$$

where $\delta$ 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 $W$ is the same as the variance-covariance matrix of vector $d$ for nucleotide sequence data generated by the Jukes-Cantor model following the model tree in figure 2A (see Rzhetsky and Nei 1992b). To generate observed distances, $d_i$'s, we computed a triangular matrix $C$ such that

$$W = C'C \quad (13)$$

(13)

(for details, see Press et al. 1988, pp. 39–45). We then computed $n$ random vectors, $x_k = (x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5}, x_{k6})$, where $k = 1, 2, \ldots, n$, and $x_k$'s are independent random variables drawn from a standard normal distribution. We generated $x_k$'s using the Box-Muller method (see Press et al. 1988, pp. 216–217). We used vectors $x_k$'s to obtain a set of random vectors $d_1, d_2, \ldots, d_n$, by

$$\hat{d}_k = d + Cx_k \sqrt{n}, \quad (14)$$

that for nucleotide sequence data (see steps 4 and 5 in the previous section). The vectors $d_1, d_2, \ldots, d_n$ 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 $PB$ 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 $S_A$, $S_B$, and $S_C$, respectively, we can introduce the following variables, $D_1$ and $D_2$:

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

Therefore, the NJ method recovers the correct tree (tree $A$) whenever both $D_1$ and $D_2$ are positive. We can compute $\hat{D}_1$ and $\hat{D}_2$ by

$$\hat{D}_1 = k_1d, \quad \hat{D}_2 = k_2d, \quad (18)$$

where

$$k_1 = (-1/4, 1/4, 0, 0, -1/4), \quad \text{and} \quad k_2 = (-1/4, 0, 1/4, 0, -1/4). \quad (19)$$

When $d_i$'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 $\mu_1 = k_1d$ and $\mu_2 = k_2d$, and the variance-covariance matrix

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

where $\sigma_1^2$ and $\sigma_2^2$ are the variances of $\hat{D}_1$ and $\hat{D}_2$, re-
spective, $\rho \sigma_1 \sigma_2$ is the covariance for $D_1$ and $D_2$, and $\rho$ is the correlation coefficient between $D_1$ and $D_2$. The elements of matrix $V$ are given by

$$
\sigma_1^2 = k_1 Wk_1^t, \quad \sigma_2^2 = k_2 Wk_2^t, \quad \text{and} \quad \rho \sigma_1 \sigma_2 = k_1 Wk_2^t,
$$

where matrix $W$ is computed by equations (11) and (12). Since the correct tree is recovered when $D_1 > 0$ and $D_2 > 0$, the probability ($P$) of obtaining the true tree (tree A in fig. 2) by the NJ method can be computed by

$$
P = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \int_0^\infty \int_0^\infty \exp \left\{ -\frac{1}{2(1-\rho^2)} \cdot \left( \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1 \sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right\} dxdy
$$

$$
= \frac{1}{2\pi \sqrt{1 - \rho^2}} \int_0^\infty \int_0^\infty \exp \left\{ -\frac{1}{2(1-\rho^2)} \cdot (x^2 - 2\rho xy + y^2) \right\} dxdy
$$

$$
= F \left( -\hat{\mu}_1 \frac{\mu_2}{\sigma_1}, -\frac{\mu_1}{\sigma_2}; \hat{\rho} \right).
$$

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 $(\hat{D}_1, \hat{D}_2)$ and variance-covariance matrix $\hat{V}$. Here, $\hat{D}_1$ and $\hat{D}_2$ are computed by equations (18) and (19) from the original set of data, and $\hat{V}$ 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

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

Our numerical study has shown that this approximation works very well even when vector $\hat{d}$ 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 2A when $\hat{b}_5$ 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 3A 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 $\hat{b}_5$), 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. 2A: $b_1 = 0.07, b_2 = 0.07, b_3 = 0.065, b_4 = 0.07$ (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. 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 $P_C$ (Appendix B), where $P_C$ is determined by $\zeta = b_5/\sigma$ (where $\sigma = \sqrt{\sigma^2}$).

To obtain an equivalent equation for $P_B$, we use equation (23), the relationship $\hat{b}_5 = \hat{D}_1 + \hat{D}_2$ (see equations [8], [18], and [19]), and $V(b_5) = \sigma^2_1 + 2\sigma_1\sigma_2 + \sigma^2_2$. To simplify the analysis, we replace $\bar{\sigma}_1$, $\bar{\sigma}_2$, and $\bar{\rho}$ in equation (23) by their expected values $\bar{\sigma}_1$, $\bar{\sigma}_2$, and $\bar{\rho}$. Then, after some algebraic manipulations we obtain the following equation:

$$\hat{P}_B \approx \frac{1}{\Psi(\zeta)2\pi(1-\rho^2)^{1/2}}$$

$$\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-(x_1+\zeta\sigma_1)^2 - 2\rho(x_1x_2+x_2^2)/2 - (1-\rho^2)\right\}$$

$$\times F[-(x_1+\zeta\sigma_1), -(x_2+\zeta\sigma_2); \rho] \, dx_1 \, dx_2,$$

(24)

where $\zeta = b_5/\sigma$, $\sigma_1 = (1+2\rho\sigma_2/\sigma_1+\sigma^2_2/\sigma^2_1)^{1/2}/2$, $\sigma_2 = (1+2\rho\sigma_1/\sigma_2+\sigma^2_1/\sigma^2_2)^{1/2}/2$, 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 $\sigma_1$ and $\sigma_2$ are very close to each other for the majority of the model trees (see Appendix A). Therefore, we can assume that $\sigma_1 \approx \sigma_2 \approx \sigma = (1+\rho)/2$.

The value of $\hat{P}_B$ for a given $\zeta$ is then determined mainly by the correlation coefficient $\rho$. The higher the $\rho$ value, the greater the $\hat{P}_B$ in equation (24). Therefore, to explain the difference in points ($\hat{P}_C$, $\hat{P}_B$) between the different model trees in figure 3, we only need to note that $\rho$ 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 $\zeta$ and, consequently, for the same value of $\hat{P}_C$ (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 $b_5$. Figure 4 indicates that the bootstrap is a 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 [1992a, 1992b, 1995].) However, the distribution of $P_C$ is uniform (fig. 4B). 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 $\hat{P}_C$ and $\hat{P}_B$ 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 $\hat{P}_C$ and $\hat{P}_B$ for the case of six-sequence trees. We consider $P_C$ to test the null hypothesis $E(b_5) = 0$, where $b_5$ is the estimate of the branch length $b_5$ of the true tree (tree $\Lambda$ in fig. 1). Here $b_5$ and $V(b_5)$ are computed by equations (8) and (9), redefining $k_0 = \{0, -\gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8, -\gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8, -\gamma_4, -\gamma_4, 0\}$ (see Rzhetsky and Nei 1992d). $P_B$ 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. 5A with fig. 3A and B). However, figure 5A 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. 5B). 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. 5B, solid circles) for the first model tree, $b_7 = 0$ and $b_9 > 0$ (fig. 5B, solid squares) for the second model tree, and $b_7 > 0$ and $b_9 > 0$ (fig. 5B, solid triangles) for the third model tree. These three types of model trees are schematically shown in figure 1D. In all three cases the value of $b_8$ was gradually increased from 0, and the values of $\bar{P}_C$ and $\bar{P}_B$ for the case of $b_8 > 0$ were averaged for each value of $b_8$.

For the first tree ($b_7=b_9=0$; solid circles in fig. 5B), we obtain an already familiar relationship in figure 5A (solid circles). In the second case ($b_7=0, b_9>0$; solid squares in fig. 5B), $\bar{P}_B$'s increase. In the third case ($b_7>0$ and $b_9>0$; solid triangles in fig. 5B), 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 5A (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. 6A with fig. 4A) (see also Zharkikh and Li, 1995). However, $P_C$ is distributed uniformly as in the case of the four sequence model tree (compare fig. 6B with fig. 4B). Therefore, it is the bootstrap test that is biased (conservative), and the interior-branch test is an unbiased test.

Figure 5B 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 5B (solid circle) corresponds to the six-sequence starlike model tree (fig. 1D[1]), but the lowest solid-square and the lowest solid-triangle points in figure 5B are obtained for the model trees that behave like a five-sequence star tree (see fig. 1D[2]) and a four-sequence star tree (see fig. 1D[3]), 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-

![Figure 5](https://example.com/fig5.png)

**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. 1A) in simulation for nucleotide-sequence data. A, Simulations for model trees with the expected exterior branch lengths $h_1 = h_2 = h_3 = h_4 = h_5 = h_6 = h_7 = h_8 = 0.07$ and $h_1 = h_9 = t$ (short branch trees, solid circles) and $b_1 = 0.31, b_3 = 0.3, b_4 = 0.29, b_5 = 0.3, b_6 = 0.28, b_3 = 0.3$, 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 $b_8 > 0$) 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: $b_7 = 0.0$ 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; $n = 100$. 

Downloaded from https://academic.oup.com/mbe/article-abstract/12/2/319/966358 by guest on 20 March 2019
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 (1992a, 1992b, 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., $d_{ij}$'s). As a result, the estimates of branch lengths, $b_i$'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. 2A, $b_3 = 0$), the conditional distribution of $b_5/s(b_5)$ for the case where the NJ tree agrees with topology $A$ in figure 2 is close to a gamma distribution (see fig. 7A) 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 $b_i$'s are excluded from consideration. We have found that the conditional distribution of $Z = b_i/s(b_i)$ varies considerably with the topology and expected branch lengths of the model tree (see fig. 7B 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 $Z_a$ be the value of $Z$ such that the probability of obtaining $Z$ greater than $Z_a$ is exactly $a$ (say, $a = 0.05$). Obviously, $Z_a$ varies with model (true) tree. We call the distribution with the highest $Z_a$ 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. 7B). 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. 7C). 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. 7C. Therefore, the four-sequence tree provides the worst-case distribution.) The distribution of $Z$ for the worst case is close to the gamma distribution

$$f(Z) = b^a T(a)^{-1} e^{-bZ} Z^{a-1},$$

(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

$$P_C' = \int_0^Z b^a T(a)^{-1} e^{-bZ} x^{a-1} dx.$$  

(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
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. 8A 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. 2A) with $b_1 = b_2 = b_3 = b_4 = 0.3$, and $b_5 = 0$ (see fig. 8A) and (2) a six-sequence model tree (fig. 1A) with $b_1 = 0.31$, $b_3 = 0.3$, $b_5 = 0.29$, $b_6 = 0.3$, $b_7 = 0.28$, $b_8 = 0.3$, and $b_9 = b_8 = b_9 = 0$ (see fig. 8B). 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 $P_C > 0.90$ 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 1992a, 1992b, 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
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

$$\bar{P}_B = \frac{B_{r+1} B_{s+1}}{B_m}, \text{ where}$$

$$B_i = \prod_{j=1}^{i-2} (2j-1), \text{ and } i \geq 3. \tag{27}$$

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]) or negative phylogenetic signal (if $\bar{P}_B = 0.33$ for $r = 2$ and $s = 3$), or positive phylogenetic signal (if $\bar{P}_B = 0.09$ for $r = 2$ and $s = 3$).

Clearly, the bias of the bootstrap estimate becomes more extreme as $\bar{P}_B$ decreases (see fig. 4A and fig. 6A). 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. 4B and fig. 6B).

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 1992a, 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 ($\tilde{D}_1, \tilde{D}_2, \tilde{D}_3, \tilde{D}_4, \tilde{D}_5$) (see equation [23]). In general, we can write $P_B = P_B(\tilde{\theta})$, where $\tilde{\theta}$ is a vector of estimates of all parameters involved. We can then consider $P_B$ as an estimate of $P_B[E(\tilde{\theta})]$, where $P_B[E(\tilde{\theta})]$ 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 $\Theta$. 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, $\bar{D}$ and $\sigma^2$ ($\bar{D} = \bar{D}_1 \simeq \bar{D}_2$, $\sigma^2 = \sigma_i^2 \simeq \sigma_0^2$, and $p = 1$). 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 for sequence clusters by this test. In practice, however $P'_C$ and $P_C$ are close to each other when a region of $P'_C$ > 0.9 or $P_C$ > 0.9 is considered.

Acknowledgments

We are grateful to Andrey Zharkikh for stimulating discussions. We also thank Susan Murphy, Bruce Lindsay, and Spencer Muse for valuable comments. This work was supported by grants from the National Institutes of Health and National Science Foundation to M.N.

APPENDIX A

Relationships between the Expected Branch Lengths of the Model Tree and the Values of $\sigma_1$, $\sigma_2$, and $\rho$

Since the values of $\sigma_1^2$, $\sigma_2^2$, and $\rho \sigma_1 \sigma_2$ (the variance and covariance of $\bar{D}_1 = S_B - S_4$ and $\bar{D}_2 = S_C - S_4$, respectively) determine the probability of recovering the correct tree (see equation [22]), it is important to understand their dependence on the branch lengths of the model tree.

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

$$\sigma_1^2 = [g(b_1, b_2) + g(b_3, b_4) + g(b_1, b_5, b_3)]$$

$$+ [g(b_2, b_5, b_3) + 4f(b_3)]/16,$$

$$\sigma_2^2 = [g(b_1, b_2) + g(b_3, b_4) + g(b_1, b_4, b_2)]$$

$$+ [g(b_2, b_5, b_3) + 4f(b_3)]/16,$$

$$\rho \sigma_1 \sigma_2 = [g(b_1, b_2) + g(b_3, b_4) + g(b_1, b_2)]$$

$$+ [g(b_2, b_5, b_3) + 4f(b_3)]/16,$$

where

$$g(h, h) = f(h_1 + h_2) - [f(h_1) + f(h_2)],$$

$$g(b_i, b_j, b_k) = f(b_i + b_j + b_k)$$

$$- [f(b_i) + f(b_j) + f(b_k)],$$

and function $f(\delta)$ is the same as that defined in equation (12). An analysis of equations (A1a–b) shows that the values of $\sigma_1$ and $\sigma_2$ tend to be close to each other, and $\rho$ approaches a value close to 1 as $b_5$ increases. This holds for a wide range of values of $b_1, b_2, b_3,$ and $b_4$.

Using equation (A1) we can directly compute the values of $\rho$ for a given expected value of $\bar{D} = b_5/s(b_5)$: $E(\bar{D}) = b_5/s$, for four-sequence model tree with "short" and "long" exterior branches (see fig. 3). Figure 9 shows that $\rho$ increases faster for trees with short exterior branches (solid circles) than for trees with long exterior branches (open circles). This explains the dif
Fig. 9.—Relationships between $\rho$ 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 $\rho$ and $Z$ increase as the length of interior branch ($b_3$) of the model tree in fig. 2A increases.

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

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

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

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

From the above equations, we obtain

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

and

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

To study the values of $\sigma_1/\sigma_2$ and $\rho$, let us rewrite the function $g(b_i, b_j)$ as

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

combining equations (A2) and (12). Considering the first three terms of Taylor’s series expansion and assuming that $b_i$ is much smaller than $3/4$, 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

$$g(b_i, b_j) = 2b_i b_j /n + O(b_i b_j), \quad (A7)$$

where $r + s = 3$. Next, we replace $b_i$ by $b + \delta_i$ ($i = 1, 2, 3, 4$), 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 $g(b_i, b_j), \sigma_1/\sigma_2$ and $\rho$.

$$g(b_i, b_j) = 2[b_i b_j + (\delta_1 + \delta_3 + \delta_2)] /n, \quad (A8)$$

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

and

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

Equation (A9) indicates that the ratio $\sigma_1/\sigma_2$ is very close to 1 whenever $\delta_1$ is close to $\delta_2$, or $\delta_3$ is close to $\delta_4$, or all $\delta_i$'s are much smaller than $b$. It follows from equation (A10) that $\rho = 1/2$ when all $\delta_i$'s are equal to 0 and that $\rho$ is close to $1/2$ whenever $\delta_i$'s are much smaller than $b$.

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

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

(Here we used the formula that was first derived by Sheppard [1899].) Note that for $\rho = 1/2$ we have $P = 1/8$. However, as $\rho$ deviates from $1/2$, the 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 $\rho$ tends to be 1 and $P$ tends to be $1/2$. 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, $\rho$ tends to be 0 and $P$ tends to be $1/4$.

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 \( \frac{1}{3} \). 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 \( i \)th interior branch and its variance with the expected values \( b_i \) and \( \sigma^2 \), 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 \( \hat{b}_i > 0 \):

\[
E(P_C | \hat{b}_i > 0) = E(P_C | Z > 0)
\]

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

\[
\text{Prob}(Z > 0) = \Phi(\xi),
\]

and

\[
\int_{0}^{\infty} P_C(x) \phi(x; \xi, 1) dx = \int_{0}^{\infty} P_C(x) \phi(x; \xi, 1) dx / \Phi(\xi),
\]

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

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

(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 \( Z = \hat{b}_i / s(\hat{b}_i) \) 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

\[
a = \frac{E(Z)}{V(Z)}, \quad \text{and} \quad b = \frac{E(Z)}{V(Z)},
\]

respectively, where \( E(Z) \) and \( V(Z) \) stand for the conditional mean and variance of \( Z \), respectively, when a data set that gives an NJ tree different from tree figure 2 are discarded.

Using the same notations as in equations (17) and (22), we can write \( Z \) as \( (D_1 + D_2)h \), where \( h = 1 / s(b_i) \). Assuming that \( D_1 \) and \( D_2 \) follow a bivariate normal distribution with mean vector \((0, 0)\) and a covariance matrix given in equation (20), we can compute the mean of \( Z \) by

\[
E(Z) = \frac{h}{2\pi\sigma_1\sigma_2P\sqrt{1 - \rho^2}} \times \int_{0}^{\infty} \int_{0}^{\infty} (x+y) \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{x^2}{\sigma_1^2} - 2\rho \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right] \right\} dx dy
\]

\[
= \frac{h(1+\rho)(\sigma_1+\sigma_2)}{2P\sqrt{2\pi}},
\]

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


TAKASHI GOJOBORI, reviewing editor

Received May 13, 1994

Accepted September 30, 1994