**Bootstrap Method of Interior-Branch Test for Phylogenetic Trees**

*Tatyana Sitnikova*

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

Statistical properties of the bootstrap test of interior branch lengths of phylogenetic trees have been studied and compared with those of the standard interior-branch test in computer simulations. Examination of the properties of the tests under the null hypothesis showed that both tests for an interior branch of a predetermined topology are quite reliable when the distribution of the branch length estimate approaches a normal distribution. Unlike the standard interior-branch test, the bootstrap test appears to retain this property even when the substitution rate varies among sites. In this case, the distribution of the branch length estimate deviates from a normal distribution, and the standard interior-branch test gives conservative confidence probability values. A simple correction method was developed for both interior-branch tests to be applied for testing the reliability of tree topologies estimated from sequence data. This correction for the standard interior-branch test appears to be as effective as that obtained in our previous study, though it is much simpler. The bootstrap and standard interior-branch tests for estimated topologies become conservative as the number of sequence groups in a star-like tree increases.

**Introduction**

There are two different types of statistical tests of the reliability of a branching pattern of phylogenetic trees. One is Felsenstein’s (1985) test, in which the stability of each sequence cluster (each partition of sequences) of a tree is evaluated by the bootstrap method (Efron 1979). The test is known to be sometimes excessively conservative (Zharkikh and Li 1992a, 1992b, 1995; Hillis and Bull 1993; Sitnikova, Rzhetsky, and Nc 1995). The other test, called the interior-branch test (Nei, Stephens, and Saitou 1985; Li 1989; Rzhetsky and Nei 1992), is to compute the standard errors of estimates of interior branch lengths and examine whether or not the branch lengths are significantly different from normal deviate test. This test has been shown to perform well if it is conducted properly (Sitnikova, Rzhetsky, and Nei 1995).

Rodrigo and Dopazo (1994) recently proposed that a bootstrap technique be used for testing the positiveness of the length of each interior branch in a given tree topology. This bootstrap test is different from Felsenstein’s. It is a nonparametric version of the interior-branch test, and we call it “the bootstrap interior-branch test.” However, the statistical properties of this test remain unclear. We have therefore studied some of the properties in comparison to those of the standard interior-branch test by using computer simulation. In this study, we examined both cases where all the assumptions of the test are satisfied and where some assumptions are violated. Before explaining details of the computer simulation, we briefly describe the bootstrap test studied here.

**Bootstrap Interior-Branch Test**

The null hypothesis of the bootstrap interior-branch test as applied by Rodrigo and Dopazo (1995) is the same as that of the standard interior-branch test, i.e., that the length of an interior branch is equal to 0. To test this null hypothesis, a bootstrap sampling of nucleotide sites from a set of sequences is conducted, and the interior branch length is estimated from each bootstrap sample. This procedure is repeated many times. Since the same topology is considered for all bootstrap samples, the estimates of the branch length can be either positive or negative. The proportion of the bootstrap samples producing positive estimates of the branch length is computed for each interior branch. This will be called the bootstrap confidence probability value.

Note that the test described above is appropriate for evaluating the reliability of a predetermined tree topology. In practice, however, an estimated topology is of special interest, and the test of this topology is slightly different from that of a predetermined topology. In the first case, a tree topology is chosen a priori and does not depend on the sequence data being analyzed, whereas in the second case a tree topology is estimated from the data. Since the statistical properties of both the bootstrap and the standard interior-branch tests are affected by the topology under study, we will consider the cases of predetermined and estimated topologies separately.

**Computer Simulation**

Both the bootstrap and the standard interior-branch tests provide the confidence probability (CP) value, which is interpreted as a complement of P value in standard statistical tests and therefore is expected to follow a uniform distribution under the null hypothesis (for the definition of P value see, e.g., Walpole and Myers 1989, p. 302). If the CP value of the test does not follow a uniform distribution, the test will give biased results. In practice, the good behavior of the tests is particularly important in the range of large (0.9–1.0) CP values, since a threshold value of 0.95 or 0.99 is commonly used to reject a null hypothesis. To study the distributions of CP values of the bootstrap and standard interior-branch tests under the null hypothesis, we conducted computer simulations.

**Methods of Computer Simulation**

We simulated the evolution of four or six nucleotide sequences of 100 nucleotides in length according to
the trees given in figure 1. We applied two methods to
test the significance of interior branch b3 in the case of
the four-sequence model tree (fig. 1A) and of interior
branch b8 for the six-sequence model tree (fig. 1B). The
procedure of computer simulations was essentially the
same as in our previous study (Sitnikova, Rzhetsky, and
Nei 1995), although in this paper several substitution
models were incorporated in generating nucleotide se-
dquences. The models were Jukes and Cantor’s (1969)
model (JC), Kimura’s (1980) two-parameter model
(K2P), and Jukes and Cantor’s (1969) model where the
rate of substitution was assumed to vary among nucle-
otide sites following a gamma distribution with param-
eter \( \alpha \) (K2P, see Golding 1983; Jin and Nei 1990). The
appropriate correction formula and calculated the ordinary
least squares (OLS) estimate of \( b_5 \) by the formula:
\[ \hat{b}_{5j} = k_j \hat{d}_j, \]
where \( \hat{b}_{5j} \) is the estimate of \( b_5 \) for the \( j \)th boot-
strap sample, \( \hat{d}_j \) is a vector of distances between se-
dquences of the \( j \)th bootstrap sample, and \( k_j \) is a vector of
the coefficients. Both \( K_j \) for \( b_5 \) in the four-sequence
model tree and \( K_8 \) for \( b_8 \) in the six-sequence model tree
were calculated as described in Rzhetsky and Nei (1992)
and given in Sitnikova, Rzhetsky, and Nei (1995). (3)
For all bootstrap samples we computed the proportion
\( P_B \) of the bootstrap replications for which the estimate
of the branch length is positive.

The CP value of the standard interior-branch test
\( P_C \) was computed as follows. (1) We estimated dis-
tances between sequences \( d \) and obtained the OLS es-
timate \( \hat{b}_5 \) of \( b_5 \) and its standard deviation \( s(\hat{b}_5) \) as
described in Rzhetsky and Nei (1992). (2) We then cal-
culated \( Z = \hat{b}_5/s(\hat{b}_5) \) and, assuming that the test statistic
\( Z \) follows the standard normal distribution, computed \( P_C \)
in the following way:
\[
P_C = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-x^2/2} \, dx.
\]

Statistical Tests for an Estimated Topology

In this simulation we tested the reliability of the
branching pattern of the tree when it was estimated from
the data by the neighbor-joining (NJ) method (Saitou
and Nei 1987). Thus, we considered only sets of se-
dquences that produced NJ trees whose topologies were
identical with that of the model tree (about one third of
all data sets in the case of the four-sequence model tree),
and discarded all other sets of sequences. In the previous
study (Sitnikova, Rzhetsky, and Nei 1995), we have de-
developed a corrected version of the standard interior-
branch test to be applied for assessing a tree estimated
from sequence data. In this corrected test, we first com-
cputed \( Z = \hat{b}_5/s(\hat{b}_5) \) as described above and then cal-
culated the corrected CP value \( P_C \) by the formula:
\[
P_C = \int_0^Z b^a \Gamma(a)^{-1} e^{-bx} x^{a-1} \, dx,
\]
where \( a = 3.17 \), and \( b = 3.06 \).

Results

Bootstrap and Interior-Branch Tests for a
Predetermined Topology

In this section the performance of the bootstrap and
standard interior-branch tests is studied by computer
simulation using the four-sequence model tree (fig. 1A).
We start with a computer simulation where the JC sub-
stitution model is applied for both generating sequences
and estimating distances. Then, the K2P and JC mod-
els are considered in the same way.

First, let us examine whether or not the assump-
tions of the tests under study are satisfied. The bootstrap
method does not make any explicit assumptions about
the shape of distribution of a test statistic (Efron and
Tibshirani 1993, p. 55). However, there is an implicit
FIG. 2.—Distributions of the estimate of interior branch length $b_2$ of the model tree in simulations according to the four-sequence model tree (fig. 1A) with $b_1 = b_2 = b_3 = b_4 = 0.3$ and $b_5 = 0$, where the JC (A), K2P with $R = 15$ (B), and JCT with $a = 0.1$ (C) substitution models were used in generating sequences, and appropriate estimators of distances were used. The distributions of $b_2$ in bootstrap samples ($\hat{b}_2$) were obtained for one data set with $P_B = 0.76$ in simulation under the JCT model with $a = 0.1$ (D). In all cases the solid line represents the normal distribution with the same mean and standard deviation as those of the distribution of $b_2$. The number of replications of simulations ($r$) was 10,000, the number of bootstrap replications ($B$) was 5,000, and the length of nucleotide sequences ($n$) was equal to 100.

Let us now examine the distributions of $P_B$ and $P_C$ obtained in computer simulations. As mentioned earlier, these distributions are expected to be close to a uniform distribution under the null hypothesis. This is indeed the case for both $P_B$ and $P_C$ (fig. 3A), although some deviation is observed in the range of very small and very large CP values. In the case of the bootstrap test, this deviation probably occurs because the $\hat{b}_2$ distribution obtained for a set of data is not necessarily a good estimate of the distribution of $b_2$ when the interior branch length estimate computed from this data set falls in the extreme tail of the distribution of $b_2$ (see a similar example in Efron and Tibshirani 1993, p. 81). A minor deviation from uniformity of the distribution of $P_C$ is likely to be a consequence of some departure of the distribution of $\hat{b}_2$ from the corresponding normal distribution. Note, however, that for the threshold values above 0.95, the bootstrap interior-branch test is less conservative than the standard interior-branch test.

The JC substitution model is not always appropriate for describing the mechanisms of molecular evolution, since in real genes transitional nucleotide substitutions tend to occur more frequently than transversional substitutions (e.g., see Nei 1987, p. 84). In addition, substitution rates may vary among the sites of sequences. In order to examine the behavior of the tests under more realistic substitution models, we conducted simulations under the K2P substitution model with a transition/trans-
version ratio ($R$) of 15 and the JCT model with a gamma parameter ($\alpha$) equal to 0.1. Note that these are rather extreme values of the parameters, since such a high $R$ value is rarely observed and $\alpha = 0.1$ is unusually small (e.g., Kocher and Wilson 1991; Yang, Goldman, and Friday 1994; S. Kumar, personal communication). Thus, we studied the performance of the tests in extreme conditions to verify their behavior in common phylogenetic analysis.

The simulation results show that for both the K2P and JCT models the assumption of a normal distribution of $b_5$ no longer applies (fig. 2B and C). Namely, in both cases the distributions of $b_5$ are narrower and have longer tails than the corresponding normal distribution. Note that the deviation from normality is much more conspicuous in the case of the JCT model than in the case of the K2P model. The distribution of bootstrap estimates ($b_5^*$) of $b_5$ for a set of sequences obtained under the JCT model often becomes more skewed than the distribution of $b_5$, though it is also narrower and has longer tails than the normal distribution (compare the shapes of the frequency distributions in fig. 2C and D).

Despite the deviation of the distribution of $b_5$ from normality, the distributions of both $P_B$ and $P_C$ appear to be close to uniform in the case of the K2P model with $R = 15$ (fig. 3B). Moreover, the bootstrap interior-branch test remains nearly unbiased in the case of the JCT model ($\alpha = 0.1$) as well, but the standard interior-branch test becomes conservative for large values ($P_C > 0.9$) of $P_C$ (fig. 3C). The distribution of $P_B$ in this case is no longer uniform but rather bimodal, whereas the $P_B$ distribution is still close to uniform.

When less extreme values of the model parameters $R$ and $\alpha$ are used (smaller values of $R$ and larger values of $\alpha$), the distribution of the branch length estimate converges to normal. Consequently, both $P_B$ and $P_C$ approach a uniform distribution (see fig. 3D for the case of the JCT model with $\alpha = 2$).

Therefore, we can conclude that the bootstrap interior-branch test is an approximate method that performs quite well under various substitution models, despite possible violation of the implicit assumption of the test.

Our study of the effects of the sequence length and the number of sequences on the performance of the interior-branch tests showed that the distribution of interior branch length estimates approaches the normal distribution as the number of nucleotides or the number of the sequences increases. This improves the performance of both tests, making their distributions closer to uniform, even in the case of the JCT model with $\alpha = 0.1$ (data not shown). By contrast, use of very short sequences (say, 10 nt long) is followed by the considerable deviation of the distribution of $b_5$ from normality. This makes both tests conservative in the range of high CP values, even in the simulation under the JC model, the bootstrap interior-branch test being less sensitive to this condition. Therefore, for applying both of the interior-branch tests, sequences at least 50 nt long should be used.

\[ \text{Interior-Branch Tests for an Estimated Topology} \]

In phylogenetic analysis of real data an investigator is often interested in testing a tree topology estimated from data rather than a predetermined topology. In this case, both bootstrap and standard interior-branch tests as described above become too liberal (data not shown). However, it is possible to develop a simple correction of the tests to overcome this problem. In analogy to deriving the formula for $P_C$ (see eq. 2), we consider the four-sequence model tree (fig. 1A) with interior branches of equal length and with interior branch equal to 0 as the null-hypothesis tree. The distribution of $Z = b_5 - b_5^* = s(b_5)$ for this tree has been found to represent the "worst case" in the sense that the tests constructed for this case will never be too liberal for any other case, even for the trees with a larger number of sequences (Sitnikova, Rzhetsky, and Nei 1995).

Let us first consider the distribution of the interior branch length estimate ($b_5$) for a predetermined topology, or the unconditional distribution (fig. 4, solid columns). The NJ-method would select a given topology among the three possible topologies for four sequences in approximately one third of the cases. Obviously, the conditional distribution of $b_5$ (fig. 4, open columns) for a given topology chosen by the NJ-method is a fraction of the unconditional distribution. Since the shape of the conditional distribution is different from the shape of the normal distribution, we aim to find a simple approximation of this distribution.

In developing a test for an interior branch of an estimated topology, Tajima (1992) adapted the truncated standard normal distribution to approximate the conditional distribution of $t = b_5/s(b_5)$, given that the considered topology is selected by the NJ-method. (Here $t$ is a statistic analogous to $Z$, though Tajima's [1992] method is used for calculating variances of branch length estimate.) Thus, in his test the probability of accepting a particular topology with critical value $c$ ($P_T(c)$) has been computed as follows:

\[ P_T(c) = \int_{t=c}^{\infty} f(t) \, dt, \]  

\[ \text{Fig. 4.—Unconditional (solid columns) and conditional (open columns) distributions of the estimated length of the interior branch $b_5$ for the four-sequence model tree (fig. 1A). The conditional distribution is a part of the unconditional distribution for the cases where the topology of the model tree is selected by the NJ method. The simulations were conducted according to the model tree with $b_1 = b_2 = b_3 = b_4 = 0.3$ and $b_5 = 0$, where the JC model was used for generating sequences and estimating distances between them. $r = 10,000$ and $n = 100$.} \]
\[
\text{where } f(t) = \begin{cases} 
\frac{\exp(-t^2/2)}{\sqrt{2\pi}} & \text{if } t \geq 0.4307, \\
0 & \text{if } t < 0.4307.
\end{cases}
\]

Here \( f(t) \) is a truncated standard normal distribution and \( P_T(c) = 1/3 \) when \( c < 0.4307 \). \( P_T(c) = 1/3 \) approximately corresponds to the probability of obtaining a tree in the case of a star-like model tree of four sequences.

To find the correction for the interior-branch tests, we make use of Tajima’s approach to describe the conditional distribution of a test statistic. The validity of such approximation can be justified in the following way. Let \( b_s \) be some value of \( b \), greater than 0. We note that the tail of the conditional distribution of \( b_s \) (fig. 4, open columns), where \( b_s > b_n \), nearly coincides with the tail of the unconditional distribution of \( b_s \) (fig. 4, solid columns). Let us fix \( b_s \) and let \( \alpha \) denote the probability of obtaining \( b_s \) higher than \( b_n \) in the unconditional distribution. Since the NJ-method recovers a given topology in one third of the cases, the conditional distribution (given that the topology is selected by the NJ-method) constitutes one third of the unconditional distribution. After normalizing the conditional distribution to make the area under the curve of the distribution equal to 1, the probability of obtaining \( b_s \) greater than \( b_n \) is about \( 3\alpha \) in the conditional distribution. Then, from \( \alpha = 1 - P \), where \( P \) is the probability of obtaining \( b_s \) smaller than \( b_n \) in the unconditional distribution, and \( 3\alpha = 1 - \tilde{P} \), where \( \tilde{P} \) is the same probability for the conditional distribution, we obtain \( \tilde{P} = 3P - 2 \). Therefore, we can introduce the following formulas to compute \( \tilde{P} \) for an observed \( b_s \) value in the test for an estimated topology:

\[
\tilde{P}_B = \begin{cases} 
(3P_B - 2) & \text{if } P_B > \frac{2}{3} \\
0 & \text{if } P_B \leq \frac{2}{3}
\end{cases}
\]

and

\[
\tilde{P}_C = \begin{cases} 
(3P_C - 2) & \text{if } P_C > \frac{2}{3} \\
0 & \text{if } P_C \leq \frac{2}{3}
\end{cases}
\]

Note that both \( \tilde{P}_C \) and Tajima’s (1992) tests make an assumption that the interior branch length estimate follows a normal distribution and, in fact, \( \tilde{P}_C = 1 - 3P_T(c) \), where \( P_T(c) \) is calculated for Tajima’s test by equation (3).

\( P_C \) can also be obtained for the two-sided interior-branch test (\( P_{C2} \)). In this case, the alternative hypothesis of the test is that the interior branch length deviates from 0. Thus, the corrected CP value of the test becomes:

\[
\tilde{P}_C = \begin{cases} 
(3P_{C2} - 1)/2 & \text{if } P_{C2} > \frac{1}{3} \\
0 & \text{if } P_{C2} \leq \frac{1}{3}
\end{cases}
\]

We examined the distributions of \( \tilde{P}_B, \tilde{P}_C \) (eq. 5), and \( P_C \) (eq. 2) in the simulations according to the four-sequence tree (fig. 1A). All tests tended to perform well for \( P_B \leq 2/3 \) and \( P_C \leq 2/3 \) are lumped in 0 for the corrected CP values. Consequently, in the region 0--0.5 both distributions diverge away from the uniform distribution. Nevertheless, these deviations can be neglected, since in practice small CP values are not used for rejecting the null hypothesis of the tests and therefore rarely draw the interest of investigators. Simulations for the six-sequence model tree (fig. 1B) with expected lengths of all interior branches equal to 0 indicate that the above three tests all become conservative as the number of sequence groups in a star tree increases (fig. 5A).
In the case of the JCT model with \( a = 0.1 \), the distributions of \( P_B \) and \( P_C \) give more serious underestimates of the true CP values in the range of large values than does the \( P_B \) distribution (fig. 5C). This is consistent with the results of the computer simulation with the JCT model for a predetermined topology, where the interior-branch test is found to be conservative for large values of \( P_C \) (see above).

Interior-Branch Tests when Model Assumptions are Violated

It is unlikely that the substitution pattern of real nucleotide sequences exactly follows any substitution model that is usually used in phylogenetic analysis. Hence, the properties of the statistical tests for the cases of biased estimators of distances between sequences are worth studying. Here we analyze the behavior of both bootstrap and standard interior branch tests when the K2P and JCT substitution models are incorporated in generating sequences, while the JC formula is used for estimation of distances.

In our simulations both interior-branch tests were affected by an incorrect choice of the distance measures to approximately equal extents. No substantial deviation of the distributions of \( P_B \) and \( P_C \) from the uniform distribution occurs when the expected value of the branch length estimate does not deviate from the true value of the branch length. This was found in the case of the four-sequence model tree with equal lengths of exterior branches (fig. 6A) where the expected value of \( b_3 \) is 0. In the case of overestimation of the interior branch length, both \( P_B \) and \( P_C \) tests become too liberal (fig. 6B), whereas they are conservative when the interior branch length is underestimated (fig. 6C).

In practice, estimates of branch lengths may be biased. Therefore, it is preferable to use the most realistic distance measures available for conducting statistical tests of the reliability of phylogenetic trees.

Discussion

Simplicity of implementation and freedom from distributional assumptions are among the most attractive properties of the bootstrap technique. One of the popular applications of the bootstrap method in statistical tests is to obtain \( P \) values. However, the bootstrap test has been shown to perform well only for normally distributed statistics or statistics that follow a symmetrical and continuous distribution. That is, it has been proven that the distribution of the bootstrap \( P \) value is asymptotically uniform, provided that the bootstrap method gives a good approximation of the real distribution (Singh and Berk 1994).

In this paper, we have shown that the bootstrap method performs well when it is employed for testing the positiveness of the length of an interior branch of a predetermined tree topology. It appears to be nearly unbiased even in the case of the simulation under the JCT substitution model with \( a = 0.1 \), where the distribution of the branch length estimate differs considerably from a normal distribution. The standard interior-branch test (the parametric analog of the bootstrap interior-branch test that incorporates an assumption of a normal distribution of the test statistic) turns out to be conservative for high values of \( P_C \) in this case.

When a topology estimated from a data set is tested, the corrected CP values for the bootstrap (\( \bar{P}_B \)) and standard interior-branch tests (\( P_C \) or \( P_C' \)) rather than \( P_B \) and \( P_C \) are recommended for use. Since \( \bar{P}_B \) and \( P_C' \) as presented in this paper are very simple to compute and as effective as \( P_C' \), they might be preferable to \( P_C' \).

The bootstrap interior-branch test can serve as a good alternative to the standard interior-branch test in evaluating the reliability of both predetermined and estimated tree topologies. Note also that, unlike the standard interior-branch test, the bootstrap test does not require computation of a covariance matrix of distance
estimates, which is rather computationally intensive when a large data set is considered.

Nevertheless, the interpretation of the results of both interior-branch tests in assessing the accuracy of a tree topology is not straightforward. This is because the test of the positiveness of an interior branch length is not equivalent to the test of the hypothesis that the partition of sequences produced by this interior branch is present in the true tree. In particular, the interior branch generating an incorrect partition in a tree may have positive expectation of the length estimate (see examples in Sitnikova, Rzhetsky, and Nei 1995). (Note that this never happens if the clustering of four sequences or four monophyletic groups is considered.) After examining a number of examples of such situations, we conjectured that a wrong tree must contain at least one interior branch of which the expectation of the length becomes negative. If this conjecture is correct, we can reject a tree under consideration when it contains an interior branch of which the length is negative and statistically significant. Conversely, the branching pattern of a topology can be recognized as reliable only if all interior branches of the tree are significantly positive.

Acknowledgments

This study was motivated by discussions with Masatoshi Nei. I would like to thank Andrey Rzhetsky and Masatoshi Nei for numerous comments and suggestions and Joaquin Dopazo for discussions on the early drafts of the manuscript. I also thank Claudia Russo, Jianzhi Zhang, and Jennifer M. Hay for comments. This work was supported by grants from NIH and NSF to M. Nei.

LITERATURE CITED


SHOZO YOKOYAMA, reviewing editor

Accepted January 8, 1996