Syst. Biol. 51(3):492–508, 2002
DOI: 10.1080/10635150290069913

An Approximately Unbiased Test of Phylogenetic Tree Selection

HIDETOSHI SHIMODAIRA

Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minatoku, Tokyo 106–8569, Japan;
E-mail: shimo@ism.ac.jp

Abstract.— An approximately unbiased (AU) test that uses a newly devised multiscale bootstrap technique was developed for general hypothesis testing of regions in an attempt to reduce test bias. It was applied to maximum-likelihood tree selection for obtaining the confidence set of trees. The AU test is based on the theory of Efron et al. (Proc. Natl. Acad. Sci. USA 93:13429–13434; 1996), but the new method provides higher-order accuracy yet simpler implementation. The AU test, like the Shimodaira–Hasegawa (SH) test, adjusts the selection bias overlooked in the standard use of the bootstrap probability and Kishino–Hasegawa tests. The selection bias comes from comparing many trees at the same time and often leads to overconfidence in the wrong trees. The SH test, though safe to use, may exhibit another type of bias such that it appears conservative. Here I show that the AU test is less biased than other methods in typical cases of tree selection. These points are illustrated in a simulation study as well as in the analysis of mammalian mitochondrial protein sequences. The theoretical argument provides a simple formula that covers the bootstrap probability test, the Kishino–Hasegawa test, the AU test, and the Zharkikh–Li test. A practical suggestion is provided as to which test should be used under particular circumstances. [Approximately unbiased test; confidence limit; Kishino–Hasegawa test; maximum likelihood; multiscale bootstrap; phylogenetics; selection bias; Shimodaira–Hasegawa test.]

Tree selection is a common practice in phylogenetics and is used to find an optimal tree from taxonomic molecular sequences. One of the widely used selection criteria is the maximum likelihood (ML) method (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981), which calculates a likelihood value for each of the candidate trees and then selects the tree with the largest likelihood value. If we have imaginary sequences of infinite length for a finite number of taxa, the optimal tree will represent the true history of evolution unless the assumed model of evolution, that is the substitution process, is extremely misspecified. In practice, however, the sequence length is finite. When the sequence length is not long enough, sampling error causes tree selection to fluctuate, and the optimal tree may not reflect the true tree. In other words, the optimal tree may have been designated optimal by chance.

Several procedures have been developed and used for assessing the confidence of tree selection. The bootstrap probability (BP; Felsenstein, 1985) and the Kishino–Hasegawa tests (KH; Linhart, 1988; Kishino and Hasegawa, 1989; Vuong, 1989) have been used widely. These methods produce for each tree a number ranging from zero to one. This number is the probability value or *P*-value, which represents the possibility that the tree is the true tree. The greater the *P*-value, the greater the probability that the tree is the true tree. Relative certainty, or uncertainty, in tree selection can also be represented as the confidence set—the set of trees that are not rejected by the tests. One expects the true tree will be included in the confidence set.

Although the BP test is very useful in practice, it is biased, as discussed in Hillis and Bull (1993), Felsenstein and Kishino (1993), Zharkikh and Li (1992), Efron et al. (1996), and Sanderson and Wojciechowski (2000). As mentioned in Shimodaira and Hasegawa (1999) and Goldman et al. (2000), selection bias is also apparent in the KH test because, typically, many trees have been compared when the choice of the ML tree is made. In other words, the likelihood value for the ML tree is biased upward because the maximum of likelihood values over all the trees can easily have a very large value by chance. The selection bias often leads to overconfidence in the wrong trees. This may result in conflicting conclusions, each claiming statistical significance; only one, however, can be true.

The selection bias of the KH test is automatically adjusted by a multiple comparisons test known as the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa, 1999; Goldman et al., 2000) when applied to tree selection. As mentioned in Remark 4 of Shimodaira and Hasegawa (1999), the selection bias is also adjusted by the weighted Shimodaira–Hasegawa (WSH) test (Shimodaira, 1993, 1998; Shimodaira and Hasegawa, 1999; Buckley et al., 2001), in which the test statistics of the SH test are standardized. There is a problem, however, in the multiple comparisons tests. Strimmer and Rambaut (2001) pointed out that the SH test may be subject to another type of bias such that the number of trees included in the confidence set tends to be very large as the number of trees to be compared increases. This conservative behavior of the SH test is alleviated, although not completely, by weighting in the WSH test.

Here, I propose a new method for reducing the bias of the BP test. The approximately unbiased (AU) test was developed for general hypothesis testing of regions; its accuracy has been confirmed (Shimodaira, 2000a) for very simple cases in which an exactly unbiased test can be obtained as a reference. The AU test provides yet another procedure for assessing the confidence of tree selection. The AU test adjusts the selection bias ignored in the BP and KH tests and is less conservative than the SH test.

The theory behind the AU test is an extension of the geometric theory of Efron et al. (1996), providing higher-order accuracy with simpler implementation. In the newly devised multiscale bootstrap procedure, several sets of bootstrap replicates are generated by changing the sequence length, which may differ from that of the original data. The number of times the hypothesis is supported by the replicates is counted for each set to obtain BP values for different sequence lengths. The AU test calculates the approximately unbiased P-value from the change in the BP values along the changing sequence length. This remarkably simple implementation of the AU test is an idea very similar to the complete-and-partial bootstrap technique of Zharkikh and Li (1995; ZL).

The test procedures mentioned above are justified in their own right. However, it is of interest to compare certain aspects of the procedures. I will discuss two aspects associated with coverage probability, namely, type-1 error and unbiasedness. A test controls a type-1 error rate if the probability of false rejection under the null hypothesis is not greater than the significance level α, say, 0.05. A test is unbiased if the probability of correct rejection under alternative hypotheses is not less than α and if it controls the rate of type-1 error. An unbiased test, then, controls the probability of a type-1 error, whereas a test may be biased even if it controls for type-1 error.

The KH test does not control for type-1 error, nor is it unbiased, given its selection bias. The SH test is excellent in terms of type-1 error, but it is heavily biased. This explains why the SH test appears conservative, especially for comparisons of many trees. Plainly, the SH test is derived under a pessimistic assumption. The AU test, on the other hand, is derived under a rather moderate assumption. In most cases it works better than the SH test but in special cases may violate type-1 error. These points will be illustrated in a simulation study as well as in the analysis of the mammalian mitochondrial (mt) protein sequences of Shimodaira and Hasegawa (1999).

Other methods such as the parametric bootstrap test, the Bayesian posterior probability test, and the likelihood ratio test against the full model, are not discussed here. These methods may have difficulty in tree selection, being sensitive to the misspecification of the probabilistic model of evolution. The true tree is often rejected at an extremely small significance level. See Shimodaira (2001) for a brief discussion, and Goldman et al. (2000) for the difference between the nonparametric and parametric versions of the bootstrap method.

Monotonicity is another desirable property of tests discussed in literature (see Appendix remark 1). This is a simple logical requirement for selection. The SH and WSH tests are monotone, but the AU test is not. Monotonicity will not be sought, however, because monotonicity and unbiasedness are not compatible. The AU test, like the test of Shimodaira (2000b), is an attempt to obtain less biased tests by relaxing the monotonicity.

To discuss the confidence of tree selection in terms of statistical testing, one must clarify the null hypotheses being tested. I do so in the next section, and then briefly review bootstrap resampling and coverage probability. On the basis of the notations and terminology given in these review sections, the AU test is described. Specific steps are provided in the section on the multiscale bootstrap. Not only tree selection but also edge selection for testing the monophyly of a specified group of taxa will be discussed. All technical details are provided in the Appendix and referenced in the text.

For calculating the P-values, the software CONSEL was developed by Shimodaira and Hasegawa (2001); it is available from the author. The software currently works in conjunction with the phylogenetic software packages Molphy (Adachi and Hasegawa, 1996a), PAML (Yang, 1997), and PAUP* (Swofford, 1998).

Methods

Null Hypotheses

In tree selection, one tries to identify the best tree from a set of competing trees, for example, tree-1, tree-2, ..., tree-M, where M is the number of the trees compared. Each tree represents a hypothetical branching order of species. M may be the number of all possible combinations of tree form of the taxa. Let us try to reduce M in advance of the analysis, if some of the trees can be eliminated by prior knowledge. For each tree, we are given a score calculated from the molecular sequences, and we select the tree with the highest score. Let Y_i be the log-likelihood of tree-i for $i = 1, \ldots, M$, and take that as the score of the tree. The log-likelihood may be expressed as

$$Y_i = \sum_{n=1}^{N} X_{i,n}, \tag{1}$$

where N is the length of the sequences and $X_{i,n}$ is the site-wise log-likelihood of tree-i at site-n. Note that calculation of $X_{i,n}$, $i = 1, \ldots, M$, $n = 1, \ldots, N$ involves numerical optimization of the tree parameters, for example, the edge lengths. The ML estimate (MLE) of these parameters is obtained by maximizing Y_i for each tree. The values of $X_{i,n}$ are produced by the phylogenetic software programs mentioned earlier.

Y_i is recognizable as a random variable because it is calculated from other random variables, namely, the molecular sequences. The observed data set of the molecular sequences is an instance of the random matrix generated by the stochastic process on the tree, which is a probabilistic model of the substitution process of molecules along the genealogy (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981). Assuming independence of the sites, the generation of Y_i is mathematically equivalent to the following simple sampling model. Although the sampling model appears to be biologically inaccurate, the mathematical equivalence justifies our use of it. In addition, its simplicity helps us to understand the statistical argument.

Consider imaginary sequences of infinite length, and take a random sample of length N'. The log-likelihood for tree-i is then expressed as

$$Y_i^* = \frac{N}{N'} \sum_{j=1}^{N'} X_{i,n_j}, \tag{2}$$

where n_j denotes a randomly selected site from the infinite length sequences, and factor N/N' is included to make Eq. 2 comparable to Y_i. Y_i^* is recognized as a random variable because its value depends on the sampling of the sites. The observed Y_i is an instance of Y_i^* in which $N' = N$. Note that the tree parameters should be reoptimized for each set of the sampled sites, but here values of $X_{i,n}$ will be treated as fixed values. This approximation is explained later.

As N' approaches infinity, Eq. 2 converges to a value denoted μ_i. In fact, μ_i is the expected value of Y_i for $i = 1, \ldots, M$. Y_i is distributed around μ_i, and thus the random vector $Y = (Y_1, \ldots, Y_M)$ is jointly distributed around the parameter vector $\mu = (\mu_1, \ldots, \mu_M)$. The tree of interest is the tree with the largest μ_i value, because it is the true tree under certain conditions (see Appendix remark 2).

The null hypotheses being tested are now described in terms of regions of the parameter space. The hypothesis that tree-i has the largest μ_i value is

$$H_i : \mu_i \geq \mu_j, \quad j = 1, \ldots, M. \tag{3}$$

This is regarded as a region in the M-dimensional parameter space, and μ is said to be included in region H_i, that is, $\mu \in H_i$, when μ_i is the largest among $\mu_1, \ldots, \mu_M$. Because the parameter space can be identified with the sampling space for Y, $Y \in H_i$ when tree-i has the largest Y_i value among $Y_1, \ldots, Y_M$. The parameter space is divided into M regions $H_1, \ldots, H_M$ corresponding to M trees, and these regions are facing each other on the boundaries where $\mu_i = \mu_j$ for at least some $i \neq j$. Which region the observed log-likelihood vector Y falls in can be determined, but ascertaining which region the parameter vector μ belongs to requires more information, as detailed in the next section.

Bootstrap Resampling

Considering that Y is distributed around μ, one might believe that the hypothesis $\mu \in H_i$ is probable when the event $Y \in H_i$ is observed. There is, however, the possibility that $Y \in H_i$ by chance, even though $\mu \in H_j$ for some $j \neq i$. In other words, the selected tree with the largest Y_i value is not necessarily the tree with the largest μ_i value. The confidence of the selection would be assessed by determining how much the selection fluctuates if Y^* of Eq. 2 is sampled from the imaginary infinite-length sequences. The frequencies with which Y^* falls in the regions $H_1, \ldots, H_M$ of M trees indicate how probable these trees are.

Because infinite-length sequences are not available in practice, let us replace them with observed sequences of length N. This is what the bootstrap resampling does. It is often the case that $N' = N$ for the bootstrap resampling, but we reserve the generality of using any value for N'. A bootstrap replicate Y_i^* of Y_i is given by Eq. 2, but the sites $n_1, \ldots, n_{N'}$ are randomly sampled from $1, \ldots, N$ with replacement. A bootstrap replicate Y^* of the vector Y is then $Y^* = (Y_1^*, \ldots, Y_M^*)$, with the same $n_1, \ldots, n_{N'}$ being used in all of the Y_i^* values so that the correlation structure among the elements of Y^* reflects that of Y. This random resampling is repeated B times, and the bootstrap replicates $Y^{*1}, \ldots, Y^{*B}$ of Y are obtained. B should be large enough, for example, 10,000, that the frequencies are calculated with marginally small sampling errors. The BP of tree-i, denoted BP_i, is the frequency obtained by counting how many times the event $Y^{*b} \in H_i$ is observed for $b = 1, \ldots, B$. The counts are divided by B so that $\mathrm{BP}_1 + \cdots + \mathrm{BP}_M = 1$. Tree-$i$ is regarded as probable when BP_i is large enough.

According to the description of the nonparametric bootstrap of Efron (1979) and Felsenstein (1985), the molecular sequences of the sites $n_1, \ldots, n_{N'}$ are to be sampled to constitute a replicate of the data set, and the MLE will be calculated for the replicate. In the bootstrap procedure described above, however, we sampled the site-wise log-likelihoods corresponding to these sites by treating them as fixed values. This approximation is the resampling of estimated log-likelihoods (RELL) method of Kishino et al. (1990), which avoids time-consuming recalculation of the MLE of the tree parameters for a large number of replicates. The approximation improves as N and N' become larger, and further improvement is possible by taking into account a higher-order term, as described in Lemma 1 of Shimodaira (2001). However, the RELL method is often accurate enough for phylogenetic analysis, as indicated in the simulation of Hasegawa and Kishino (1994), because the sequence length currently is very large.

Given that Y^* is expressed as the sum of independently sampled N' components in Eq. 2, it follows from the central limit theorem that the distribution of Y^* approaches the multivariate normal as N' approaches infinity. This justifies the normal approximation of the distribution of $Y_i^* - Y_j^*$ used commonly in the KH test.

Although independence of the sites is assumed throughout in this paper, modifying the method to take into account the short-range correlations among the sites is not very difficult. Block resampling for time series is useful for this case (see, for example, the bibliographic notes of Davison and Hinkley, 1997:427).

Coverage Probability

The P-value of testing H_i, denoted P_i, is calculated from the site-wise log-likelihoods $X_{i,n}$, $i = 1, \ldots, M$, $n = 1, \ldots, N$. Several definitions of P-values other than BP_i are available in the literature. Although those may be justified under some circumstances, I argue for the desired properties of P-values below.

In the practice of statistical testing, H_i is rejected when $P_i < \alpha$ for a prespecified level of significance $0 < \alpha < 1$, say, 0.05. In other words, the possibility is rejected that, among the candidate trees, tree-i has the largest μ_i value. The rejection probability of tree-i is

$$\beta_i(\mu) = \Pr\{P_i < \alpha; \mu\}, \tag{4}$$

where μ in Eq. 4 indicates that the probability depends on μ. $\beta_i(\mu)$ is sometimes called the power function.

The confidence set of trees is obtained by collecting trees with $P_i \geq \alpha$. When $\mu \in$

H_i, it is desirable for P_i to be $\geq \alpha$ so that tree-i is included in the confidence set. This probability

$$1 - \beta_i(\mu), \quad \mu \in H_i \tag{5}$$

is called the coverage probability of the confidence set. The coverage probability should not be smaller than the confidence coefficient $1 - \alpha$. In other words, the probability of false rejection should satisfy

$$\beta_i(\mu) \leq \alpha, \quad \mu \in H_i. \tag{6}$$

The test controls for type-1 error when inequality 6 holds.

The test can be very conservative, even if it controls for the type-1 error, when the rejection probability is small for the false hypotheses. When μ is not included in H_i, that is, $\mu \notin H_i$, it is desirable for $\beta_i(\mu)$ to be as large as possible. The test of H_i is said to be unbiased when

$$\beta_i(\mu) \geq \alpha, \quad \mu \notin H_i \tag{7}$$

and when inequality 6 holds at the same time. Often $\beta_i(\mu)$ changes continuously as μ moves, and thus the unbiasedness implies "similarity" (Lehmann, 1986) on the boundary

$$\beta_i(\mu) = \alpha, \quad \mu \in \partial H_i, \tag{8}$$

where ∂H_i denotes the boundary of H_i. ∂H_i is the hypersurface of H_i facing the outside of H_i (Fig. 1). If this is the case, P_i is a random variable distributed uniformly on [0, 1] when $\mu \in \partial H_i$. The similarity is checked rather easily and is used as a substitute for unbiasedness. Unbiasedness is one of the desired properties, though not a mandatory one, of P-values.

The SH test and the WSH test satisfy inequality 6 as explained in Shimodaira and Hasegawa (1999). However, the equality possible in inequality 6, that is, $\beta_i(\mu) = \alpha$, occurs only at the least favorable configuration where $\mu_1 = \cdots = \mu_M$, the inequality holding strictly for the other $\mu \in H_i$. The power can be much smaller than α for general values of $\mu \notin H_i$, and Eq. 8 does not hold at all. This effect is multiplied as M increases, which makes the SH test and the WSH test look conservative. The AU test described next takes this problem into consideration.

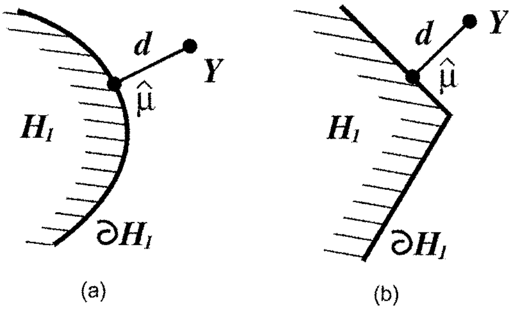

FIGURE 1. Region H_1 with boundary ∂H_1. Y is the data point, $\hat{\mu}$ is the projection, and d is the signed distance. The asymptotic theory behind the AU test assumes (a) a smooth boundary, which approximates (b) a nonsmooth boundary. In fact, the boundary is not smooth for the selection problem, where region H_1 forms a polyhedral convex cone in M-dimensional space. The curvature is zero everywhere but becomes infinite at the vertex and the edges where $\mu_j = \mu_1$ for more than one $j \neq 1$.

Approximately Unbiased Test

The AU test for regions with general smooth boundaries has been developed based on the theory of Efron et al. (1996). Consider a simplified model of the multivariate normal distribution with an identity covariance matrix

$$Y \sim N_M(\mu, I_M), \tag{9}$$

or equivalently, $Y_i, i = 1, \ldots, M$ are independently distributed as $N(\mu_i, 1)$, the normal distribution with mean μ_i and variance 1. This model appears to be oversimplified, because the log-likelihoods Y_i are correlated with each other in practice. The transformation-invariant property of the BP, however, justifies the following argument. For example, the normal model with an arbitrary covariance matrix is brought back to Eq. 9 by linear transformation, yet the BP values are invariant. One has only to assume the existence of such a smooth, possibly nonlinear, transformation to bring the problem back to Eq. 9; it is not necessary to know what the transformation is.

Suppose for the moment that region H_1 has smooth boundary ∂H_1, as shown in Figure 1a, where vector Y is indicated as a point. The "signed distance," denoted d, is the distance from Y to ∂H_1 with a positive or negative sign when Y is outside or inside of H_1, respectively. Thus,

$$d = \pm\sqrt{(Y_1 - \hat{\mu}_1)^2 + \cdots + (Y_M - \hat{\mu}_M)^2},$$

where $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_M)$ is the point on ∂H_1 closest to Y, that is, the projection of Y onto ∂H_1. If ∂H_1 is not curved at all, d is distributed normally with a variance of one, and its mean becomes zero for $\mu \in \partial H_1$. This is easily understood by considering the invariance of Eq. 9 under the rotation of the axes and taking $d = Y_1$ and $H_1 \colon \mu_1 \leq 0$ after the change of variables. Thus, the appropriate P-value of the test of H_1 will be

$$\mathrm{KH}_1 = 1 - \Phi(d), \tag{10}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. This is the case when $M = 2$, and Eq. 10 corresponds to the KH test (see Appendix remark 3).

If ∂H_1 is curved, however, the P-value of Eq. 10 is no longer valid. It follows from the Corollary of Efron (1985) that the appropriate P-value is

$$\mathrm{AU}_1 = 1 - \Phi(d - c), \tag{11}$$

where the number $c = c_1 - dc_2$ is related to the curvature of ∂H_1. Actually, c_1 and c_2 are geometric constants of magnitude $O(N^{-1/2})$ and $O(N^{-1})$, respectively, and are defined by the shape of the boundary at $\hat{\mu}$ (see Eq. 2.16 of Efron and Tibshirani [1998] and Appendix remark 4). Note here that $O(N^{\alpha})$ denotes "proportional to N^{α}." If ∂H_1 is flat, $c = 0$, and Eq. 11 reduces to Eq. 10. As the region becomes convex, the curvature c increases so that AU_1 effectively becomes larger. This is analogous to adjusting the selection bias in the SH test, as will be mentioned again later. Eq. 11 is known to be third-order accurate for similarity; that is, it controls Eq. 8 asymptotically up to the order $O(N^{-1})$ with an error of only $O(N^{-3/2})$.

Comparison of Tests

Comparing the existing P-values in light of the AU test, let us consider a class of the generalized AU P-values parameterized by κ

$$\mathrm{AU}_1(\kappa) = 1 - \Phi(d - \kappa c), \tag{12}$$

where the curvature c is multiplied by weight κ; $\kappa = 1$ corresponds to the AU test, and $\kappa = 0$ corresponds to the KH test. Note that the replicate Y^* with $N' = N$ for a given Y is distributed as

$$Y^* \sim N_M(Y, I_M), \tag{13}$$

and that by ignoring the sampling error of $O(B^{-1/2})$, the BP is expressed as

$$\mathrm{BP}_1 = 1 - \Phi(d + c), \tag{14}$$

which follows from the argument of Efron and Tibshirani (1998) (see Eq. 2.19 therein). This differs from Eq. 11 only by the sign of c, and thus a $\kappa = -1$ corresponds to the BP.

Zharkikh and Li (1995) provided a P-value by using two sets of bootstrap replicates: "complete" bootstrap with sequence length $N' = N$, and "partial" bootstrap with sequence length $N' < N$. Their P-value is intended to be less biased than the usual BP. Appendix remark 5 shows that the ZL test corresponds approximately to $\kappa = 3$.

The curvature c has the magnitude of order $O(N^{-1/2})$, so that P-values of the form in Eq. 12, with $\kappa \neq 1$, are first-order accurate: correct asymptotically up to the order $O(1)$ with error $O(N^{-1/2})$. Only the choice $\kappa = 1$ makes the AU test third-order accurate for a smooth boundary. For long sequences, $N^{-3/2} \ll N^{-1/2}$; therefore, third-order accurate P-values are less biased than first-order accurate P-values.

Applying the AU test to the selection problem raises another problem, however. Region H_i defined in Eq. 3 cannot be brought back to a region with a smooth boundary by any smooth transformation, because ∂H_i is not smooth at the vertex and the edges, as shown in Figure 1b. Region H_1 of the selection problem is actually a polyhedral convex cone in M-dimensional space. Although this singularity occasionally leads to a serious violation of Eq. 8, as shown later, the AU test is often useful in practice for the selection problem.

The SH test is not expressed in Eq. 12. However, it may be useful to see the curvature as the cause of selection bias that connects the

SH test to the AU test. The SH test adjusts the type-1 error at the least favorable configuration, where the selection bias is maximized, corresponding to the vertex of ∂H_1 shown in Figure 1b. This implies that SH_1 may be expressed by Eq. 11 but with c evaluated at the point where the curvature is maximized if the boundary is smooth everywhere. Thus, the SH test overestimates the selection bias unless μ is at the vertex.

Multiscale Bootstrap

Although Eq. 11 is highly accurate, the calculation of d and c is problematic in practice. This was first solved by Efron et al. (1996), using a second-level bootstrap. Let BP_0 be the BP calculated from the replicates around $\hat{\mu}$ instead of Y; $Y^* \sim N_M(\hat{\mu}, I_M)$. This corresponds to $d = 0$ so Eq. 14 leads to $\mathrm{BP}_0 = 1 - \Phi(c_1)$, from which c_1 is estimated. BP_1 gives the estimate of $d + c$, from which one can calculate

$$\mathrm{EH}_1 = 1 - \Phi(d - c'), \qquad (15)$$

where $c' = c_1 + dc_2$. This P-value is equivalent to that of Efron et al. (1996), if ignoring $O(N^{-3/2})$ terms. EH_1 is different from Eq. 11 by the sign of dc_2 and is second-order accurate; correct asymptotically up to the order $O(N^{-1/2})$ with error $O(N^{-1})$. The second-order accuracy is often accurate enough in practice, but the calculation of $\hat{\mu}$ is occasionally complicated in the implementation. This motivated the development of a new method.

The calculation of $\hat{\mu}$ is avoided by the newly devised multiscale bootstrap technique. In addition, the P-value becomes third-order accurate by taking full advantage of Eq. 11. The steps described below use d and c indirectly and thus are not very susceptible to changes in the model of Eq. 9.

Let $r = N'/N$ be the relative sequence length of the bootstrap replicate. If r is altered to differ from unity, the covariance matrix in Eq. 13 becomes I_M/r, and the scale becomes $1/\sqrt{r}$ of the usual bootstrap. The problem reduces to $r = 1$ by multiplying Y^* by the factor $\sqrt{r}$, but this causes d and c to become $d\sqrt{r}$ and $c/\sqrt{r}$, respectively. As a result of the scaling, Eq. 14 becomes

$$\mathrm{BP}_1(r) = 1 - \Phi(d\sqrt{r} + c/\sqrt{r}). \qquad (16)$$

A quite simple set of steps allows calculation of the P-value of the AU test. The key idea is to fit Eq. 16 to the BP values of different sequence lengths.

Step 1. Specify the scaling constants $r_1, \ldots, r_K$ and the number of replicates $B_1, \ldots, B_K$ for $K \geq 2$ sets of bootstrap replicates. In all of the examples shown later, $r_1 = 0.5$, $r_2 = 0.6, \ldots, r_{10} = 1.4$, and $B_1 = \cdots = B_{10} = 10,000$ for $K = 10$. See Appendix remark 6 for the choice of scales and remark 7 for the effective number of replicates.

Step 2. Generate B_k bootstrap replicates with sequence length $N' = r_k N$ for $k = 1, \ldots, K$,

$$Y^{*1}(r_k), \ldots, Y^{*B_k}(r_k),$$

and calculate the BP values

$$\mathrm{BP}_1(r_k) = \#\{Y^{*b}(r_k) \in H_1; \; b = 1, \ldots, B_k\}/B_k.$$

This resampling method is termed multiscale bootstrap, because several values of scale $1/\sqrt{r_k}$ are used. The rescaling approximation described in Appendix remark 8 is useful to reduce the computation.

Step 3. Estimate d and c by the weighted least squares (WLS) method, that is, by minimizing the residual sum of squares (RSS)

$$\mathrm{RSS}(d, c) = \sum_{k=1}^{K} \nu_k^{-1}\{d\sqrt{r_k} + c/\sqrt{r_k} - \Phi^{-1}[1 - \mathrm{BP}_1(r_k)]\}^2,$$

where the weight for each k is the inverse of the variance given by

$$\nu_k = \mathrm{BP}_1(r_k)[1 - \mathrm{BP}_1(r_k)]/(\phi\{\Phi^{-1}[\mathrm{BP}_1(r_k)]\}^2 B_k).$$

Note that $\Phi^{-1}(\cdot)$ and $\phi(\cdot)$ are the quantile function and the density function, respectively, of the standard normal distribution. Alternatively, the MLE of d and c can be used (see Appendix remark 9). When RSS is very large, the AU test should not be used (see Appendix remark 10).

Step 4. Calculate the P-value according to Eq. 11.

Edge Selection

Although only tree selection has been discussed, the methods described thus far can also be used for other types of hypotheses. The focus below is on the selection of edges, instead of trees, to test the monophyly of a specified group of taxa. Other problems, such as the test of congruence of phylogenies for several genes, are treated similarly.

Monophyly is rejected if the clade is not found in any of the nonrejected trees. This idea, as well as its modifications, is described formally as follows. An edge, often called a branch in phylogenetics, of unrooted trees divides the taxa into two groups and thus determines the clade as the one not including the outgroup. An edge in one tree is identified with one in another tree when these two edges determine the same split of leaves and thus determine the same clade. Consider m competing edges, say, edge-1, edge-2, ..., edge-m. For $e = 1, \ldots, m$, edge-e is included in some of tree-1, ..., tree-M, and the set of the indices of these trees is defined as S_e. The clade determined by edge-e is true if tree-i is true for some $i \in S_e$. In other words, the hypothesis corresponding to edge-e, denoted H_e, is rejected if all of $H_i, i \in S_e$ is rejected by $P_i < \alpha$ for $i \in S_e$. Thus, the P-value of edge-e, denoted P_e, is obtained as

$$P_e = \max_{i \in S_e} P_i. \qquad (17)$$

Equation 17 is applicable to the AU test, the SH test, and the WSH test. It is easy to check that P_e controls inequality 6 for H_e, when all $P_i, i \in S_e$ controls inequality 6:

$$\Pr\{P_e < \alpha\} \leq \Pr\{P_i < \alpha\} \leq \alpha,$$

when $\mu \in H_i$ for some $i \in S_e$, because $P_e \geq P_i$.

Although Eq. 17 is valid, P_e is no longer approximately unbiased, even if the AU$_i$ values are used in Eq. 17. To improve the test for monophyly, the AU test can be applied directly to H_e in exactly the same way as it is applied to H_i. The region corresponding to edge-e is the union of the regions of the associated trees, so that $H_e = \cup_{i \in S_e} H_i$. The P-value is denoted AU$_e$. Often smaller than the value obtained from Eq. 17, this P-value is less conservative.

It is also possible to improve the SH test for H_e (see Appendix remark 11).

Results and Discussion

Testing the Spherical Region

The spherical region H_1:$\sqrt{\mu_1^2 + \cdots + \mu_M^2} \leq R$ in M-dimensional space is simple enough to obtain the exact P-value and BP analytically: $P_1 = 1 - F_{M,R^2}[(R+d)^2]$, and $\mathrm{BP}_1 = F_{M,(R+d)^2}(R^2)$, where $F_{M,R^2}(\cdot)$ is the cumulative distribution function of the noncentral χ^2 with degrees of freedom $= M$ and noncentrality $= R^2$. In Table 1, the P-values BP$_1$, KH$_1$, AU$_1$, EH$_1$, and ZL$_1$ are calculated for several combinations of M and R, the values of d having been chosen to give an exact P-value of 0.05.

Only AU$_1$, which is third-order accurate, is very close to the exact value for all the combinations of M and R. EH$_1$, with second-order accuracy, is relatively good, except for the large curvature. The first-order–accurate P-values BP$_1$, KH$_1$, and ZL$_1$ are not very close to 0.05, except for the flat boundary $c = 0$. As c becomes larger, BP$_1$ and KH$_1$ tend to violate type-1 error, whereas ZL$_1$ becomes conservative. This agrees with the weight κ, which is multiplied to the extent of the selection bias measured by the curvature c.

TABLE 1. P-values (×100) for spherical regions.

				BP$_1$	KH$_1$	AU$_1$		ZL$_1$
M	R	d	c	(−1)	(0)	(1)	EH$_1$	(3)
1	5	1.64	0.00	5.00	5.00	5.00	5.00	5.00
4	10	1.78	0.14	2.73	3.73	5.01	5.25	8.58
4	5	1.90	0.26	1.55	2.88	5.05	6.03	13.1
10	10	2.05	0.41	0.69	2.00	5.02	5.90	20.6
10	5	2.38	0.75	0.09	0.87	5.16	9.46	44.6
20	10	2.49	0.85	0.04	0.64	5.05	7.48	52.5
30	10	2.92	1.28	0.00	0.18	5.08	9.87	82.0

d was chosen so that the exact P-value is 0.05. The MLE of d (not shown) and c were calculated by the multiscale bootstrap technique, and used for KH$_1$, AU$_1$, and ZL$_1$ by using Eq. 12 with the κ value indicated parenthetically. EH$_1$ was calculated by Eq. 15. Because the BP values were obtained analytically without simulation, the number of replicates $B = \infty$ for all P-values.

Simulation for the Normal Model

I performed a simulation to illustrate the P-value differences in the selection problem; the results are shown in Figure 2 and in Table 2. The vector Y of $M = 10$ is generated from Eq. 9, in which the log-likelihoods Y_i are not correlated with each other. When there are correlations among Y_i, as is the case in practical tree selection, the number of trees M in the simple model for simulation effectively changes. For example, positive correlations may effectively decrease M so that the selection bias becomes small.

The five types of P-values were calculated for tree-1. AU_1 and ZL_1 were calculated according to Eq. 12, with $\kappa = 1$ and $\kappa = 3$, respectively, and using d and c estimated by the multiscale bootstrap. The other three P-values, BP_1, KH_1, and SH_1, were calculated by their own definitions with the usual bootstrap of $r = 1$. In this simulation, the variance of $Y_j - Y_i$ is 2 for all the pairs, and thus the WSH test is equivalent to the SH test. Three simulation sets, each consisting of 10,000 repetitions, have been performed with different configurations of the means:

$$\mu_1 = \mu_2 \geq \mu_3 = \cdots = \mu_M,$$

where $M = 10$ and $\mu_1 - \mu_3$ is 5, 1, or 0. These configurations represent the points on the boundary of H_1.

For the three cases of the configuration, tree-1 and tree-2 are equally good with respect to μ_i values. If the model of evolution is correctly specified, and if the true tree is one of the M candidate trees, then the tie implies that the two trees are regarded as "true." Thus, the actual true tree is the consensus tree obtained from the "true" trees by shrinking the edges not shared by the two trees. However, this is not what is intended by the configurations. Rather, the tie represents the misspecification of the model of evolution. Assume that tree-1 is the true tree and the rest of the trees are not. Then $\mu_1 > \mu_2$ under the correct specification. As the model deteriorates, the difference $\mu_1 - \mu_2$ decreases (or possibly increases if the misspecification goes the other way), and eventually $\mu_1 = \mu_2$. Further misspecification makes μ across the boundary so that $\mu_1 < \mu_2$; in such a case, one may mistakenly conclude that tree-2 is better than tree-1 even when the infinite length sequences are available. What is intended by the configurations with $\mu_1 = \mu_2$ are the situations in which the misspecification is the worst in some direction within the limits that reasonable inference is possible. This is illustrated in a geometric argument of Shimodaira (2001). The geometry is shown clearly for a simple case in Yang (2000).

When $\mu_1 - \mu_3 = 5$, the difference is large enough that tree-3, ..., tree-10 are rarely comparable in Y_i value to tree-1 and tree-2. In most cases, the observed Y_i values imply that tree-3, ..., tree-10 are obviously worse than tree-1 and tree-2. Thus, in effect, only tree-1 and tree-2 are being compared, and the selection bias becomes marginal. As expected, the KH test works perfectly in this case. Figure 2 shows that KH_1 is distributed uniformly, and that the equality in Eq. 8 holds at this configuration for any $0 < \alpha < 1$. The unbiasedness requires Eq. 8 at any configuration on the boundary. It was confirmed that the equality actually holds at this configuration, as demonstrated in Table 2 for $\alpha = 0.05$, where the probability of rejecting tree-1 is estimated at 0.052. The same applies to the BP, the AU test, and the ZL test. However, the SH test behaves very differently. The distribution of SH_1 is heavily skewed to the right. The rejection probability of tree-1 is only 0.008, which is much smaller than $\alpha = 0.05$. Thus, the SH test is wasting the power and tends to include more trees in the confidence set than necessary. The average number of trees in the confidence set is 2.4 for the SH test, which is not very different from that for the AU test in this case. But the difference is multiplied for large M values, as will be evident in the real data analysis shown later.

When $\mu_1 - \mu_3 = 1$, the difference is not large enough to separate the two groups of trees. This situation will be typical when many similarly good trees are compared. As Figure 2 shows, the distribution of BP_1 is heavily skewed to the left. The probability of rejecting tree-1 is 0.250, which is much larger than $\alpha = 0.05$. This is alleviated in the KH test, but the rejection probability is still 0.086 for tree-1, indicating overconfidence in the wrong trees. The situation is markedly alleviated in the AU test, where the rejection probability becomes 0.057. The distribution of AU_1 is very close to the uniform distribution, though not perfect. The

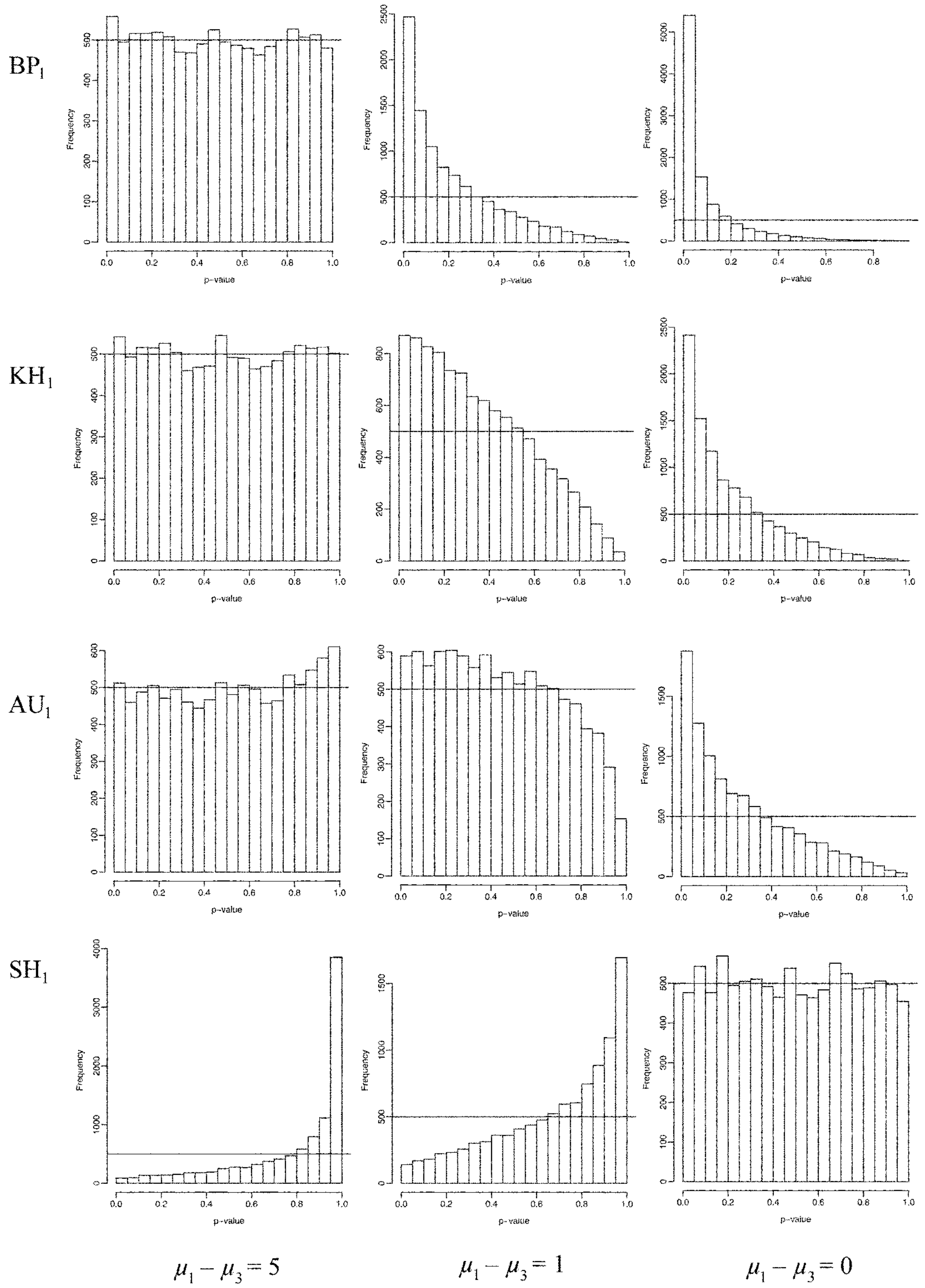

FIGURE 2. Distributions of P-values for tree-1. Ten trees are compared, and the means are $\mu_1 = \mu_2 \geq \mu_3 = \cdots = \mu_{10}$. Three configurations of the means with $\mu_1 - \mu_3 = 5$, 1, or 0 were examined by simulation of 10,000 repetitions for each. The frequencies of the P-values are shown for 20 bins of width 0.05. The expectation of the frequency for each bin should be 500 if the test is unbiased.

TABLE 2. Rejection probabilities at $\alpha = 0.05$.

	$\mu_1 - \mu_3$							
	5			1			0	
	Trees			Trees			Trees	
	1,2	3–10	#T	1,2	3–10	#T	1–10	#T
BP_i	0.053	0.999	1.9	0.250	0.649	4.3	0.538	4.6
KH_i	0.052	0.995	1.9	0.086	0.348	7.0	0.237	7.6
AU_i	0.049	0.994	2.0	0.057	0.287	7.6	0.190	8.1
SH_i	0.008	0.954	2.4	0.014	0.097	9.2	0.050	9.5
ZL_i	0.046	0.904	2.7	0.008	0.054	9.5	0.027	9.7

$\Pr\{P_i < \alpha\}$ was calculated from 10,000 repetitions, and the results were averaged over the trees having the same μ_i value. #T indicates the average number of trees in the confidence set.

observations on BP_1, KH_1, and AU_1 indicate that

$$\Pr\{\mathrm{BP}_1 < \alpha\} \geq \Pr\{\mathrm{KH}_1 < \alpha\}$$
$$\geq \Pr\{\mathrm{AU}_1 < \alpha\} \approx \alpha.$$

This agrees with the asymptotic argument of the smooth boundary. These P-values are expressed by Eq. 12 with κ values of -1, 0, and 1, respectively, and thus

$$\mathrm{BP}_1 \leq \mathrm{KH}_1 \leq \mathrm{AU}_1$$

because region H_1 is convex, and the curvature c is often positive in the selection problem.

The distribution of SH_1 is still skewed to the right for $\mu_1 - \mu_3 = 1$. The rejection probability of tree-1 for SH_1 is 0.014 at $\alpha = 0.05$, and that for ZL_1 is even smaller. This is not what one may wish, because the rejection probability of the wrong trees is also smaller than those of the other three methods, and the confidence set of trees becomes larger; the SH test and the ZL test still overestimate the selection bias in this case.

When $\mu_1 - \mu_3 = 0$, all 10 trees are equally good with respect to μ_i values. As explained earlier, that does not imply that all 10 trees are correct, but that the misspecification gave the wrong trees the same μ_i values as μ_1. This is rather an extreme situation in practice. The distributions of BP_1, KH_1, and AU_1 are left-skewed, and their rejection probabilities for tree-1 are much larger than $\alpha = 0.05$. Thus, these three tests are invalid in this case, whereas the ZL test still appears conservative. On the other hand, the SH test works perfectly here. SH_1 is distributed uniformly, and the rejection probability becomes 0.050. This is no surprise, because the SH test assumes the configuration $\mu_1 = \cdots = \mu_{10}$ to estimate the selection bias. This is the least favorable configuration at which the selection bias is maximized.

The failure of the AU test comes from the approximation of the nonsmooth H_1 boundary by the smooth boundary, as shown in Figure 1. As μ approaches the least favorable configuration, the extent of the failure as well as the selection bias increases. At this point the AU test underestimates the selection bias. The AU test is not exactly unbiased; it is only approximately unbiased, and it works well in cases where the selection bias is not extreme.

Analysis of Mammalian Mt Protein Sequences

The mammalian mt protein sequences from Shimodaira and Hasegawa (1999) were reanalyzed by using the same model of evolution as in that paper. They include the mt protein sequences of $N = 3{,}414$ amino acids for six mammalian species (human, seal, cow, rabbit, mouse, and opossum). The software package PAML (Yang, 1997) was used to calculate the site-wise log-likelihoods for the trees. The mtREV model (Adachi and Hasegawa, 1996b) was used for amino acid substitutions, and the site-heterogeneity was modeled by the discrete-gamma distribution (Yang, 1996). The results are shown in Tables 3–5.

The clade {seal, cow} was significantly supported in preliminary analysis, and thus only the 15 bifurcating trees with this clade are considered initially. Table 3 is essentially a reproduction of Table 1 of Shimodaira and Hasegawa (1999), except for the newly added

TABLE 3. Fifteen trees and the P-values of mammalian sequences.

Tree	T_i	PP$_i$	BP$_i$	KH$_i$	AU$_i$	SH$_i$	WSH$_i$	Tree form
1	−2.7	**0.934**	**0.579**	**0.639**	**0.789**	**0.944**	**0.948**	(((1(23))4)56)
2	2.7	**0.065**	**0.312**	**0.361**	**0.516**	**0.799**	**0.791**	((1((23)4))56)
3	7.4	0.001	0.036	**0.122**	**0.114**	**0.575**	**0.422**	(((14)(23))56)
4	17.6	0.000	0.013	0.044	**0.075**	**0.178**	**0.210**	((1(23))(45)6)
5	18.9	0.000	0.035	**0.066**	**0.128**	**0.149**	**0.299**	(1((23)(45))6)
6	20.1	0.000	0.005	0.049	0.029	**0.114**	**0.105**	(1(((23)4)5)6)
7	20.6	0.000	0.017	**0.051**	**0.101**	**0.112**	**0.252**	((1(45))(23)6)
8	22.2	0.000	0.001	0.032	0.009	**0.073**	**0.050**	((15)((23)4)6)
9	25.4	0.000	0.000	0.003	0.000	0.032	0.015	(((1(23))5)46)
10	26.3	0.000	0.003	0.019	0.028	0.034	**0.124**	(((15)4)(23)6)
11	28.9	0.000	0.000	0.010	0.003	0.018	**0.069**	(((14)5)(23)6)
12	31.6	0.000	0.000	0.003	0.001	0.006	0.033	(((15)(23))46)
13	31.7	0.000	0.000	0.003	0.001	0.006	0.034	(1(((23)5)4)6)
14	34.7	0.000	0.000	0.001	0.005	0.003	0.013	((14)((23)5)6)
15	36.2	0.000	0.000	0.001	0.002	0.002	0.009	((1((23)5))46)

Considered are the 15 trees with the clade of seal and cow, numbered by increasing order of T_i, the log-likelihood difference from the largest among the others (see Eq. 23). The log-likelihood values were calculated by PAML (Yang, 1997) software, and all the P-values were calculated by CONSEL (Shimodaira and Hasegawa, 2001) software. PP denotes the approximate Bayesian posterior probability taken from Table 1 of Shimodaira (2001). BP, KH, SH, and WSH were calculated from $B = 10{,}000$ replicates of $r = 1$. AU was calculated from the multiscale bootstrap with total $\sum B_k = 100{,}000$. The SH test and WSH (the weighted SH test) are referred to as MC and MS, respectively, in Shimodaira and Hasegawa (1999). The P-values that are not significant at $\alpha = 0.05$ are emphasized in bold type. The labels for the taxa used in the tree forms are 1 = *Homo sapiens* (human), 2 = *Phoca vitulina* (seal), 3 = *Bos taurus* (cow), 4 = *Oryctolagus cuniculus* (rabbit), 5 = *Mus musculus* (mouse), and 6 = *Didelphis virginiana* (opossum).

PP$_i$ and AU$_i$ columns. PP$_i$ is the approximate Bayesian posterior probability (PP) taken from Table 1 of Shimodaira (2001); the PP values are calculated by the Bayesian information criterion (BIC) approximation (Schwarz, 1978; Hasegawa and Kishino, 1989). For AU$_i$, the MLE is used to estimate d and c in Step 3 (above). The WLS gave practically the same results. The columns BP$_i$, KH$_i$, SH$_i$, and WSH$_i$ correspond to columns BP, KH, MC, and MS, respectively, of Shimodaira and Hasegawa (1999). The differences in the P-values between the two tables are due to the seeds of the random number generation. WSH denotes the weighted SH test suggested in Remark 4 of Shimodaira and Hasegawa (1999). In the WSH test, each difference $Y_j - Y_i$ is divided by the estimate of the standard error so that these terms are weighted equally for taking the maximum in Eq. 23.

The confidence set of trees obtained by the AU test at $\alpha = 0.05$ is $\{1, 2, 3, 4, 5, 7\}$, which is between those values obtained by the KH test and those by the SH test. The KH test rejects tree-4, whereas the SH test does not reject trees-6 and -8. If we want to be safe, we should take the eight trees obtained by the SH test; this assumes maximum selection bias. If we do not need to be so cautious, we take the six trees obtained by the AU test; this assumes moderate selection bias. If we are

TABLE 4. Ten edges and the P-values of mammalian sequences.

Edge	T_e	PP$_e$	BP$_e$	KH$_e$	AU$_e$	SH$_e$	WSH$_e$	Clade	Trees
1	−17.6	**1.000**	**0.927**	**0.956**	**0.954**	**0.994**	**0.991**	{1234}	1, 2, 3
2	−2.7	**0.934**	**0.592**	**0.639**	**0.749**	**0.910**	**0.921**	{123}	1, 4, 9
3	2.7	**0.065**	**0.318**	**0.361**	**0.469**	**0.754**	**0.735**	{234}	2, 6, 8
4	7.4	0.001	0.036	**0.122**	**0.111**	**0.567**	**0.411**	{14}	3, 11, 14
5	17.6	0.000	**0.065**	0.044	**0.075**	**0.177**	**0.253**	{45}	4, 5, 7
6	18.9	0.000	0.040	**0.066**	**0.088**	**0.147**	**0.277**	{2345}	5, 6, 13
7	20.6	0.000	0.019	**0.051**	**0.070**	**0.112**	**0.227**	{145}	7, 10, 11
8	22.2	0.000	0.004	0.032	0.016	**0.072**	**0.113**	{15}	8, 10, 12
9	25.4	0.000	0.000	0.003	0.000	0.032	0.031	{1235}	9, 12, 15
10	31.7	0.000	0.000	0.003	0.000	0.006	0.032	{235}	13, 14, 15

Listed are 10 edges that actively specify the 15 trees of Table 3, numbered by increasing order of $T_e = \min_{i \in S_e} T_{e,i}$ (see Eq. 21). For each edge, the P-values and the clade are shown. The list of trees including the edge-e, denoted S_e earlier, is also given. KH$_e$ is the KH test of T_e. There are $2^{6-1} - 1 = 31$ possible edges for the six taxa. In addition to the 10 edges, 15 edges actively specify the possible 105 trees of the six taxa. There are six other edges for clades {1}, {2}, {3}, {4}, {5}, and {12345}, which are always included in the trees.

TABLE 5. Best 20 of 105 trees of the mammalian sequences.

Tree	T_i	PP$_i$	BP$_i$	KH$_i$	AU$_i$	SH$_i$	WSH$_i$	Tree form
1	−2.7	**0.934**	**0.579**	**0.639**	**0.792**	**0.989**	**0.992**	(((1(23))4)56)
2	2.7	**0.065**	**0.312**	**0.361**	**0.517**	**0.930**	**0.908**	((1((23)4))56)
3	7.4	0.001	0.036	**0.122**	**0.115**	**0.841**	**0.594**	(((14)(23))56)
4	17.6	0.000	0.013	0.044	**0.076**	**0.577**	**0.338**	((1(23))(45)6)
5	18.9	0.000	0.035	**0.066**	**0.131**	**0.549**	**0.449**	(1((23)(45))6)
6	20.1	0.000	0.005	0.049	0.030	**0.506**	**0.175**	(1(((23)4)5)6)
7	20.6	0.000	0.017	**0.051**	**0.103**	**0.499**	**0.390**	((1(45))(23)6)
8	22.2	0.000	0.001	0.032	0.009	**0.458**	**0.082**	((15)((23)4)6)
9	25.4	0.000	0.000	0.003	0.000	**0.384**	0.024	(((1(23))5)46)
10	26.3	0.000	0.003	0.019	0.028	**0.363**	**0.193**	(((15)4)(23)6)
11	28.9	0.000	0.000	0.010	0.003	**0.309**	**0.108**	(((14)5)(23)6)
12	31.6	0.000	0.000	0.003	0.001	**0.258**	0.048	(((15)(23))46)
13	31.7	0.000	0.000	0.003	0.001	**0.255**	**0.052**	(1(((23)5)4)6)
14	34.7	0.000	0.000	0.001	0.005	**0.203**	0.017	((14)((23)5)6)
15	36.2	0.000	0.000	0.001	0.002	**0.177**	0.011	((1((23)5))46)
16	48.5	0.000	0.000	0.000	0.001	**0.050**	0.003	((((13)2)4)56)
17	49.2	0.000	0.000	0.000	0.000	0.047	0.002	((((12)3)4)56)
18	65.8	0.000	0.000	0.000	0.000	0.004	0.002	(((13)2)(45)6)
19	65.9	0.000	0.000	0.000	0.000	0.005	0.002	(((12)3)(45)6)
20	67.4	0.000	0.000	0.000	0.000	0.003	0.003	(((1(45))2)36)

Tree-21 through -105 have marginally small P-values. Tree forms and bold emphasis as in Table 3.

brave enough, we take the five trees obtained by the KH test; this assumes no selection bias.

ZL$_i$ values for the 15 trees in Table 3 are 0.922, 0.711, 0.270, 0.249, 0.336, 0.112, 0.382, 0.071, 0.003, 0.159, 0.028, 0.021, 0.004, 0.167, and 0.027, and the confidence set is $\{1, 2, 3, 4, 5, 6, 7, 8, 10, 14\}$. The ZL test appears to be too conservative in this case. On the other hand, the confidence set obtained by BP values consists of only two trees, presumably indicating overconfidence in these particular, but wrong, trees; tree-7 is best supported as the ML tree in recent analyses using the updated sequence data (Cao et al., 2000; Madsen et al., 2001; Murphy et al., 2001). The PP values are even more extreme.

There are 10 possible hypotheses of the monophyly for the 15 trees; their P-values are shown in Table 4. The AU test rejects edges-8, -9, and -10, which correspond to the clades {human, mouse}, {human, seal, cow, mouse}, and {seal, cow, mouse}. In addition, the KH test rejects edge-5, corresponding to the controversial clade {rabbit, mouse} for Glires (Lagomorpha + Rodentia), whereas the SH test does not reject edge-8, which corresponds to {human, mouse}. The differences in conclusions reflect the assumptions as to the extent of the selection bias.

Three edges in Table 4 are rejected by the AU test with $\mathrm{AU}_e < 0.05$. However, this does not necessarily imply strong support for the other seven edges. In fact, only edge-1 is strongly supported by the AU test. The AU P-value for the hypothesis that edge-e is not true is $1 - \mathrm{AU}_e$ as described in the Appendix, remark 12. This is significant only for edge-1 with $1 - 0.954 = 0.046 < 0.05$. Nothing can be stated with statistical significance for edges-2, -3, -4, -5, -6, and -7.

There seems to be inconsistency in the above argument. The strong support for edge-1 is incompatible with nonrejection of edges-5, -6, and -7. Given the logical relations among the hypotheses, special consideration is necessary. In fact, edge-1 is not strongly supported any more when all the H_e and the reverse of H_e, $e = 1, \ldots, m$, are tested simultaneously (see Appendix remark 13).

Tree selection changes interestingly when all 105 possible trees of the six taxa are compared. Comparing Table 5 with Table 3 shows almost no change for BP$_i$, KH$_i$, and AU$_i$, but SH$_i$ and WSH$_i$ change significantly. The SH test selected 8 of the 15 trees, but 16 of the 105 trees; Strimmer and Rambaut (2001) also calculated the confidence set in their Table 1. The effect is not as large in the WSH test, where the number of selected trees changed from 10 to 11. The increase in the size of the confidence set observed in the SH and WSH tests is attributable to the nature of the multiple comparisons. No matter how bad the added trees are, the size of the confidence set increases as the number of the candidate trees grows larger. The effect is compensated for, although not completely, by weighting.

Conclusions

Having discussed several tests and associated P-values, I summarize here recommendations as to when they are best used.

The AU test is recommended for general tree selection problems. It satisfies the requirement for unbiasedness at least approximately and thus controls for type-1 error in most cases. The AU confidence set is not susceptible to an increase in the number of candidate trees. However, the AU test must be used with caution when many of the best trees are nearly equally as good; one might miss the true tree in the confidence set by having overconfidence in the wrong trees. A breakdown of the AU test may be detected by the diagnostics mentioned in Appendix remark 10.

The SH test is safe to use and is a good option when the number of candidate trees is not very large. Control of type-1 error is excellent though conservative, and the conclusions are drawn safely; one will not miss the true tree in the confidence set, which is often larger than that for the AU test. A practical difficulty is that the SH test is very susceptible to an increase in the number of candidate trees. This is alleviated by weighting in the WSH test.

The AU test is computationally practical for trees numbering only a few thousand when the ML is used as the criterion. When more trees are compared, the KH test is the only option. After screening bad trees by the KH test at small α values, say, 0.001, apply the AU test. This two-step procedure may make sense, because the AU test is not affected very much by the number of candidate trees.

The AU test is applicable to other problems when the BP values are available. For example, it can be applied directly to tree selection based on other criteria such as parsimony, minimum evolution, or least squares. One has to count only how many times each tree is selected for sets of the replicates of several sequence lengths, and then proceed to step 3, as described earlier. The AU test is not confined to the selection problem; it is useful for general hypothesis testing of regions. The calculation of the confidence limits of a real parameter, say, μ_i, is also straightforward by the inversion of the significance tests, as implemented in the CONSEL software program.

Acknowledgments

The idea of the multiscale bootstrap came from a discussion with Ziheng Yang and was developed during my stay at Stanford University arranged by Brad Efron. The biological arguments are mostly those of Masami Hasegawa. Constructive comments were provided by the associate editor Nick Goldman and anonymous reviewers. I thank them as well as Joe Felsenstein, Hirohisa Kishino, and Susan Holmes for helpful discussions on phylogenetic problems. This work was supported by a Grant-in-Aid from JSPS (12780179).

References

Adachi, J., and M. Hasegawa. 1996a. MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28. Institute of Statistical Mathematics, Tokyo.

Adachi, J., and M. Hasegawa. 1996b. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42:459–468.

Buckley, T. R., C. Simon, H. Shimodaira, and G. K. Chambers. 2001. Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. Mol. Biol. Evol. 18:223–234.

Cao, Y., M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene 259:149–158.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 32:550–570.

Davison, A. C., and D. V. Hinkley. 1997. Bootstrap methods and their application. Cambridge Univ. Press, Cambridge, UK.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Ann. Statist. 7:1–26.

Efron, B. 1985. Bootstrap confidence intervals for a class of parametric problems. Biometrika 72:45–58.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13434.

Efron, B., and R. Tibshirani. 1998. The problem of regions. Ann. Statist. 26:1687–1718.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

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.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Hasegawa, M., and H. Kishino. 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution 43:672–677.

HASEGAWA, M., AND H. KISHINO. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. Mol. Biol. Evol. 11:142–145.

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

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

KISHINO, H., T. MIYATA, AND M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 30:151–160.

LEHMANN, E. L. 1952. Testing multiparameter hypotheses. Ann. Math. Stat. 23:541–552.

LEHMANN, E. L. 1986. Testing statistical hypotheses, 2nd edition. Wiley, New York.

LINHART, H. 1988. A test whether two AIC's differ significantly. South Afr. Stat. J. 22:153–161.

MADSEN, O., M. SCALLY, C. J. DOUADY, D. J. KAO, R. W. DEBRY, R. ADKINS, H. M. AMRINE, M. J. STANHOPE, W. W. DE JONG, AND M. S. SPRINGER. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

MARCUS, R., E. PERITZ, AND K. R. GABRIEL. 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660.

MURPHY, W. J., E. EIZIRIK, W. E. JOHNSON, Y. P. ZHANG, O. A. RYDER, AND S. J. O'BRIEN. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.

SANDERSON, M. J., AND M. F. WOJCIECHOWSKI. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). Syst. Biol. 49:671–685.

SCHWARZ, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

SHIMODAIRA, H. 1993. A model search technique based on confidence set and map of models. Proc. Inst. Stat. Math. 41:131–147 (in Japanese).

SHIMODAIRA, H. 1998. An application of multiple comparison techniques to model selection. Ann. Inst. Stat. Math. 50:1–13.

SHIMODAIRA, H. 2000a. Another calculation of the p-value for the problem of regions using the scaled bootstrap resamplings. Tech. Rep. No. 2000-35. Stanford Univ., Palo Alto, California.

SHIMODAIRA, H. 2000b. Approximately unbiased one-sided tests of the maximum of normal means using iterated bootstrap corrections. Tech. Rep. No. 2000-7. Stanford Univ., Palo Alto, California.

SHIMODAIRA, H. 2001. Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. Commun. Stat. A Theory Methods 30:1751–1772.

SHIMODAIRA, H., AND M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

SHIMODAIRA, H., AND M. HASEGAWA. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

STRIMMER, K., AND A. RAMBAUT. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. R. Soc. London B 269:137–142.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

VUONG, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57:307–333.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

YANG, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556.

YANG, Z. 2000. Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. London B 267:109–116.

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

ZHARKIKH, A. AND W.-H. LI. 1995. Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. 4:44–63.

First submitted 23 Sep. 2001; Revision submitted 19 Jan. 2002; Final acceptance 5 Feb. 2002

Associate Editor: Nick Goldman

APPENDIX

This appendix is a collection of remarks and technical details.

1. *Monotone tests.* Consider two vectors of log-likelihoods Y and Y' such that

$$Y'_j - Y_j \geq Y'_1 - Y_1, \quad j = 2, \ldots, M,$$

where notations are defined in Methods (see Eqs. 1 and 3). To reject H_1, Y' is regard as stronger evidence than Y. H_1 is rejected by Y' if H_1 is rejected by Y. This property is said to be the monotonicity of the test. It follows from Lehmann (1952) that the monotone test with minimum bias is the WSH test. Thus, the WSH test is the optimal method within the class of the monotone tests. It also follows from Lehmann (1952) that an exactly unbiased test does not exist for the selection problem.

2. *Consistency of tree selection.* For two densities $f(x)$ and $g(x)$, the Kullback–Leibler divergence is defined by $D(f, g) = \int f(x)(\log f(x) - \log g(x))\,dx \geq 0$, which takes the minimum value zero only when $g(x) \equiv f(x)$. In other words, the expected value of the log-likelihood $\log g(x)$ under the true model $f(x)$ is maximized when $g(x) \equiv f(x)$. This implies that the tree of the largest μ_i value, denoted tree-i_0, represents the true history of evolution, provided one of the M trees is correct and the probabilistic model of the evolution is correctly specified. For a given N', the tree selection selects tree-i^*, which maximizes Y^*_i. It follows from the law of large numbers that Y^*_i converges to the expected value μ_i as $N' \to \infty$, and thus the probability of selecting the true tree converges to 1, that is, $i^* \to i_0$, provided the trees are identifiable, that is, i_0 is uniquely defined. In practice, the evolution model is misspecified, but the consistency still holds if the misspecification is

minor, in which case the true tree has the largest μ_i value.

3. *Shortcut to the AU test.* It follows from Eqs. 10, 11, and 14 that another representation of AU_1 is

$$\text{AU}_1 = \Phi(2\Phi^{-1}(\text{KH}_1) - \Phi^{-1}(\text{BP}_1)), \qquad (18)$$

which implies that many published results of tree selection can be readily reanalyzed by the AU test without using the multiscale bootstrap. Unfortunately, Eq. 18 does not work very well, because Eq. 10 is not a very precise approximation for the selection problem.

4. *Curvature of the boundary.* One can take the local coordinates $(u_1, \ldots, u_{M-1}, \nu)$ at $\hat{\mu}$ so that $\hat{\mu}$ is expressed as $(0, \ldots, 0)$, the observed vector Y is expressed as $(0, \ldots, 0, d)$, and the hypersurface is expressed as $\nu = -(\lambda_1 u_1^2 + \cdots + \lambda_{M-1} u_{M-1}^2) + O(N^{-1})$. The geometric constants are then given by $c_1 = \lambda_1 + \cdots + \lambda_{M-1}$ and $c_2 = \lambda_1^2 + \cdots + \lambda_{M-1}^2$.

5. *ZL test.* Zharkikh and Li (1995) provided a P-value closely related to that of the AU test. Their complete-and-partial bootstrap is equivalent to the multiscale bootstrap with $K = 2$ and can be reformulated for any $K \geq 2$. First, Eq. 16 here is replaced by their Eq. 34, expressed as

$$\text{BP}_1(r) = 1 - \Phi(Z\sqrt{r} - \Phi^{-1}(1/M')\sqrt{1+r}), \qquad (19)$$

where Z, M', and r correspond to the $-Z$, K, and $1/r$, respectively, of Zharkikh and Li. By the WLS method of step 3, Z and $\Phi^{-1}(1/M')$ are estimated, and the P-value is given by $\text{ZL}_1 = 1 - \Phi(Z)$, which was denoted W_1. Comparing Eqs. 16 and 19, and considering the Taylor expansion around $r = 1$, yields $Z = d - 3c$ and $\Phi^{-1}(1/M') = -2\sqrt{2}c$, and thus $\kappa = 3$.

6. *Choice of scale parameters.* The choice of scale parameters,—K, r_k, and B_k—is another practical issue and is a matter of the experimental design. For example, the computational cost of the multiscale bootstrap will be expressed as $\sum_{k=1}^{K} a_k B_k$, using as the cost per sample $a_k > 0$, and the optimal choice of the scale parameters is sought to minimize the standard error of AU_1 while fixing the cost. I tried this for a simple problem by a numerical optimization using d and c obtained by a preliminary multiscale bootstrap; the standard error was minimized when $K = 2$, r_1 is close to zero, and r_2 is slightly larger than unity. However, taking a large range of r_k leads to a breakdown of the theory, and $K \geq 3$ is necessary to detect it. The optimal choice is not simple in practice. Instead, a fixed set of scale parameters is used in the examples.

7. *Effective number of replicates.* Considering binomial sampling, the standard error of BP_1 is

$$\text{se}(\text{BP}_1) = \sqrt{\text{BP}_1(1 - \text{BP}_1)/B},$$

where B is the number of the bootstrap replicates for $r = 1$. On the other hand, the standard error of AU_1 is given by

$$\text{se}(\text{AU}_1) = \phi(\Phi^{-1}(\text{AU}_1))\sqrt{\nu_{11} - 2\nu_{12} + \nu_{22}},$$

where ν_{11}, ν_{12}, ν_{22} are the elements of the matrix

$$\left(\sum_{k=1}^{K} \nu_k^{-1} \begin{pmatrix} r_k & 1 \\ 1 & r_k^{-1} \end{pmatrix}\right)^{-1}.$$

This is converted to the effective number of replicates

$$B_{\text{eff}} = \frac{\text{AU}_1(1 - \text{AU}_1)}{\text{se}(\text{AU}_1)^2},$$

which represents the number of replicates as if AU_1 were obtained the same way as BP_1. In the simulation for the normal model with the configuration $\mu_1 - \mu_3 = 1$, the average of the B_{eff} for tree-1 is 2,126, when the actual total number of replicates $\sum_{k=1}^{K} B_k$ is 100,000. This further reduces to 781 if only those having $\text{AU}_1 < 0.1$ are averaged. This implies that the multiscale bootstrap needed a hundred times more bootstrap replicates than the usual bootstrap to have the same standard error. In this sense, the other implementation of the AU test given by Efron et al. (1996) may be advantageous to the multiscale bootstrap; it requires a much smaller number of replicates once $\hat{\mu}$ is obtained. However, the computation appears not to be a practical problem currently if an efficient method such as RELL is used for resampling. Using a 400 MHz pentium computer took less than 30 min to obtain all the P-values of the mammalian sequences.

8. *Rescaling approximation.* For a large problem, the following rescaling approximation is useful to reduce the computation of multiscale bootstrap. First, generate Y^{*b}, $b = 1, \ldots, B_0$, with, say, $r_0 = 1$ and $B_0 = 10,000$. Then, in step 2 (above), we calculate

$$Y^{*b}(r_k) = \bar{Y} + \sqrt{r_0/r_k}\,(Y^{*b} - \bar{Y}), \quad b = 1, \ldots, B_0,$$

where $\bar{Y} = \sum_{b=1}^{B_0} Y^{*b}/B_0$, and $B_k = B_0$ for $k = 1, \ldots, K$. The total number of replicates is KB_0, but resampling is performed only B_0 times. The P-values calculated by this approximation are 0.798, 0.511, 0.130, 0.069, 0.110, 0.034, 0.124, 0.012, 0.000, 0.020, 0.000, 0.002, 0.000, 0.000, and 0.000 for the 15 trees in Table 3, and 0.958, 0.759, 0.461, 0.126, 0.071, 0.078, 0.076, 0.014, 0.001, and 0.000 for the 10 edges in Table 4. Thus, the approximation gives practically equivalent results but reduces the computation to one tenth.

9. *MLE of* d *and* c. Step 3 may be replaced by the MLE instead of the WLS. Parameters d and c are estimated by maximizing the log-likelihood for the binomial distributions

$$L(d, c) = \sum_{k=1}^{K} B_k\{\text{BP}_1(r_k)\log \pi_k(d, c) + [1 - \text{BP}_1(r_k)]\log[1 - \pi_k(d, c)]\},$$

where $\pi_k(d, c) = 1 - \Phi(d\sqrt{r_k} + c/\sqrt{r_k})$. The Newton–Raphson method is used for the optimization, with the initial value obtained by the WLS. Although the MLE and the WLS often provide practically equivalent results, the MLE may perform better than the WLS when BP_1 is very close to zero or unity. Both the WLS and the MLE are implemented in the CONSEL software application.

10. *Diagnosing the breakdown of the theory.* A large RSS value in step 3 indicates a breakdown of the asymptotic theory; RSS is distributed as χ^2 with $K-2$ degrees of freedom.
11. *SH test for edge selection.* The SH test of H_e is improved by the following argument, but the P-value will not differ much from that obtained by Eq. 17. First note that the hypothesis of edge-e is expressed as

$$H_e : \cup_{i \in S_e} \{\mu_i \geq \mu_j,\ j \notin S_e\}, \tag{20}$$

where the subscript j runs through $1, \ldots, M$, except for the elements in S_e. The SH test applied to each member in the union of Eq. 20 uses the test statistic

$$T_{e,i} = \max_{j \notin S_e}(Y_j - Y_i). \tag{21}$$

The P-value corresponding to $T_{e,i}$, denoted $P_{e,i}$, is calculated from the replicates of Y generated with "centering" as described in Shimodaira and Hasegawa (1999);

$$T_{e,i}^{*b} = \max_{j \notin S_e}\left(\left(Y_j^{*b} - \bar{Y}_j\right) - \left(Y_i^{*b} - \bar{Y}_i\right)\right),$$

where $\bar{Y}_i = B^{-1} \sum_{b=1}^{B} Y_i^{*b}$, or simply let $\bar{Y}_i = Y_i$, and thus

$$P_{e,i} = \#\{T_{e,i} > T_{e,i}^{*b}, b = 1, \ldots, B\}/B.$$

Because H_e is rejected if and only if all the members in Eq. 20 are rejected, that is, $P_{e,i} < \alpha$ for all $i \in S_e$, this yields the P-value of H_e

$$\mathrm{SH}_e = \max_{i \in S_e} P_{e,i}. \tag{22}$$

Moreover this reduces to Eq. 17 if Eq. 21 is replaced with the test statistic of the SH test for tree-i.

$$T_i = \max_{j \neq i}(Y_j - Y_i). \tag{23}$$

12. *Reversing hypothesis.* The geometry of Figure 1 does not change even if the roles of the null and alternative hypotheses are exchanged but only the signs of d are c are reversed. Thus, the generalized AU P-value for the null hypothesis $\mu \notin H_1$ against alternative $\mu \in H_1$ is obtained as $1 - \mathrm{AU}_1(\kappa)$ from Eq. 12. This almost obvious formula does not apply to the SH and WSH tests, for which the least favorable configuration is not the same for the reversed problem.
13. *A closed testing procedure.* When tree-7, say, is true, edges-5 and -7 as well as the reversed hypotheses of edges-1, -2, -3, -4, -6, -8, -9, and -10 are true. The type-1 family-wise error (FWE) rate is the probability of rejecting any of these true hypotheses. The AU test controls the type-1 error rate for each of the true hypotheses separately, but not simultaneously. The closure method of Marcus et al. (1976) is applied to the AU test for controlling the type-1 FWE through simply using Eq. 17. The P-value for the reversed hypothesis of edge-1 becomes 0.128, the maximum of AU_i, $i = 4, \ldots, 15$.