Tests for Two Trees Using Likelihood Methods

Edward Susko*[1]

[1]Department of Mathematics and Statistics & Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author:** E-mail: susko@mathstat.dal.ca.

Associate editor: Oliver Pybus

Abstract

This article considers two similar likelihood-based test statistics for comparing two fixed trees, the Kishino-Hasegawa (KH) test statistic and the likelihood ratio (LR) statistic, as well as a number of different methods for determining thresholds to declare a significant result. An explanation is given for why the KH test, which uses the KH test statistic and normal theory thresholds, need not give correct type I error probabilities under the appropriate null hypothesis. Simulations show that the KH test tends to give much smaller type I error probabilities than expected. The article presents a computationally efficient normal-theory parametric bootstrap method for determining better KH test statistic thresholds. For the LR statistic, existing mixture of chi-squares results for determining thresholds are extended to cases in which a tree with two or three zero edge-lengths exhibits the two trees being compared. The resulting chi-bar test and use of the KH test statistic with normal bootstrap are shown through simulation to give good performance but are more difficult to implement than the KH test. Two conservative approaches are presented which require only log likelihoods and simple chi-square thresholds. While they did not perform as well as chi-bar and normal bootstrap methods in the simulations considered, they gave better performance than the KH test and have just as simple an implementation. As a by-product of parametric bootstrap considerations, an adjustment to the Swofford-Olsen-Waddell-Hillis (SOWH) test is proposed.

Key words: **maximum likelihood, topology test, KH test, SOWH test.**

Introduction

The Kishino-Hasegawa (KH) test developed in Hasegawa and Kishino (1989) and Kishino and Hasegawa (1989) is the most widely used test of whether there is significant evidence for a particular hypothesized tree under the null hypothesis that a different but fixed tree is correct. It has been implemented in a number of popular packages including TREE-PUZZLE (Schmidt et al. 2002) and PAML (Yang 1997, 2007). Shimodaira and Hasegawa (1999) and Goldman et al. (2000) point out that the KH test is inappropriate if one of the two trees is chosen in a data-dependent manner; this is sometimes referred to as selection bias. The most frequent data-dependent choice is the estimated tree which, in the case of maximum likelihood (ML) estimation as used in the KH test, leads to a test with higher than expected false-positive or type I error rate.

Although two-tree tests like the KH test can be inappropriate in some settings due to selection bias, they continue to be widely reported and there are a number of legitimate reasons for continued interest in them. First, cases do arise where a priori hypotheses about correct trees can be made; for example, the Coelomata and Ecdysozoa trees considered in Dopazo and Dopazo (2005). Second, the distinction between data-dependent tree choice and data-independent choice is not always clear. A tree can be both the ML tree and be considered a priori reasonable. The KH test provides useful supplementary information in such cases about what inference would be made had the tree been fixed. Third, two-tree tests can provide relevant information even if a selection bias is present. For a given two-tree test, if a Tree 2 is not rejected when the test is applied to it and the ML tree, that Tree 2 would not have been rejected had a selection bias correction been made to the test. Finally, two-tree tests can be viewed as a basis for tests that adjust for selection bias. For instance, the SH test of Shimodaira and Hasegawa (1999) is the selection bias-adjusted version of the KH test and the Swofford-Olsen-Waddell-Hillis (SOWH) test described originally in Swofford et al. (1996) and considered in greater detail in Goldman et al. (2000) can be viewed as a selection bias-adjusted version of two-tree parametric bootstrapping. Anisimova and Gascuel (2006) illustrate how Bonferroni adjustment can be used to give a selection bias-adjusted version of the Ota et al. (2000) test for a significant split.

This article examines two-tree tests based on either the KH test statistic or likelihood ratio (LR) statistic. The two trees being tested are assumed fixed a priori so that the issue of selection bias does not need to be corrected for. An argument is given in this article that the motivation for the KH test is problematic and that the test cannot be expected to give correct type I error probabilities. Results indicate that the test can be grossly conservative. Such conservativeness is not due to KH test difficulties discussed in Shimodaira and Hasegawa (1999) and Goldman et al. (2000) as there is no selection bias in the cases considered. The result also helps to explain why KH test *P* values tend to be larger than SOWH *P* values; the SOWH adjustment for selection suggests the

Article

opposite. Determining thresholds for the KH test statistic through parametric bootstrapping, the two-tree analog of the SOWH test, is more appropriate. However, a familiar mammalian mitochondrial data example will be used to explain how small P values from the SOWH test, and its fixed two-tree analog, can also be due to the choices of the edge-lengths in the trees used for simulation.

An alternative method for KH test statistic threshold determination is presented here which uses less intensive normal simulations as a proxy for parametric bootstrapping. Extension of the Ota et al. (2000) LR statistic test, which will be referred to as the chi-bar test, provides an alternative test in some cases. These tests are more difficult to implement than the KH test, but their performance is found to be much better. Alternative testing methods that are conservative and use either the LR statistic or KH test but with simple chi-square thresholds are also considered. Although they are not as powerful as either the chi-bar test or using the KH test statistic with normal simulation, they are found to be more powerful than the KH test and just as easy to implement.

Theory and Methods

One-sided tests are considered: whether there is significant evidence for a particular Tree 1 by comparison with a particular Tree 2. All of the methods that will be considered use a log LR as a test statistic. Let k denote a site pattern. For instance, with four taxa, $k = ACGG$ gives a nucleotide site pattern where an A, C, G, and G were observed for taxa 1–4, respectively. Let $p_k(\boldsymbol{\theta}; j)$ denote the probability of site pattern k for tree j and parameter $\boldsymbol{\theta}$, including edge-lengths and possibly other substitution parameters. Then, assuming a conventional model where evolution at sites is independent, the maximized log likelihood for tree j is

$$\sum_{i=1}^{n} \log p_{k_i}(\widehat{\boldsymbol{\theta}}_j; j)$$

where $\widehat{\boldsymbol{\theta}}_j$ is the ML estimate for tree j based on all sites, k_i is the site pattern for site i, and n is the number of sites.

The KH Test

The original version of the KH test applied a z-test to the site log likelihood differences for the two trees

$$d_i = \log p_{k_i}(\widehat{\boldsymbol{\theta}}_1; 1) - \log p_{k_i}(\widehat{\boldsymbol{\theta}}_2; 2) \quad (1)$$

Recall that the one-sample z-test of $H_0 : E[d_i] = 0$ against $H_A : E[d_i] > 0$ has P value

$$p = P(Z > \bar{d}) \quad (2)$$

where $\bar{d}$ is the mean d_i and Z has a $N(0, s_d^2/n)$ distribution; here s_d^2 denotes the sample variance of the d_i. Let $\Lambda_2 = n\bar{d}$ denote the KH test statistic. Then, alternatively, the KH test checks whether the KH test statistic is larger than expected from a $N(0, ns_d^2)$ distribution.

A variation of the KH test replaces the $N(0, s_d^2/n)$ distribution in equation (2) by a bootstrap distribution (Kishino et al. 1990). Specifically, d_i are sampled with replacement to obtain a new $\bar{d}^*$ that is the mean d_i sampled. The process is commonly referred to as RELL for resampling estimated log likelihoods and is repeated a large number of times yielding a large number of $\bar{d}^*$. The p in equation (2) is replaced by the proportion of $\bar{d}^* - \text{ave}_b\bar{d}^*$ that are larger than $\bar{d}$. Here, $\text{ave}_b\bar{d}^*$ is the average $\bar{d}^*$, averaged over bootstrap samples. The $\bar{d}^* - \text{ave}_b\bar{d}^*$ is a centering step to make sure that the log likelihood difference follows the expected value for the differences to be 0, as assumed for the null distribution. If n is relatively large and enough bootstrapped $\bar{d}^*$ are obtained, it turns out that RELL will give P values that are approximately the same as those using the normal distribution in equation (2), no matter what the data (Bickel and Freedman 1981). In the examples considered, n was large enough that results with RELL were almost identical to results using equation (2). As a consequence, simulations only considered performance of the KH using the normal distribution in equation (2).

The Null Hypothesis

The null hypothesis of the KH test states only that $H_0 : E[d_i] = 0$. The Goldman et al. (2000, p. 664) review of the KH test specified a tree as part of the null hypothesis, but the original reference of Kishino and Hasegawa (1989), which only briefly discusses the testing setting, does not clearly indicate that Tree 2 is assumed under the null. As a statement of significant support in favor of Tree 1 over Tree 2 is not very meaningful when a Tree 3 that is very different from Tree 1 generates the data, the null hypothesis throughout this article includes that Tree 2 is correct. More specifically, the null tree throughout the article is a strict consensus tree that sets as many edge-lengths to zero as is needed to make Tree 1 and 2 equivalent. For instance, if Tree 1 and 2A are being compared in figure 1, this leads to Tree 3A which sets one edge-length to 0. If comparison is between Trees 1 and 2B, two edge-lengths are set to 0, giving Tree 3B and comparison between Trees 1 and 2C gives the star Tree 3C. Although this choice differs somewhat from the null of KH test, it is approximately equivalent as will be established below.

There are a number of arguments that the consensus tree gives the correct choice of null tree. First, it is conventional in hypothesis testing to evaluate type I error at a point on the intersection of the boundaries of the null and alternative hypothesis; type I error for the z-test of mean hypothesis $H_0 : \mu \leq 0$ against $H_A : \mu > 0$ is evaluated at $\mu = 0$. Consensus trees are the points on the intersection of the boundary of tree space for Tree 1 and the boundary of tree space for Tree 2; see supplementary material, Supplementary Material online, for additional details. Second, an α-level test should satisfy that the probability of false rejection under the null is at most α. In the present case, the null is that Tree 2 is correct and any sensible test will be less likely to reject for positive edge-lengths than if edge-lengths are set to 0 in such a way that Tree 1 and Tree 2 are equivalent. Thus, to ensure a false-positive rate at most α, one needs to ensure that the false-positive rate is at most α for the tree giving both Tree 1 and Tree 2. Third, if Tree 2 is estimated, it will certainly not be rejected in favor of Tree 1. One needs only be concerned with

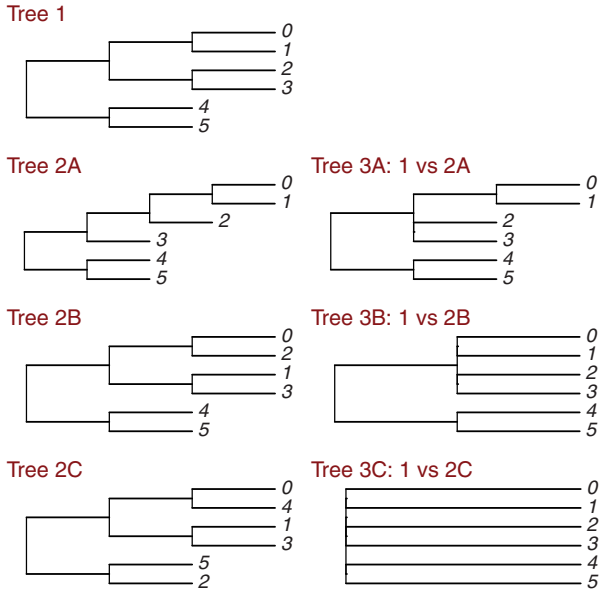

Fig. 1. Tree 1 gives the tree that significant evidence is being sought for when compared with Trees 2A–2C. The Trees 3A–3C give the corresponding null trees obtained by collapsing edges in Tree 1 to make it equivalent to Trees 2A–2C, respectively.

thresholds for rejection when Tree 1 is estimated, in which case it is natural to determine these thresholds for the tree satisfying the null hypothesis with smallest distance to Tree 1. For both the branch score distances of Kuhner and Felsenstein (1994) and geodesic distances of Billera et al. (2001), and any given Tree 1, the least distant version of Tree 2 is one with edge-lengths set to 0 to make the two trees equivalent; see supplementary material, Supplementary Material online, for additional details.

That a consensus tree null hypothesis is approximately equivalent to the KH null hypothesis follows from arguments associated with consistency of ML estimation. Suppose the null hypothesis includes in addition to $E[d_i] = 0$ that the generating tree is Tree 2, but not necessarily the consensus tree. Then, under the null, for large samples, due to the consistency of ML estimation, the ML estimates of parameters for Tree 2 will be approximately the same as the generating parameters with large probability. Arguments of White (1982) indicate that the ML estimates of parameters for Tree 1 will converge to particular parameters as well. As a consequence, $p_k(\widehat{\boldsymbol{\theta}}_j; j) \approx p_k(\boldsymbol{\theta}_{j0}; j)$, where $\boldsymbol{\theta}_{j0}$ denotes the parameters converged upon. As expectations are calculated using the generating distribution, $p_k(\boldsymbol{\theta}_{20}; 2)$, for large samples,

$$E[d_i] \approx \sum_k p_k(\boldsymbol{\theta}_{20}; 2) \log[p_k(\boldsymbol{\theta}_{20}; 2)/p_k(\boldsymbol{\theta}_{10}; 1)] \quad (3)$$

where the sum is over all possible site patterns. However, a key property giving rise to consistency of ML estimation is that the sum in equation (3), which is sometimes referred to as the Kullback-Leibler divergence, is positive unless $p_k(\boldsymbol{\theta}_{20}; 2) = p_k(\boldsymbol{\theta}_{10}; 1)$ for all patterns k (cf. Section 9.3 of Pawitan 2001). For conventional continuous-time Markov models as well as a number of models that allow variation in rates across sites (cf. Chang 1996; Allman et al. 2008, 2012), it has been shown that the only way in which probabilities of all data patterns can be equal for two trees is if the two trees are equivalent. To do this, the edge-length corresponding to incompatible splits need to be set to 0. Thus for any generating tree different than this tree, the expected difference in maximized log likelihoods will eventually differ substantially from 0 and the KH null hypothesis will not be satisfied. The only generating Tree 2 that will approximately have a zero expected log likelihood difference is the one that sets edge-lengths for incompatible spits to 0.

LR Tests

The null hypothesis of interest is a special case of Tree 1 with some edge-lengths constrained to make it equivalent to Tree 2. The problem thus can be recast as a conventional parametric testing problem. Tree 1 is fixed, the null hypothesis is that a particular subset of its edges have zero length (those that are 0 in the consensus tree of Trees 1 and 2), and the alternative hypothesis that some of these are positive. These are nested hypotheses and LR tests can be applied. The LR statistic is $2\Lambda_3$ where

$$\Lambda_3 = \sum_{i=1}^{n} \log p_{k_i}(\widehat{\boldsymbol{\theta}}_1; 1) - \sum_{i=1}^{n} \log p_{k_i}(\widehat{\boldsymbol{\theta}}_3; 2) \quad (4)$$

and $\widehat{\boldsymbol{\theta}}_3$ denotes the ML estimate for Tree 2 (or equivalently Tree 1) when edge-lengths are constrained to be 0 to make Trees 1 and 2 equivalent.

Ota et al. (2000) consider the LR test and note that, because of the boundary constraints that some edges are 0, the appropriate null distribution is not a chi-square distribution but rather a mixture of chi-squares. In particular, they consider the case where the null tree has a single zero edge-length and show that the large sequence-length distribution of Λ_3 is $(1/2)\chi_0^2 + (1/2)\chi_1^2$. Consistent with Susko (2013) and references therein, this test and the extensions presented later on are referred to as chi-bar tests.

KH Test Statistic Thresholds from Full Bootstrapping

The final variations of the KH test that reviewed involve bootstrap methods (Efron 1982) whose application to two-tree tests is described in Goldman et al. (2000). The nonparametric version is as follows:

1) Generate sites with replacement to obtain a bootstrapped data set.
2) Obtain ML estimates $\widehat{\boldsymbol{\theta}}_1^*$ and $\widehat{\boldsymbol{\theta}}_2^*$ for the bootstrapped data and use these to calculate the KH test statistic Λ_2^* for the bootstrapped data.
3) Repeat 1-2 a large number of the times to obtain a large number of Λ_2^*.
4) Center the Λ_2^* by subtracting the average Λ_2^*, averaged over all bootstrapped data sets.
5) Calculate a P value as the proportion of Λ_2^* larger than Λ_2.

The process is referred to in Goldman et al. (2000) as full bootstrapping, as the RELL version of the KH test can

alternatively be characterized as nonparametric bootstrapping with the time-saving device of replacing $\widehat{\theta}_j^*$ by $\widehat{\theta}_j$ in step 2.

Parametric bootstrapping replaces step 1 with generation from the fitted Tree 2 with parameters $\widehat{\theta}_2$. As Tree 2 is the generating tree under the null hypothesis, the version mentioned in Goldman et al. (2000) skips the centering step 4. Parametric bootstrapping with centering will also be considered in what follows.

As the null hypothesis considered here is a consensus tree, an alternative form of parametric bootstrapping not considered in Goldman et al. (2000) is to replace step 1 with generation from the fitted consensus Tree 2 with parameters $\widehat{\theta}_3$. This can be considered with or without centering.

KHns: KH Test Statistic Thresholds from Normal Simulation

Full parametric or nonparametric bootstrapping provides valid replacements for the normal distribution used to obtain the P value in equation (2). Their computational cost can, however, be substantial. In this section, a faster simulation method is presented.

The method description requires a bit of notation. With parameters ordered so that the first ones are the edge-lengths that are 0 under the null hypothesis (to make the two trees equivalent), the same true generating value, θ_0, applies to both trees. For tree j, let $\mathbf{S}^{(j)}(\theta_0)$ denote the vector of derivatives of the log likelihood, which has two blocks of entries, $\mathbf{S}_b^{(j)}$ and $\mathbf{S}_r^{(j)}$, where $\mathbf{S}_b^{(j)}$ gives the derivatives with respect to the first zero-valued edge-length parameters and $\mathbf{S}_r^{(j)}$ the rest of the derivatives. Let $-nI^{(j)}(\theta_0)$ denote the expected matrix of second derivatives of the log likelihood. Then, $I^{(j)}(\theta_0)$ can be similarly decomposed as

$$I^{(j)}(\theta_0) = \begin{bmatrix} I_b^{(j)} & I_{br}^{(j)} \\ I_{rb}^{(j)} & I_r^{(j)} \end{bmatrix}$$

Finally, let

$$I_b^{(jc)} = I_b^{(j)} - I_{br}^{(j)}[I_r^{(j)}]^{-1} I_{rb}^{(j)}$$

Lemma 2 in the supplementary material of Susko (2013) gives that

$$2\Lambda_2 \approx [\boldsymbol{a}^{(1)}]^T I_b^{(1c)} \boldsymbol{a}^{(1)} - [\boldsymbol{a}^{(2)}]^T I_b^{(2c)} \boldsymbol{a}^{(2)} \quad (5)$$

where $\boldsymbol{a}^{(j)}$ is the minimizer of

$$(\mathbf{Z}_b^{(j)} - \boldsymbol{a})^T I_b^{(jc)} (\mathbf{Z}_b^{(j)} - \boldsymbol{a}) \quad (6)$$

subject to the constraint that $\boldsymbol{a} \geq \mathbf{0}$. Here, $\mathbf{Z}_b^{(j)}$ are the components of $n^{-1/2}[I^{(j)}(\theta_0)]^{-1}S^{(j)}(\theta_0)$ that correspond to the zero edge-lengths.

It is well known that, with regularity conditions, $\mathbf{S}^{(j)}(\theta_0)$ has mean 0 and covariance matrix $I^{(j)}(\theta_0)$ (cf. Proposition 3.4.4 of Bickel and Doksum 2007). As $\mathbf{S}^{(j)}(\theta_0)$ is proportional to a mean ($\mathbf{S}^{(j)}(\theta_0) = \sum_i \mathbf{S}^{(j)}(k_i; \theta_0)$, where $\mathbf{S}^{(j)}(k_i; \theta_0)$ is the contribution from the ith site), it follows from the Central Limit Theorem that $\mathbf{S}^{(1)}(\theta_0)^T$ and $\mathbf{S}^{(2)}(\theta_0)$ are jointly approximately multivariate normal with mean 0 and a joint covariance matrix $\Sigma^{(s)}$ that can be calculated and will be discussed in this study. As $\mathbf{Z}_b^{(1)}$ and $\mathbf{Z}_b^{(2)}$ are linear transformations of $\mathbf{S}^{(1)}(\theta_0)$ and $\mathbf{S}^{(2)}(\theta_0)$, it follows that they too are approximately multivariate normal with a joint covariance matrix, $\Sigma^{(z)}$, which can be calculated from $\Sigma^{(s)}$ using expressions for covariance matrices of linear transformations; full implementation details are given in supplementary material, Supplementary Material online.

In summary, the main result for simulation is that $2\Lambda_2$ has the approximate distribution of the right-hand side of equation (5) where $\boldsymbol{a}^{(j)}$ is the minimizer of (6) and $\mathbf{Z}_b^{(j)}$ is part of a vector having a multivariate normal distribution with mean 0 and covariance matrix, $\Sigma^{(z)}$. Thus, the distribution of $2\Lambda_2$ can be approximated with the following simulation strategy.

1) Repeatedly generate $\mathbf{Z}_b^{(1)*}$ and $\mathbf{Z}_b^{(2)*}$, normal random vectors with mean 0, and joint covariance matrix $\Sigma^{(z)}$.
2) For each $\mathbf{Z}_b^{(1)*}$ and $\mathbf{Z}_b^{(2)*}$, determine $\boldsymbol{a}^{(j)*}$ as the minimizer of $(\mathbf{Z}_b^{(j)*} - \boldsymbol{a})^T[I_b^{(jc)}](\mathbf{Z}_b^{(j)*} - \boldsymbol{a})$ with the constraint $\boldsymbol{a} \geq \mathbf{0}$.
3) Given the result from 2, calculate $2\Lambda_2^*$ from equation (5).

Given the Λ_2^* generated according to this scheme, probabilities involving Λ_2, like $P(\Lambda_2 > c)$, can be approximated from proportions of Λ_2^* satisfying the condition. As θ_0, the parameters generating the data, are unknown, estimates of these, obtained through ML with the null hypothesis tree, are plugged in to obtain $\Sigma^{(z)}$ and I_b^{jc}.

Returning to the calculation of $\Sigma^{(s)}$, the joint covariance matrix of $S^{(1)}(\theta_0)$ and $S^{(2)}(\theta_0)$, it is of the form

$$\Sigma^{(s)} = n\begin{bmatrix} I^{(1)}(\theta_0) & I^{(12)}(\theta_0) \\ I^{(12)}(\theta_0)^T & I^{(2)}(\theta_0) \end{bmatrix}$$

The covariance matrix $I^{(12)}(\theta_0)$ is calculated as

$$I^{(12)}(\theta_0) = \sum_k p_k S_k^{(1)}(\theta_0) S_k^{(2)}(\theta_0)^T \quad (7)$$

where the sum is over all patterns k and $S_k^{(j)}(\theta_0)$ is the contribution to $S^{(j)}(\theta_0)$ that arises from site pattern k if it is in the alignment. Similarly,

$$I^{(j)}(\theta_0) = \sum_k p_k S_k^{(j)}(\theta_0) S_k^{(j)}(\theta_0)^T \quad (8)$$

Because the number of possible patterns gets large with a large number of taxa, it may be necessary to approximate equations (7) and (8). This can be done by either replacing p_k with the observed frequency of pattern k in the original alignment or an alignment simulated under the null tree. In this case, the sum would be over at most as many site patterns as there are sites. The covariance matrix $\Sigma^{(z)}$ in step 1 can be calculated from equations (7) and (8). In practice, a further linear transformation is applied to the $\mathbf{Z}_b$ for simpler normal generation and to use the NNLS routine of Lawson and Hanson (1974) in step 2. Full details are given in the supplementary material, Supplementary Material online.

The main result leading to steps 1–3 makes some assumptions. Derivatives of any order need to be obtainable and expected values must exist. These conditions are satisfied for standard continuous-time Markov models. It is also assumed that the covariance matrix $\Sigma^{(s)}$ is invertible. This has

been true for all settings considered. As mentioned previously, identifiability of edge-lengths and other parameters is assumed and has been shown to hold for standard models.

Chi-Square Results

The final set of methods presented use the LR statistic $2\Lambda_3$, where Λ_3 is defined in equation (4), as a test statistic and chi-square results for threshold determination. They include extensions of the Ota et al. (2000) chi-bar test to cases where the null tree has two or three zero edge-lengths. In addition, the conservative LR statistic thresholds of the conditional test of Susko (2013) are discussed. These can be more easily implemented and apply for null trees with more zero edge-lengths. For the results to hold, an additional assumption is that, except for those edge-lengths set to 0 to make the trees equivalent, no other parameters are set to their boundary values.

The methods discussed differ primarily in the threshold used to declare significance or, equivalently, used for P value calculation. With the conditions discussed, the results of Shapiro (1985) imply that $2\Lambda_3$ has the approximate distribution of a mixture of chi-square random variables

$$P(2\Lambda_3 > y) = \sum_{j=0}^{p} w_j P(\chi_j^2 > y) \tag{9}$$

where χ_j^2 has a chi-square distribution with j degrees of freedom; χ_0^2 is 0 with probability 1. Here, p is the number of edge-lengths that were constrained to be 0 to make the two tree equivalent.

The weights of the mixture are positive and sum to 1. In general, they are difficult to calculate directly and depend on the true parameters in the generating model. As noted in Ota et al. (2000), however, when only one edge-length needs to be set to 0 to make the trees equivalent, $w_0 = w_1 = 1/2$. In the case that two edge-lengths need to be set to 0 to make the trees equivalent, the weights are

$$w_1 = 1/2,\ w_2 = \cos^{-1}\left\{I^c_{b12}/\sqrt{I^c_{b11}I^c_{b22}}\right\}/(2\pi),\ w_0 = 1/2 - w_2 \tag{10}$$

This follows roughly from case 7 of Self and Liang (1987) but adjusts for the presence of additional parameters beyond the two edge-lengths set to 0.

The weights for the case that three edge-lengths are set to 0 are more complicated. Let $A = [I^c_b]^{-1}$, $\rho_{ij} = A_{ij}/\sqrt{A_{ii}A_{jj}}$ and

$$\rho_{ij\cdot k} = (\rho_{ij} - \rho_{ik}\rho_{jk})(1-\rho_{ik}^2)^{-1/2}(1-\rho_{ik}^2)^{-1/2}.$$

Then,

$$\begin{aligned} w_3 &= [2\pi - \cos^{-1}(\rho_{12}) - \cos^{-1}(\rho_{13}) - \cos^{-1}(\rho_{23})]/(4\pi)\\ w_2 &= [3\pi - \cos^{-1}(\rho_{12\cdot 3}) - \cos^{-1}(\rho_{13\cdot 2}) - \cos^{-1}(\rho_{23\cdot 1})]/(4\pi)\\ w_1 &= 1/2 - w_3,\ w_0 = 1/2 - w_2 \end{aligned} \tag{11}$$

While I^c_b depends on unknown parameters, estimates of these can be used in practice. Derivations of equations (10) and (11) are given in the Appendix. When more than three edge-lengths need to be set to zero to make the trees equivalent, explicit formulas are no longer available.

The equation (9) is the correct one to use in calculating P values for the LR test. When the number of zero edge-lengths, P, required to make the two trees equivalent is larger than 3, however, explicit formulas for the w_j are no longer available. A conservative approach that will be referred to as the naive test treats $2\Lambda_3$ as if its distribution, under the null hypothesis, is chi-square with p degrees of freedom. It is naive in the sense that it is the usual LR test one would apply if unaware that edge-lengths being set to 0 under the null creates difficulties for ML theory. Because chi-square distributions with smaller degrees of freedom have smaller probabilities of exceeding a threshold, under the null hypothesis,

$$\begin{aligned} P(2\Lambda_3 > y) &= \sum_{j=0}^{p} w_j P(\chi_j^2 > y) \le \sum_{j=0}^{p} w_j P(\chi_p^2 > y)\\ &= P(\chi_p^2 > y) \end{aligned} \tag{12}$$

which implies that P values calculated using a chi-square distribution with p degrees of freedom will always be larger than the correct P value calculated using equation (9). As a consequence, the test will be conservative in that if one rejects whenever a P value is less than 0.05, the type I error probability will be less than 0.05.

Another conservative test, discussed in Susko (2013) and referred to as the conditional test, calculates P values using a chi-square distribution with degrees of freedom V, the number of edge-lengths that are 0 under the null hypothesis but estimated to be positive. As discussed in Susko (2013), under the null hypothesis, the conditional distribution of $2\Lambda_3$, given $V = v$, is chi-squared with v degrees of freedom. It follows that the type I error of a 0.05-level conditional test, given that $V > 0$, is 0.05. The test is conservative because there is some positive probability that none of the edge-lengths that are 0 under the null hypothesis will be estimated as positive, in which case $2\Lambda_3 = 0$ and any reasonable test will not reject. That is, the type I error is 0 when $V = 0$.

Using the KH Test Statistic with LR Statistic Thresholds

Both Λ_2 and Λ_3 are differences in log likelihoods between the maximized log likelihood for Tree 1 and a maximized log likelihood for Tree 2. As the Tree 2 log likelihood for Λ_3 is maximized subject to some edge-lengths being constrained to 0, it is never larger than the Tree 2 log likelihood for Λ_2, which is maximized without constraint. As Tree 2 log likelihoods are subtracted, it follows that $\Lambda_3 \le \Lambda_2$. As a consequence, if Λ_2 is used in place of Λ_3 in any of the LR statistic P value calculations, conservative or otherwise, a larger P value will be obtained than with Λ_3. Thus, such a substitution will give a conservative method. The reason such a replacement is worth considering is that all phylogenetic software implementations of ML allow ML estimation of edge-lengths, so that Λ_2 can be calculated, whereas few allow edge-lengths to be constrained to 0 as is required by Λ_3.

Results

Difficulties with the KH Test Motivation

The reason that the usual KH test need not give correct type I error probabilities using normality or RELL is that when the correct null hypothesis is considered, the argument for the approximate normal distribution of Λ_2 given in Kishino and Hasegawa (1989) no longer applies. What they argue is that the ML estimates will converge upon some parameters $\boldsymbol{\theta}_{j0}$, $j = 1,2$, and thus the log likelihoods with ML estimates should be approximately the same as when these parameters are substituted:

$$\sum_{i=1}^{n} \log p_{k_i}(\widehat{\boldsymbol{\theta}}_{j0}; j) = \sum_{i=1}^{n} \log p_{k_i}(\boldsymbol{\theta}_{j0}; j) + r_j(\boldsymbol{\theta}_{j0}) \qquad (13)$$

Because the remainder term $r_j(\boldsymbol{\theta}_{j0})$ is small relative to the first term, they argue that it can be ignored and indeed never consider its form. They point out that each $\log p_{k_i}(\boldsymbol{\theta}_{j0}; j)$ is independent and identically distributed. Thus by the Central Limit Theorem, the first term in equation (13) has an approximate normal distribution. They correctly conclude that the log likelihoods with ML parameters are approximately normal.

It may appear that the same argument can be applied to the differences in log likelihoods and Kishino and Hasegawa (1989) argue that it follows in the same way. However, the correct null hypothesis is a tree with edges collapsed to make the two trees equivalent. As ML estimation is consistent and the null tree is a version of both Trees 1 and 2, the parameters converged upon $\boldsymbol{\theta}_{j0}$, $j = 1,2$, will be the true generating parameters. Consequently, the log likelihoods will be the same for Trees 1 and 2 when $\boldsymbol{\theta}_{j0}$ is substituted. So,

$$\begin{aligned}\Lambda_2 &= \sum_{i=1}^{n} \log p_{k_i}(\boldsymbol{\theta}_{10}; 1) + r_1(\boldsymbol{\theta}_{10}) \\ &\quad - \sum_{i=1}^{n} \log p_{k_i}(\boldsymbol{\theta}_{20}; 2) + r_2(\boldsymbol{\theta}_{20}) \\ &= r_1(\boldsymbol{\theta}_{10}) - r_2(\boldsymbol{\theta}_{20})\end{aligned}$$

While the log likelihoods separately are larger than the remainder terms and are approximately normal, they cancel and the distribution ends up depending on the remainder terms, whose form was never considered.

The Mammalian Mitochondrial Data

The mammalian mitochondrial data considered previously in Goldman et al. (2000) and Shimodaira (2002) provide a useful illustrative example. For this data set with 6 taxa and 3,414 sites, a mtREV model (Adachi and Hasegawa 1996) was fit with gamma rates-across-sites variation and 8 rate categories (Yang 1994). The two trees to be tested and their consensus tree are in figure 2. The *P* values for significant evidence for Tree 1 over Tree 2 are given in table 1. The *P* value for the KH test was, up to the precision indicated, the same regardless of whether the normal distribution was used to calculate *P* values or RELL with 10,000 bootstrap samples. For the other bootstrap methods, 1,000 bootstrap samples were considered in each case. As Tree 1 was the ML tree, Goldman et al. (2000) considered it in their examples of the SOWH test, where they similarly reported a *P* value of 0.00.

Table 1. The *P* values for the Mammalian Mitochondrial Data.

Test	*P* value
KH	0.45
KHns	0.05
Parametric bootstrap, Tree 3, uncentered	0.05
Parametric bootstrap, Tree 3, centered	0.05
Parametric bootstrap, Tree 2, uncentered	0.00
SOWH	0.00
Parametric bootstrap, Tree 2, centered	0.49
Nonparametric bootstrap (centered)	0.45
Chi-bar	0.00
Chi-bar(Λ_2)	0.06
Naive	0.00
Naive(Λ_2)	0.12

NOTE.—Here, naive and chi-bar refer to likelihood ratio tests using different thresholds: a chi-square threshold with maximum degrees of freedom (naive), with degrees of freedom dependent on the number of estimated zero edge-lengths (cond) and with a mixture of chi-square threshold (chi-bar). The suffix Λ_2 indicates that the KH test statistic, Λ_2, was used in place of the constrained log LR, Λ_3.

The KHns test and parametric bootstrapping from Tree 3, with or without centering, give identical *P* values up to the precision indicated. For large sequence lengths, $E[d_i] \approx 0$ under Tree 3, so centered or uncentered bootstrapping should give comparable results. The parametric bootstrap uses the distribution of Λ_2 under Tree 3 with fitted parameters and the KHns test uses the large sample approximation to that distribution. Thus, substantial agreement between these approaches is expected. That all of these *P* values, which should have approximately correct type I error rates, are much smaller than the KH *P* value suggests the KH test may give conservative results.

The *P* values from the SOWH test is 0.00 which is smaller than the KHns *P* value and much smaller than the KH *P* value. This is surprising as the SOWH test can be viewed as a correction to the KH test when Tree 1 is not known in advance but is rather the ML tree. The likelihood of the ML tree is always at least as large as the likelihood of a fixed Tree 1. Thus, for the same null generating tree, the distribution of KH test statistics calculated with the ML tree in place of a fixed Tree 1 should be shifted to the right of the distribution of KH test statistics. Consequently, one expects larger *P* values from the SOWH test than a two-tree test with the same observed KH test statistic. The explanation of the discrepancy thus does not have to do with the SOWH test correcting for selection bias. It must be due to the SOWH test using a different generating tree in bootstrapping. Indeed, the small SOWH test *P* value is not inconsistent with the *P* value from uncentered parametric bootstrapping from Tree 2 which is similarly 0.00. By comparison, the *P* value from the parametric bootstrap with generating Tree 3 is 0.05 which is comparable to the KHns *P* value. The ML edge-lengths for Trees 2 and 3 were, on average, within 0.005 of each other except for the edge-length

that was constrained to 0. Without constraint it was estimated to be 0.02 which is the smallest edge-length in the tree but still not very small. The small SOWH *P* value is a consequence of the fact that one does not expect large log likelihood differences for one tree (ML or fixed) against Tree 2 when the generating Tree 2 is well resolved.

The nonparametric bootstrap and parametric bootstrap from Tree 2 with centering gave large values. The mean of these distributions is 0 by design, so the large values relative to parametric bootstrapping from Tree 3 with centering implies that these distributions showed extra variance over what is expected under the consensus tree null. The chi-bar and naive *P* values are small. By comparison, the *P* values from these tests with Λ_2 in place of Λ_3 are not small, indicating that the difference is due to the differing test statistics used.

Simulation Results

To further illustrate the properties of the tests under the null and alternative hypothesis, consider simulation from the six-taxa tree, Tree 1 in figure 1. Testing was of Tree 1 against Trees 2A–2C, leading to the null Trees 3A–3C. Accordingly, null simulations were from Trees 3A–3C with all nonzero edge-lengths set to 0.1. Similarly, simulations to study power properties were from Tree 1 with all of the edge-lengths that are nonzero under both null and alternative hypothesis set to 0.1. Depending on the setting, some or all of those edge-lengths that are 0 under the null hypothesis were positive in power simulations. All such positive edge-lengths were set to a common positive value which was allowed to vary as proportions of rejections are reported over a range of values. For each simulation setting, 1,000 simulated data sets of 1,000 sites were generated from the HKY model (Hasegawa et al. 1985) with κ parameter of 2 and frequencies of *A*, *C*, *G*, and *T* equal to 0.1, 0.2, 0.3, and 0.4.

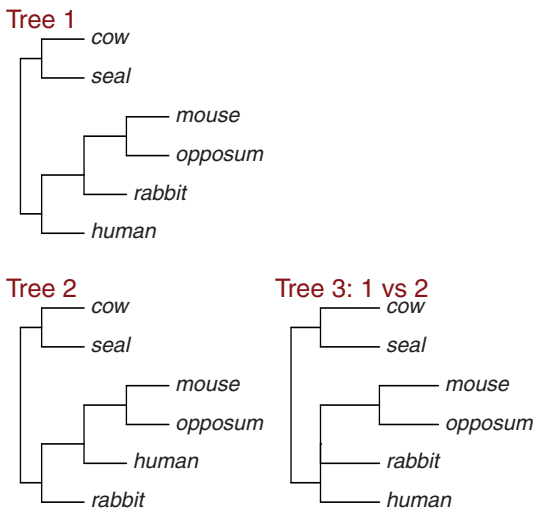

Fig. 2. Tree 1 gives the tree that significant evidence is being sought for in comparison to Tree 2 when using the mammalian mitochondrial data. Tree 3 gives the null tree that makes Trees 1 and 2 equivalent.

The results of simulations under the null hypothesis are given in table 2. The KH test has a far smaller than necessary false-positive rate, almost never rejecting. As expected by theory, using the naive or conditional test thresholds gives a smaller than expected false-positive rate although the difference is not substantial. The false-positive rates of KHns and the chi-bar test are close to the expected rate of 0.05.

Figure 3 gives the results of simulations to investigate power. Across conditions, the power of the KH test is substantially smaller than the other tests. The KHns test has performance almost identical to the chi-bar test, both of which are the best performers. As expected, given the similarities and differences between the conditional and naive tests, the conditional test always does better. In simulation settings where all of the edge-lengths that are 0 under the null were set to a positive value, the improvement is small, particularly as edge-lengths get larger. This makes sense as the number of estimated positive edge-lengths in such cases should tend to be close to the maximum degrees of freedom, so that the two tests frequently use the same degrees of freedom. Similarly, when only one of the edge-lengths was positive, the conditional test shows a substantial improvement over the naive test and is more comparable to the KHns test. In this case, the conditional test frequently uses only 1 degree of freedom where the naive test always uses 3.

Figure 4 gives the result for the conditional test using either the correct log LR $2\Lambda_3$ where edge-lengths are constrained to 0 (this was what was used in fig. 3) or $2\Lambda_2$ where the edge-lengths estimated under Tree 2 are unconstrained. The latter

Table 2. The Numbers of False Positives for 0.05-Level Tests over 1,000 Simulated Data Sets from Null Trees 3A–3C.

Tree	*t*	KHns	Chi-Bar	Chi-Bar(Λ_2)	Cond	Cond(Λ_2)	Naive	Naive(Λ_2)	KH
3A	0.02	30	26	25			16	15	0
	0.10	48	44	40			24	23	0
	0.50	36	32	32			13	13	1
3A	0.02	39	31	27	33	29	13	11	0
	0.10	50	39	31	35	28	18	15	0
	0.50	42	47	44	26	24	9	7	0
3C	0.02	37	27	21	44	32	6	6	0
	0.10	40	31	23	29	22	10	6	0
	0.50	51	45	30	33	20	12	8	1

NOTE.—The nonzero edge-lengths in the generating trees were each set to *t*. Here, cond refers to the conditional test and cond(Λ_2) to the conditional test using Λ_2 in place of Λ_3. For Tree 3A, the conditional test is the same as the naive test.

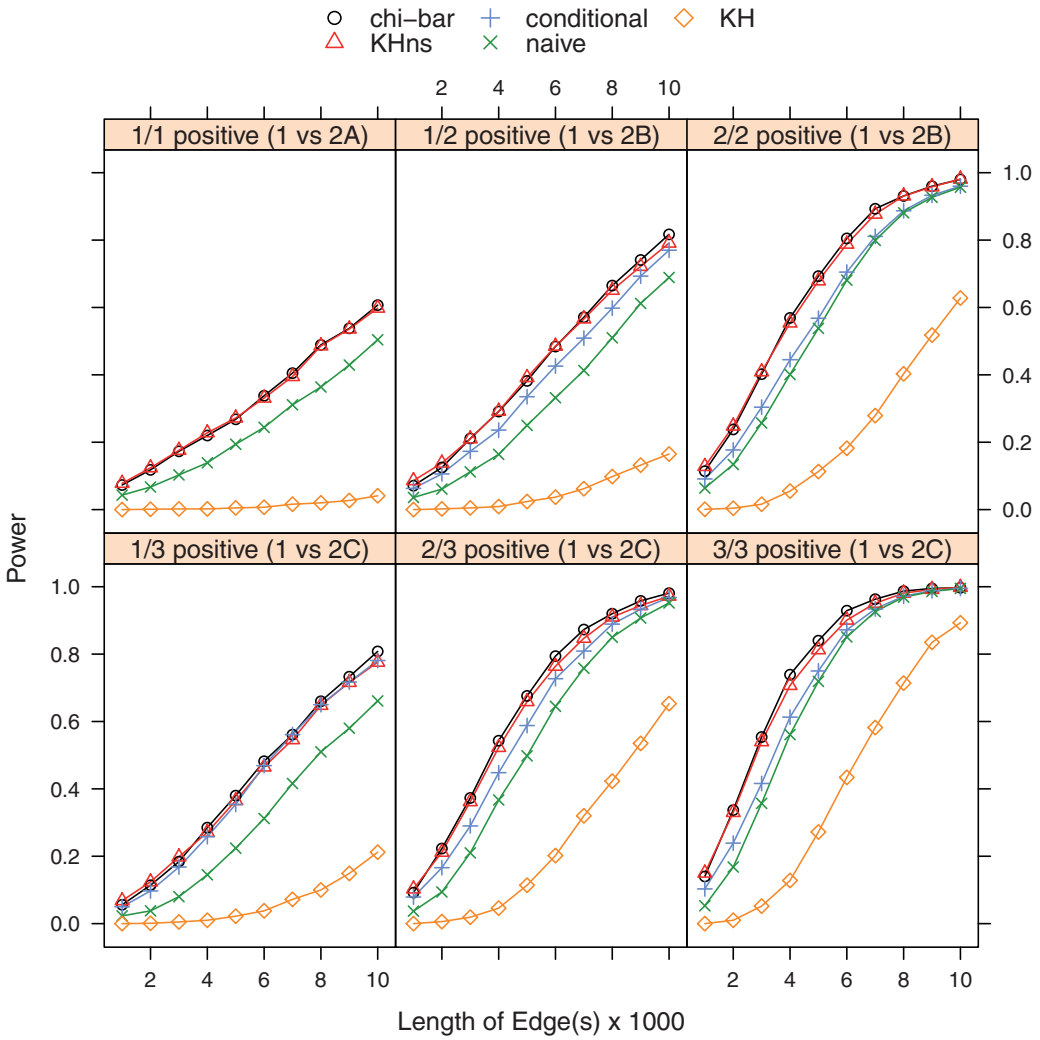

Fig. 3. Power of the tests of Tree 1 vs. 2A–C. Each panel and *x* axis value corresponds to a single simulation setting where the generating tree set some subset of the edges that were 0 in the null tree to the *x* axis value. The numbers of edges in the subset is indicated in the label. When only one edge is of zero length under the null, the conditional and naive tests are equivalent.

will always give lower power. It is of interest because most phylogenetic implementations of ML estimation do not allow estimation with zero edge-lengths. The conservative versions of conditional and naive tests can be obtained with such software by using $2\Lambda_2$ in place of $2\Lambda_3$. The loss in power is smallest when all of the unresolved edges under the null are positive in the generating tree. This makes some sense as when all the edge-lengths in Tree 1 are positive, those edges associated with splits in Tree 2 that are not present in Tree 1 will naturally be more likely to be estimated close to 0 and thus give likelihoods more comparable to those when a zero constraint is imposed.

Discussion

The chi-bar and KHns tests are expected to generally have approximately correct type I error probabilities, whereas the conditional and naive tests are expected to be conservative. Simulation results were consistent with this but suggested that the KH test will be much more conservative than the other tests. Because a number of these tests used the same KH test statistic and because the KH null is approximately the same as the consensus tree null, the main reason for the KH test difficulties must be that thresholds tend to be too large.

The chi-bar and KHns tests gave almost identical power in the simulations considered and better power than the other tests. These are results that can be expected to generalize. The chi-bar and KHns test can be expected to give similar performance as they are both based on good approximations to, respectively, the null distributions of $2\Lambda_3$ and $2\Lambda_2$, two similar test statistics. The other tests are based on the same test statistics but are known to be conservative with a consequent loss of power.

The KH null hypothesis is not exactly the same as the consensus tree null of the other tests. However, as a generating Tree 2 satisfying the exact KH null but differing from the consensus tree would not be compatible with any Tree 1, smaller KH test statistic values can be expected under such a null than under the consensus null. Thus, tests with approximately correct or conservative type I errors under the consensus null can reasonably be expected to be conservative

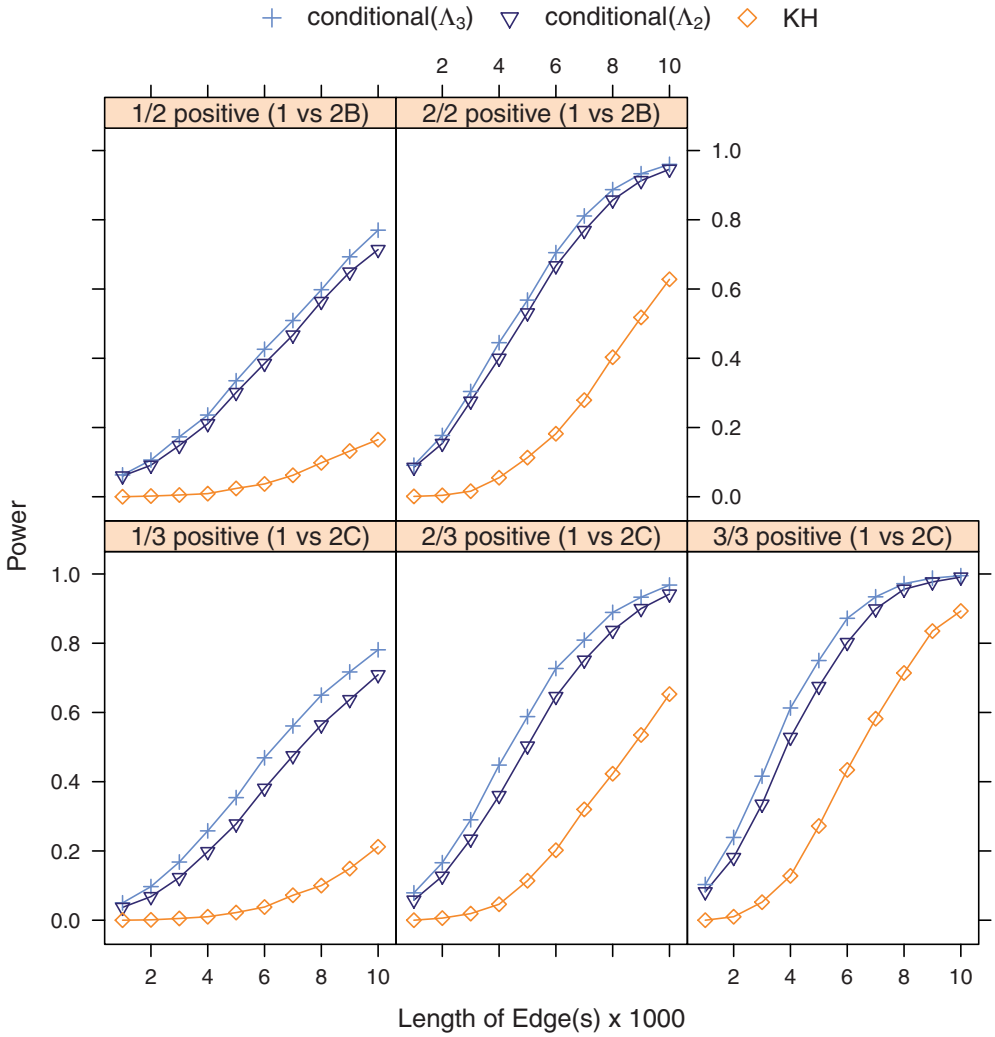

Fig. 4. Power for tests of Tree 1 vs. 2A–C. The settings are the same as for figure 3. Power for the conditional test is calculated when the log LR, Λ_3, is used and when the KH test statistic, Λ_2, is used. Power for the KH test is the same as in figure 3 and is included for comparison.

under an exact KH null. Consequently, the better power of all of the alternatives to the KH test is not likely a consequence of inflated type I error under an exact KH null.

The chi-bar test has the best performance and is simple to implement in the case that there is a single zero edge-length in the null tree. In the case that two or three edge-lengths are zero under the null, information matrix calculations are required similar to those of the KHns test. When more than three edge-lengths are set to zero under the null, however, the chi-bar test is no longer feasible.

The conditional test has the complication of determining whether an edge-length is 0 where most software implementations do not allow zero edge-lengths. A simple adjustment is to use a small threshold. In practice, 1.0e−9 was used, but experience looking at differences in log likelihoods when edge-lengths are close to 0 suggests the behavior of the test would be similar if edge-lengths less than 1.0e−6 were treated as 0. As the conditional test always does better than the naive test, there is little reason to use the latter. Serious discussion of it has been included here because one can envision some settings where it might prove useful. One example would be the reanalysis of literature results where likelihoods were reported for two trees and where the number of zero edge-lengths required to obtain the null tree could be determined from the graphical or Newick representation of the two trees, but where estimated edge-lengths could not be obtained without refitting the model.

A difficulty with KH implementations was noted in the course of this work. Because simulation was often from a poorly resolved Tree 1 or an unresolved Tree 3, there were a number of instances where the estimated edge-lengths for Trees 1 and 2 were such that they were either the same or almost the same trees. Indeed, when there was a single zero-length edge in a generating Tree 3, the ML Tree 1 and Tree 2 were identical a substantial proportion of the time, as predicted by theory. Such cases were recorded, appropriately, as failures to reject Tree 2. The site log likelihoods in this case are identical so that both $\bar{d} = 0$ and $s_d = 0$. For practical implementations of the KH test, this creates substantial difficulties. Due to round-off error or lower bounds on allowable edge-lengths, $\bar{d}$ and s_d will differ slightly from 0, making the normalized KH test statistic, $\sqrt{n}\bar{d}/s_d$, a numerically unstable value that can be large and positive, leading to an incorrect conclusion of rejection.

In simulations, the KH test performed much worse than the other tests. Its false-positive rate was much smaller than necessary, leading to a consequent loss of power. Relative to KHns, it is easy to implement. However, the conservative chi-square tests are as easy to implement. The much better performance of these tests suggests they are to be recommended in cases where KHns is difficult to implement.

Parametric and nonparametric boostrapping provide alternative but computationally more expensive approaches to testing. For the mammalian data set, parametric bootstrapping from a tree with zero-constrained edge-lengths gave larger P values than the SOWH test and parametric bootstrapping from Tree 2 with edge-lengths estimated without constraint. Bootstrapping from the zero-constrained tree gives a fairer chance to the null hypothesis and likely should be part of most SOWH implementations. In theory, however, if the consensus null tree is the generating tree, edge-lengths that are 0 in the consensus tree should be estimated as approximately 0 and bootstrapping from the fitted Tree 2 should give similar P values.

Nonparametric bootstrapping is expected to give rise to approximately correct type I error probabilities with large enough sequences. It is interesting to note that the RELL version of the KH test, which gives very similar results to the normal theory version, can be viewed as a time-saving approximation to nonparametric bootstrapping that skips the step of estimating parameters for each bootstrap sample. The difficulties of the KH test then suggest that difficulties can arise due to RELL.

One-sided tests have been considered throughout the article. This is consistent with implementations of the KH test in software like PAML and TREE-PUZZLE. It can sometimes be the case that interest is in a two-sided hypothesis: whether there is significant evidence for a well-resolved version of either Tree 1 or Tree 2. The KHns test extends naturally to this case. Rejection at the α-level occurs when the observed Λ_2 is less than the $\alpha/2$th quantile of the Λ_2^* or greater than the $(1-\alpha/2)$th quantile. Extension of the chi-bar test is more difficult but a simple, conservative approach is to apply a Bonferroni correction to the two separate tests: Tree 1 against Tree 2 and vice versa.

Software for the conditional, chi-bar, and KHns tests is available at http://www.mathstat.dal.ca/~tsusko (last accessed January 23, 2014) for many commonly used phylogenetic models. Although the conditional and naive tests cannot be expected to perform as well as KHns, they have the advantage of being relatively simple to implement. Software is also provided which computes the degrees of freedom given a tree file containing the estimated Tree 1 and a version of Tree 2. If $2\Lambda_2$ is used in place of $2\Lambda_3$, this software allows such tests to be applied to the output of any existing phylogenetics software.

Supplementary Material

Supplementary Material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The author thanks Eli Levy Karin, Tal Pupko, Andrew Roger, and two anonymous referees for helpful comments and discussion. This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix: The Weights of the Chi-Square Mixture

Theorem 2.1 of Shapiro (1985) gives as a special case that a mixture of chi-squares distribution applies to

$$(\boldsymbol{y}-\widehat{\boldsymbol{\eta}})^T I_b^c(\boldsymbol{y}-\widehat{\boldsymbol{\eta}}) \tag{14}$$

where $\widehat{\boldsymbol{\eta}}$ minimizes equation (14) over $\boldsymbol{\eta} \geq \boldsymbol{0}$ and q-dimensional $\boldsymbol{y}$ has a multivariate normal distribution with mean $\boldsymbol{0}$ and variance–covariance matrix $[I_b^c]^{-1}$.

As explained in the proof of Theorem 1 of Susko (2013), the limiting distribution of $2\Lambda_3$ is the same as that of $\widehat{\boldsymbol{\eta}}^T I_b^c \widehat{\boldsymbol{\eta}}$. It too has a mixture of chi-squares distribution but whenever the degrees of freedom of its distribution is ν, the degrees of freedom of the distribution of equation (14) is $q-\nu$. It follows that the weights of the mixture w_ν in equation (9) are the same as the weights $w_{\nu,q}(U)$ in equation (4.9) of Shapiro (1985). In the case that $q=3$ edges are set to 0 in the null tree, this directly gives equation (11). In the case that $q=2$ edges are set to 0, using the formula for the inverse of a 2×2 matrix and that $\cos^{-1}(x)=\pi-\cos^{-1}(-x)$, an expression is obtained that is more similar to that of Case 7 in Self and Liang (1987). Let $A=[I_b^c]^{-1}$ then

$$\begin{aligned} w_2 &= [\pi-\cos^{-1}(A_{12}/\sqrt{A_{11}A_{22}})]/(2\pi) \\ &= [\pi-\cos^{-1}(-I^c_{b12}/\sqrt{I^c_{b11}I^c_{b22}})]/(2\pi) \\ &= \cos^{-1}(I^c_{b12}/\sqrt{I^c_{b11}I^c_{b22}})/(2\pi). \end{aligned}$$

References

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

Allman ES, Ané C, Rhodes JA. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv Appl Prob.* 40:229–249.

Allman ES, Rhodes JA, Sullivant S. 2012. When do phylogenetic mixture models mimic other phylogenetic models. *Syst Biol.* 61: 1049–1059.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol.* 55: 539–552.

Bickel PJ, Doksum KJ. 2007. Mathematical Statistics: basic ideas and selected topics. Upper Saddle River (NJ): Prentice Hall.

Bickel PJ, Freedman DA. 1981. Some asymptotic theory for the bootstrap. *Ann Stat.* 9:1196–1217.

Billera LJ, Holmes SP, Vogtmann K. 2001. Geometry of the space of phylogenetic trees. *Adv Appl Math.* 27:733–767.

Chang JT. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math Biosci.* 137: 51–73.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.* 6:R41.

Efron B. 1982. The jacknife, the bootstrap and other resampling plans. CBMS-NF regional conference series in applied mathematics,

Vol. 38. Philadelphia (PA): Society for Industrial and Applied Mathematics.
Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol.* 49:652–670.
Hasegawa M, Kishino H. 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* 43:672–677.
Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22: 160–174.
Kishino H, Hasegawa M. 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, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 31: 151–160.
Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11:459–468.
Lawson CL, Hanson RJ. 1974. Solving least squares problems. Englewood Cliffs (NJ): Prentice-Hall.
Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol.* 17:798–803.
Pawitan Y. 2001. In all likelihood: statistical modeling and inference using likelihood. Oxford: Oxford University Press.
Schmidt HA, Strimmer K, Vingron M, von Haesler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc.* 82:605–610.
Shapiro A. 1985. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* 72: 133–144.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 5:492–508.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16:1114–1116.
Susko E. 2013. Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika* 100:1019–1023.
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland (MA): Sinauer Associates. p. 407–514.
White H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–25.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.
Yang Z. 1997. PAML: a program for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
Yang Z. 2007. PAML 4: a program for phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.