Syst. Biol. 49(4):652–670, 2000

Likelihood-Based Tests of Topologies in Phylogenetics

NICK GOLDMAN,[1] JON P. ANDERSON,[2] AND ALLEN G. RODRIGO[3]

[1]*University Museum of Zoology, Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK; E-mail: N.Goldman@zoo.cam.ac.uk*
[2]*Department of Molecular Biotechnology, University of Washington, Seattle, Washington USA*
[3]*School of Biological Sciences, University of Auckland, Auckland, New Zealand*

Abstract.—Likelihood-based statistical tests of competing evolutionary hypotheses (tree topologies) have been available for approximately a decade. By far the most commonly used is the Kishino–Hasegawa test. However, the assumptions that have to be made to ensure the validity of the Kishino–Hasegawa test place important restrictions on its applicability. In particular, it is only valid when the topologies being compared are specified a priori. Unfortunately, this means that the Kishino–Hasegawa test may be severely biased in many cases in which it is now commonly used: for example, in any case in which one of the competing topologies has been selected for testing because it is the maximum likelihood topology for the data set at hand. We review the theory of the Kishino–Hasegawa test and contend that for the majority of popular applications this test should not be used. Previously published results from invalid applications of the Kishino–Hasegawa test should be treated extremely cautiously, and future applications should use appropriate alternative tests instead. We review such alternative tests, both nonparametric and parametric, and give two examples which illustrate the importance of our contentions. [Kishino–Hasegawa test; maximum likelihood; phylogeny; Shimodaira–Hasegawa test; statistical tests; tree topology.]

Hasegawa and Kishino (1989) and Kishino and Hasegawa (1989) developed methods for estimating the standard error and confidence intervals for the difference in log-likelihoods between two topologically distinct phylogenetic trees representing hypotheses that might explain particular aligned sequence data sets. The method initially was introduced to compute confidence intervals on posterior probabilities for topologies in a Bayesian analysis (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1989). Attention quickly turned to using the same ideas to perform nonparametric likelihood ratio tests (LRTs) of the statistical significance of topologies (Kishino and Hasegawa, 1989), as currently implemented in the popular PHYLIP (Felsenstein, 1995), MOLPHY (Adachi and Hasegawa, 1996), PUZZLE (Strimmer and von Haeseler, 1996; Strimmer et al., 1997), and PAUP* (Swofford, 1998) software packages. Tests based on the ideas of Kishino and Hasegawa (hereafter referred to as KH-tests) could overcome some of the peculiar properties of phylogenetic estimation (see, e.g., Yang et al., 1995), for example, that different topologies do not share the same sets of parameters and do not generally represent nested hypotheses (one hypothesis being a special case of another).

Kishino and Hasegawa originally devised and applied their methods for trees that were specified a priori, that is, trees that corresponded to phylogenetic hypotheses derived independently of the data at hand. However, since then, the KH-test has in the majority of cases been used to compare the maximum likelihood (ML) tree derived from the data to one or more a priori–specified trees or to one or more a posteriori–specified trees (e.g., the trees with second-, third-, and so forth highest likelihoods). Such applications are facilitated, even encouraged, by various software packages. We contend that these latter applications of the KH-test are incorrect. We are not the first to note this: Swofford et al. (1996) observed that the KH-test should be applied only when the trees in question are specified a priori. Shimodaira and Hasegawa (1999) recently made the same point and described a correct nonparametric test for the case in which the ML tree is one of the tested topologies.

In this paper, we review the published literature on likelihood-based tests of topologies in phylogenetics. We wish to draw attention to the inapplicability of the KH-test for many common situations, despite the fact that it has become the most widely used and generally accepted way of testing alternative

hypotheses of evolutionary relationship. We first look in detail at the KH-test, in particular reviewing the conditions for its correct application and considering various methods that can be used to generate the null hypothesis distribution of the KH-test statistic. We then consider alternative applications of the KH-test, which represent the majority of published applications, and discuss why these are incorrect. We contend that the bias inherent in the KH-test when applied incorrectly warrants that the results of such applications be treated extremely cautiously. Next, we describe both the nonparametric test introduced by Shimodaira and Hasegawa (1999) and a parametric LRT of topologies, described briefly by Swofford et al. (1996) and based on the parametric bootstrap or Monte Carlo simulation approach to hypothesis testing using likelihood ratio statistics advocated by Goldman and others (e.g., Goldman, 1993; Hillis et al., 1996; Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997).

All of the methods described in this paper are equally relevant to ML analyses of DNA and amino acid (aa) data. We provide two analyses comparing the results of different tests, one using DNA sequences from HIV-1 isolates and one using aa sequences from mammalian mitochondrial (mt) proteins. These examples illustrate the importance of understanding precisely what hypotheses are being tested, the importance of performing statistically valid tests, and the differing power of parametric and nonparametric tests. We conclude with a discussion of the importance of our and others' recent work on the theory of tests of topologies.

The likelihood-based test developed from the work of Hasegawa and Kishino (1989) and Kishino and Hasegawa (1989) is quite closely related to a parsimony-based test described by Templeton (1983). Kishino and Hasegawa (1989 : 177) also discussed parsimony-based analogs of their likelihood methods. Such parsimony-based tests are subject to precisely the same misuses described in this paper for likelihood-based tests. In their unmodified forms, they too should not be used with phylogenies that are not selected a priori. Derivation of their corrected forms could follow the same general approaches described below (Buckley et al., unpubl.).

METHODS

Terminology and Notation

For the purposes of this paper, we are interested in the topologies of phylogenetic trees, not in the lengths of the branches on those trees. We use T_j to denote the topology of the jth prespecified tree, and T_{ML} for the topology of the ML tree for a given data set. Implicitly, because we focus on likelihood-based models, we assume that some model of evolution that allows us to define the probability of character-state changes can be specified before tree reconstruction. The vector of parameters for branch lengths (plus any other free parameters, e.g., in the model of nucleotide substitution being used) is written θ_x for topology T_x. H_0 and H_A are, respectively, null and alternative hypotheses for statistical testing. L_x is the log-likelihood of T_x or H_x for a given data set, generally maximized over all possible values of θ_x, and $L_x(k)$ is the sitewise log-likelihood for site k out of a total of S sites, so $L_x = \sum_{k=1}^{S} L_x(k)$. For parametric or nonparametric bootstrapped data, $L_x^{(i)}$ is the log-likelihood of T_x for the ith replicate data set, and $L_x^{(i)}(k)$ is the corresponding sitewise log-likelihood. We use δ to denote the difference in log-likelihoods between topologies ($\delta^{(i)}$ for the value from the ith replicate data set), $\mathrm{E}[X]$ to denote the expectation of a statistic X, and $\mathcal{N}(\mu\ \sigma^2)$ to indicate a normal (gaussian) distribution with mean μ and variance σ^2.

Fundamentals of the KH-Test

Suppose we have two hypotheses (tree topologies) T_1 and T_2, selected a priori, and we want to test whether they are equally well supported by a data set. Intuitively, for any one data set we would expect that stochasticity and sampling would ensure $L_1 \neq L_2$, even when the null hypothesis is true. However, if we were somehow able to obtain several data sets, we would expect that "on average" $L_1 = L_2$ when the null hypothesis is true. Writing $\delta \equiv L_1 - L_2$, this intuition corresponds to $\mathrm{E}[\delta] = 0$. In terms of a statistical test, our hypotheses are:

$$H_0: \quad E[\delta] = 0$$
$$H_A: \quad E[\delta] \neq 0.$$

In this section we describe a nonparametric approach to performing this comparison. This method was essentially given by Hasegawa and Kishino (1989), although they did not use the procedure for significance testing.

To perform a test of H_0 versus H_A, we need to know the distribution of δ under the null hypothesis. Working nonparametrically, we cannot derive the exact distribution of δ because H_0 does not specify a distribution for the data from which it is calculated. We are able, however, to implement the nonparametric bootstrap (Efron, 1982; Felsenstein 1985; Efron and Tibshirani, 1986), the use of which requires the assumption that the data are a representative and independent sample from the true distribution of data.

Table 1 summarizes all of the statistical tests discussed in this paper and explains the mnemonic naming system we have adopted. Table 2 relates this naming system to various published tests and software implementations (all described below). The procedure for the fundamental KH-test is now as follows:

Test *priNPfcd*:

- Calculate the test statistic $\delta \equiv L_1 - L_2$.
- Resample data (repeated nonparametric bootstrap data sets i).
- Reestimate any free parameters θ_1, θ_2 (branch lengths, and so forth) to get maximized log-likelihoods $L_1^{(i)}$ and $L_2^{(i)}$ under T_1 and T_2, respectively.
- Hence calculate bootstrap values of $\delta^{(i)} \equiv L_1^{(i)} - L_2^{(i)}$.
- For the values of $\delta^{(i)}$ to conform to H_0, we require $E[\delta^{(i)}] = 0$; to ensure this, replace the $\delta^{(i)}$ by $\tilde{\delta}^{(i)} \equiv \delta^{(i)} - \bar{\delta}^{(i)}$, where $\bar{\delta}^{(i)}$ is the mean over bootstrap replicates i of $\delta^{(i)}$. This procedure is known as "centering" (see below), and the resulting set of values $\tilde{\delta}^{(i)}$ gives an estimate of the distribution of δ under H_0.
- Test whether the attained value of δ (from the original data) is a plausible sample from the distribution of the $\tilde{\delta}^{(i)}$ by seeing if it falls within the confidence interval for $E[\delta]$, given, for example, by the 2.5% and 97.5% points of the ranked list of the $\tilde{\delta}^{(i)}$. A two-sided test is appropriate (because we have no a priori expectation as to whether T_1 or T_2 should be preferred); in this example, a 5% significance level is being used.

Hope (1968) and Marriott (1979) consider the amount of resampling that is needed for reliable statistical properties. It seems reasonable to hope that 100 data sets will give a sufficiently accurate estimate of the distribution of δ for testing at the 5% level, for both nonparametric (resampled data) and parametric (simulated data; see below) tests. More stringent tests will require more replicate data sets.

Hall and Wilson (1991) and Westfall and Young (1993 : 35) explain more fully the need for the centering procedure ($\tilde{\delta}^{(i)} \equiv \delta^{(i)} - \bar{\delta}^{(i)}$; Table 1, level 4, option *c*) to ensure conformity to the null hypothesis in this and other nonparametric tests. Comparing the observed value δ with the distribution of the $\delta^{(i)}$ (as is appropriate in parametric tests; see below and Table 1, level 4, option *u*) would give an invalid test. A test comparing the expected value of δ (i.e., 0) with the (uncentered) distribution of the $\delta^{(i)}$ is less inappropriate but is equivalent to comparing $\bar{\delta}^{(i)}$ with the distribution of the $\tilde{\delta}^{(i)}$. This in turn is equivalent to using $\bar{\delta}^{(i)}$ as an estimate of the test statistic δ, which is inefficient and without any redeeming advantage (see also Efron et al., 1996). We note that under the more powerful normal approximations discussed below (Table 1, level 5, options *a* and *s*) a test comparing 0 with the distribution of the $\delta^{(i)}$ becomes equivalent to the test that compares δ with the distribution of the $\tilde{\delta}^{(i)}$.

The first stages of this test (up to the calculation of the $\delta^{(i)}$) form the procedure at the heart the work of Hasegawa and Kishino (1989). In that paper, however, significance testing of phylogenies is not contemplated (there is no mention of the hypothesis $E[\delta] = 0$) and instead the estimated distribution of δ is used to compute confidence intervals on posterior probabilities of different (a priori) topologies in a Bayesian analysis. The idea of using the distribution of the $\delta^{(i)}$ to perform a significance test of phylogenies based on $E[\delta] = 0$ was introduced by Kishino and Hasegawa (1989), with some methods being partially described in Hasegawa et al. (1988). To our knowledge, the above form (*priNPfcd*) of the KH-test has never been implemented. At the time of this test's introduction, one

TABLE 1. Components of statistical tests described in this paper. Each test is composed of one option chosen at each of the five levels. Not all combinations are valid, as indicated. Mnemonic codes are derived by concatenating the italicized letters labeling each option; the derivation of these mnemonics is indicated by underlining. See text for further details.

Level 1: choice of trees to test	Level 2: statistical approach	Level 3: optimization method for bootstrapped data	Level 4: test statistic and distribution against which it is compared	Level 5: how test is performed or confidence intervals are generated
pri: trees chosen a priori	*NP*: nonparametric	*f*: full: all parameters estimated from data (and optimization over multiple topologies with *pos* option at level 1)	*c*: centered: attained δ vs. distribution of centered nonparametric estimates $\tilde{\delta}^{(i)} \equiv \delta^{(i)} - \hat{\delta}^{(i)}$ (only with *NP* option at level 2)	*d*: comparison of test statistic directly with its estimated distribution
pos: includes tree(s) selected a posteriori, from analysis of data to be used for testing	*P*: parametric	*p*: partial: some parameters fixed at values estimated from data	*u*: uncentered: attained δ vs. distribution of parametric estimates $\delta^{(i)}$ (only with *P* option at level 2)	*n*: assumption of normal distribution for test statistic (applicable only with *pri* option at level 1)
		n: no optimization for bootstrapped data (gives rise to RELL methods with *d* or *n* options at level 5)		*a*: additional normal assumption (variance of δ estimated from variance of sitewise $\delta(k)$; applicable only with both *pri* and *n* options at levels 1 and 3, respectively) *s*: stronger assumption of normal distribution for sitewise $\delta(k)$ (applicable only in combination with both *pri* and *n* options at levels 1 and 3, respectively)

TABLE 2. Relationships of statistical tests previously described and popular software implementations with tests described in this paper, with additional notes.

In literature/computer implementation	In notation of this paper	Notes
Kishino–Hasegawa test, fundamental concept	*priNPfcd*	Nonparametric test; never before published/implemented in this form[a]
Hasegawa and Kishino (1989)	*priNPf•*[b]	Distribution of δ derived in a different context; no statistical test specified
Kishino and Hasegawa (1989)	*priNPn•*[b]	RELL and normal approximations introduced; statistical tests only briefly discussed
PHYLIP (Felsenstein, 1995) and PUZZLE (Strimmer and von Haeseler, 1996; Strimmer et al., 1997) implementations	*priNPnca*	Use additional normal assumption and perform (two-sided) z-test
MOLPHY implementation (Adachi and Hasegawa, 1996)	*priNPnca*	Uses additional normal assumption; estimates variance of δ but performs no test
PAUP* implementation (Swofford, 1998)	*priNPncs*	Uses stronger normal assumption and performs (two-sided) paired t-test
Shimodaira and Hasegawa (1999) implementation of Kishino–Hasegawa test	*priNPncd*	Used in an example (with RELL and a one-sided test), for comparison with *posNPncd*
Shimodaira–Hasegawa test, fundamental concept	*posNPfcd*	Nonparametric test; never before implemented in this form[a]
Shimodaira and Hasegawa (1999)	*posNPncd*	Uses RELL; test described and used in an example
SOWH-test, fundamental concept	*posPfud*	Parametric test; originally described by Swofford et al. (1996)
SOWH-test, alternative implementation in this paper	*posPpud*	Used in an example (some approximations under H_A); see text for details

[a] These tests, amongst others, are implemented in this paper for the HIV-1 data set.

[b] Dots represent components of testing procedures that were not specified in the corresponding publications.

main reason for this was probably the computation time required for the step in which free parameters θ_1, θ_2 are reestimated for each bootstrap data set. (Nowadays this computational demand would not be problematic but this form of the test is still not used, probably because interest is typically in phylogenies, at least one of which has been chosen a posteriori—see below.) Accordingly, methods were devised to reduce the computational burden.

The KH-Test: Time-Saving Approximations

The fundamental KH-test (*priNPfcd* above) has the disadvantage that likelihood maximization is performed for each bootstrap replicate data set *i*. Although no maximization over topologies is required, because in this case only two a priori–specified topologies are being considered, Kishino and Hasegawa (1989) were concerned about the computation time needed. To reduce the computational burden, they developed a resampling estimated log-likelihood (RELL) technique (Table 1, level 3, option *n*; see also Kishino et al., 1990). In brief, they showed that instead of performing time-consuming likelihood optimizations for each bootstrap data set, one could use values of $\delta^{(i)}$ calculated with the optimized parameter values $(\hat{\theta}_1, \hat{\theta}_2)$ from the original data set. Certain asymptotic conditions (correctly specified evolutionary models and sufficiently large amounts of data) are required for this approximation to be valid; the RELL method has been shown to perform well in the estimation of bootstrap probabilities of phylogenies (Hasegawa and Kishino, 1994; see also below for discussion of the possible effects of approximations to log-likelihood scores in LRTs). The necessary likelihood calculations now require no optimization after the initial analysis of the original data, which saves a large amount of computational effort. By using a prime symbol ($'$) to denote this form of approximation where parameters are not re-optimized for replicate data sets, the resulting test can be described as follows:

Test *priNPncd*:
- Calculate the test statistic $\delta \equiv L_1 - L_2$.
- Resample data (repeated bootstrap data sets *i*).
- Using the ML estimates of any free parameters $\hat{\theta}_1, \hat{\theta}_2$ (branch lengths and so forth) derived from the original data set, compute log-likelihoods $L_1'^{(i)}$ and $L_2'^{(i)}$ under T_1 and T_2, respectively (the resampling is effectively being made from the sitewise log-likelihoods $L_1(k)$ and $L_2(k)$ estimated under H_0; hence the RELL mnemonic).
- Calculate bootstrap values of $\delta'^{(i)} \equiv L_1'^{(i)} - L_2'^{(i)}$.
- As before, perform the centering procedure $\tilde{\delta}'^{(i)} \equiv \delta'^{(i)} - \delta'^{(i)}$.
- Test whether the attained value of δ (from the original data) is a plausible sample from the distribution of the $\tilde{\delta}'^{(i)}$ by seeing if it falls within the confidence interval for $E[\delta]$ given by appropriate points of the ranked list of the $\tilde{\delta}'^{(i)}$ (two-sided test).

Kishino and Hasegawa (1989) further showed that the difference (δ) in log-likelihoods between two topologies specified a priori would follow a normal distribution; the mean and variance of which could be specified in terms of the differences in log-likelihoods ($\delta^{(i)}$) calculated for nonparametric bootstrap data sets *i*. This approximation, based on the Central Limit Theorem, requires the same asymptotic conditions as the RELL method for its validity. An alternative to using the direct comparison of the attained δ with the distribution of the $\tilde{\delta}'^{(i)}$, as in the last step of test *priNPncd*, is then to utilize this normal approximation for δ (Table 1, level 5, option *n*):

Test *priNPncn*:
- Proceed as in test *priNPncd* above, but replace the final step with the following:
- Compute the variance of the $\tilde{\delta}'^{(i)}$ (denote this by v^2) and test whether the attained value of δ (from the original data) is a plausible sample from a $\mathcal{N}(0, v^2)$ distribution, by seeing if it falls within the confidence interval for $E[\delta]$ given, for example, by $0 \pm 1.96v$ (two-sided test; 5% significance level in this example).

In practice, these tests have rarely been implemented, to our knowledge. More usually, an additional assumption is also made: that the variance of δ can be estimated from the variance (over sites $k = 1, 2, \ldots, S$) of the sitewise log-likelihood differences $\delta(k) \equiv$

$L_1^{(k)} - L_2^{(k)}$ (Table 1, level 5, option *a*; Kishino and Hasegawa, 1989). In this case a test can be made without any resampling, thus giving an even greater saving in time:

Test *priNPnca*:

- Calculate the test statistic $\delta \equiv (L_1 - L_2)$.
- Using the ML estimates of any free parameters $\hat{\theta}_1, \hat{\theta}_2$ (branch lengths, and so forth) derived from the original data set, compute sitewise log-likelihoods $L_1(k)$ and $L_2(k)$, under T_1 and T_2, respectively, for the sites k of the original data set.
- Calculate the values $\delta(k) \equiv L_1(k) - L_2(k)$ and hence the centered values $\tilde{\delta}(k) \equiv \delta(k) - \bar{\delta}(k)$ and an estimate of their variance $v^2 = \sum_k (\tilde{\delta}(k))^2/(S-1)$ (clearly, the variances of the $\delta(k)$ and the $\tilde{\delta}(k)$ are identical). Because $\delta \equiv \sum_k \tilde{\delta}(k) + S\bar{\delta}(k)$, and $\bar{\delta} = 0$ under H_0, Sv^2 is an estimate of the variance of δ.
- Test whether the attained value of δ calculated from the original data is a plausible sample from a $\mathcal{N}(0, Sv^2)$ distribution, for example, by comparing it with the confidence interval $0 \pm 1.96\sqrt{S}v$ (two-sided test; 5% significance level in this example).

This is the method implemented in various programs of the PHYLIP (Felsenstein, 1995), MOLPHY (Adachi and Hasegawa, 1996), and PUZZLE (Strimmer and von Haeseler, 1996; Strimmer et al., 1997) packages. (Programs in MOLPHY compute $\sqrt{S}v$ but leave statistical testing to the user).

In Swofford's (1998) PAUP* program, a stronger assumption is made (D. Swofford, pers. comm.): that the sitewise log-likelihood differences $\delta(k)$ are themselves normally distributed (Table 1, level 5, option *s*). This assumption, if accurate, guarantees the accuracy of the normal approximations described above (Table 1, level 5, options *n* and *a*) and permits a different test to be performed directly on the sitewise log-likelihoods:

Test *priNPncs*:

- Proceed as in test *priNPnca* above, but replace the final two steps with the following:
- Perform a paired-*t*-test of the $L_1(k)$ and $L_2(k)$ (pairs $\{L_1(1), L_2(1)\}$, $\{L_1(2), L_2(2)\}, \ldots, \{L_1(S), L_2(S)\}$) to determine if the means of the $\{L_1(k)\}$ and $\{L_2(k)\}$ are equal (two-sided test).

We know of no theoretical justification for this additional assumption, and it does not give rise to any significant saving in computation time. However, we would expect that it will only make the smallest of differences in real applications (with a large number of sites S), and the significance levels reported by PAUP* and by the DNAML program of PHYLIP are invariably very similar (J. Felsenstein, pers. comm.; D. Swofford, pers. comm.; J.P. Anderson, unpubl.). Note that both the *priNPnca* and *priNPncs* tests require no resampling and so are even faster than the *priNPncd* and *priNPncn* tests, which use RELL methods.

Incorrect Usage of the KH-Test

Many of the arguments in the derivations of the statistical tests above are strongly dependent on the topologies T_1 and T_2 having been selected independently of any analysis of the data used for the testing. In particular, this assumption is necessary to justify the fundamental hypothesis H_0: $E[\delta] = 0$. Unfortunately, when the selection of topologies has been made with reference to the data, especially if they have been selected by a criterion linked to their likelihood scores, this expectation is no longer justified. Two particular cases in which it is not reasonable to expect $E[\delta] = 0$ are (1) the comparison of the tree found to have the maximal likelihood, T_{ML}, with an a priori tree, T_1, and (2) the comparison of T_{ML} with a tree selected for having the second- (or third-, and so forth) highest likelihood. In fact, in both of these cases $E[\delta] > 0$. T_{ML} is selected exactly because its likelihood for the data at hand is greater than that of any other tree: In other words, it is guaranteed that L_{ML} will be at least as large as L_1 or the log-likelihood of any other topology.

This is not a minor discrepancy: no result suggests that $E[\delta]$ will be even near to 0 in these cases. Further, given that necessarily $E[\delta] > 0$, two-sided tests are no longer appropriate: we are interested in assessing deviations only in one "direction" from expectation, and one-sided tests are required. In our experience, however, situations such as those just described in which trees are selected a posteriori are precisely those for

which the KH-test has most often been used. Indeed, even Hasegawa and coworkers appear to omit consideration of whether $E[\delta] = 0$ is a valid assumption (Hasegawa et al., 1988:8; Kishino and Hasegawa, 1989:175), even though Kishino and Hasegawa (1989 : 177, equation 20 and following) recognize that $E[\delta]$ depends on how the topologies being assessed were chosen.

We believe that the results of all such analyses using the KH-test are invalid and require recomputation by methods such as those described later in this paper. We stress that this is not a minor, obscure, or purely hypothetical point, of interest only to theoretical statisticians. Although the KH-test is suitable for the questions it was designed to answer, it is entirely inappropriate for use in testing the significance of trees selected by ML from the data which are to be used for the testing. We can find only one possible adjustment applicable to the results of incorrect applications of KH-tests that may render them useful; this is discussed below.

To help illustrate why we can no longer base tests on the hypothesis $E[\delta] = 0$, we use an analogy of running races. The coach of a running squad is interested to know whether two runners, Matthew and Mark, differ significantly in their average running times for the 100 m sprint. To determine this, he times the two runners when they participate in several races. For each race, the coach calculates the difference in running times t between Matthew and Mark: δ(Matthew, Mark) $\equiv t$(Matthew) $- t$(Mark). Note that δ(Matthew, Mark) can sometimes be positive and sometimes be negative, depending on who runs faster in any given race; in fact, if Matthew and Mark are equally good at the 100 m sprint, then the average value of δ(Matthew, Mark) over many races will tend to 0. In fact, as the team statistician approvingly explains, the data the coach has collected can be used to estimate the variance of δ(Matthew, Mark) and, consequently, to test the following hypothesis:

$$H_0: \quad E[\delta(\text{Matthew, Mark})] = 0$$

$$H_A: \quad E[\delta(\text{Matthew, Mark})] \neq 0.$$

This is analogous to the KH-test for phylogenies, with Matthew and Mark corresponding to two a priori topologies T_1 and T_2, each race equivalent to a sample of data, t(Matthew) and t(Mark) equivalent to the log-likelihoods L_1 and L_2, respectively, and δ(Matthew, Mark) corresponding to the difference in log-likelihoods δ.

The coach is also interested in another runner, Luke, whom he believes to be the fastest runner on his squad. Given his success at collecting data for the earlier hypothesis test, the coach decides to do something similar with Luke. He obtains running times for Luke, t(Luke), over several races amongst the squad members, as well as the fastest time for each race, t(fastest). For each race he computes the difference between these times, δ(Luke, fastest) $\equiv t$(Luke) $- t$(fastest), arguing that if Luke truly is the fastest then, over many races, the average of δ(Luke, fastest) will be zero. However, as the team statistician points out, this assumption is necessarily false. The reason for this is simple. If Luke truly is the fastest, then we may expect that in the majority of the races he participates in, his time is the fastest time, i.e., t(Luke) = t(fastest) and δ(Luke, fastest) = 0. However, we also expect that some other squad members will manage to win some races, if only infrequently, so that δ(Luke, fastest) >0. Note that it is never possible that δ(Luke, fastest) <0, and consequently the average of δ(Luke, fastest) over several races (the majority with δ(Luke, fastest) = 0 and some with δ(Luke, fastest) >0) must necessarily be >0. Variations in the squad members' performances in different races ensures that even if none is systematically faster than Luke, there is a chance that someone of equal or lower ability will appear to outperform Luke in any one race. In fact, the bigger the squad, the more likely this is, and the greater the level of outperformance one can expect. The statistical test used should reflect this fact and cannot be based on $E[\delta(\text{Luke, fastest})] \equiv E[t(\text{Luke}) - t(\text{fastest})] = 0$.

This example is analogous to the common but incorrect application of the KH-test, when T_{ML} and T_1 are identified with the faster runner and Luke, respectively, L_{ML} and L_1 with t(fastest) and t(Luke); and δ with δ(Luke, fastest). It would be possible, but of little interest here, to devise a correct test based on runners selected a posteriori. Instead, we will revert to discussing phylogenetic examples and will describe two tests that can be used in place of the KH-test when it is not applicable.

The Shimodaira–Hasegawa Test: A Corrected Nonparametric Test of Topologies

Although it has been noted in the past that the KH-test is suitable only for cases where E[δ] = 0 (Swofford et al., 1996), it appears that Shimodaira and Hasegawa (1999) are the first to publish a full explanation of why the test is not appropriate when one or more topologies under test were selected with reference to the same data being used for testing. Our arguments in the preceding section are essentially an extended version of those of Shimodaira and Hasegawa (1999). Based on earlier work by Shimodaira (1993, 1998), Shimodaira and Hasegawa (1999) have proposed a nonparametric test similar to the KH-test but making the appropriate allowance for the method by which topologies are usually selected for statistical comparison. The Shimodaira–Hasegawa test (SH-test) simultaneously compares all topologies in some set $\mathcal{M}$ and makes appropriate allowance for these multiple comparisons. It is necessary that $\mathcal{M}$ contains every topology that can possibly be entertained as the true topology, to ensure that the true topology is always "available" to be the ML topology for any bootstrap data set; if this condition is not met, the significance levels computed will be inaccurate (Westfall and Young, 1993:48). In addition, selection of topologies for the set $\mathcal{M}$ should be made a priori and not with reference to the observed data; otherwise, significance levels will again be inaccurate. Choosing $\mathcal{M}$ to be the set of all possible topologies is always safe, if conservative.

We cannot write the null hypothesis as E[δ] = 0 for this test; instead, the hypotheses tested are as follows:

H_0: all $T_x \in \mathcal{M}$ (including T_{ML}, the ML tree) are equally good explanations of the data

H_A: some or all $T_x \in \mathcal{M}$ are not equally good explanations of the data

and the test may proceed as follows:

Test *posNPfcd:*

- Calculate a test statistic δ_x for each topology $T_x \in \mathcal{M}$: δ_x is the attained value of $L_{ML} - L_x$.
- Generate nonparametric bootstrap replicate data sets i and for each one maximize likelihoods over parameters θ_x for each permitted topology T_x, giving optimal log-likelihood values $L_x^{(i)}$.
- For each topology T_x, form the adjusted log-likelihood $\tilde{L}_x^{(i)} \equiv L_x^{(i)} - \bar{L}_x^{(i)}$ by subtracting $\bar{L}_x^{(i)}$, the mean over replicates i of $L_x^{(i)}$, from each value of $L_x^{(i)}$—this is the centering method devised by Shimodaira (1998), which is appropriate for enforcing that the resampled data conform to H_0 for this a posteriori test.
- For each replicate i, find $\tilde{L}_{ML}^{(i)}$, the maximum over topologies T_x of the adjusted log-likelihoods $\tilde{L}_x^{(i)}$, and form bootstrap replicate statistics $\delta_x^{(i)} \equiv \tilde{L}_{ML}^{(i)} - \tilde{L}_x^{(i)}$; this allows for the a posteriori selection of T_{ML}.
- For each topology T_x, test whether the attained δ_x is a plausible sample from the distribution (over replicates i) of the $\delta_x^{(i)}$ by seeing if it falls within the confidence interval for E[δ_x] given, for example, by the interval between 0 and the 95% point of the ranked list of the $\delta_x^{(i)}$. Such a one-sided test is appropriate, because we know that only $\tilde{L}_{ML}^{(i)} \geq \tilde{L}_x^{(i)}$ is possible; in this example, a 5% significance level is being used.

(We have used some notation different from that of Shimodaira and Hasegawa [1999], to maintain a consistent style within this paper. Shimodaira and Hasegawa [1999] use $T_\alpha\, \alpha = 1, 2, \ldots, M$; $\tilde{L}_{\alpha \cdot i}$; $\tilde{R}_{\alpha \cdot i}$; and $\tilde{S}_{\alpha \cdot i}$, where we have used $\delta_x T_x \in \mathcal{M}$; $L_x^{(i)}$; $\tilde{L}_x^{(i)}$; and $\delta_x^{(i)}$, respectively.)

Time-saving approximations are also possible with this test. Shimodaira and Hasegawa (1999) propose the use of the RELL method for finding approximate values of $L_x^{(i)}$ without having to re-optimize θ_x for each replicate data set. This test, implemented by Shimodaira and Hasegawa (1999) and Buckley et al. (unpubl.), can be described as follows:

Test *posNPncd*:

- Calculate a test statistic δ_x for each topology $T_x \in \mathcal{M}$: δ_x is the attained value of $L_{ML} - L_x$.
- Generate nonparametric bootstrap replicate data sets i; for each one, and for each tree T_x, use the ML estimates $\hat{\theta}_x$ of any free parameters derived for each tree

T_x from the original data set to compute log-likelihoods $L_x^{\prime(i)}$, which approximate the optimized values $L_x^{(i)}$ in test *posNPfcd* above.

- For each topology T_x, form the adjusted log-likelihood $\tilde{L}_x^{\prime(i)} \equiv L_x^{\prime(i)} - \bar{L}_x^{\prime(i)}$ (centering).
- For each replicate i, find $\tilde{L}_{\mathrm{ML}}^{\prime(i)}$, the maximum over topologies T_x of the adjusted log-likelihoods $\tilde{L}_x^{\prime(i)}$, and form bootstrap replicate statistics $\delta_x^{\prime(i)} \equiv \tilde{L}_{\mathrm{ML}}^{\prime(i)} - \tilde{L}_x^{\prime(i)}$.
- For each topology T_x, test whether the attained δ_x is a plausible sample from the distribution (over replicates i) of the $\delta_x^{\prime(i)}$ by seeing if it falls within the confidence interval for $\mathrm{E}[\delta_x]$ given, for example, by 0 and the 95% point of the ranked list of the $\delta_x^{\prime(i)}$ (one-sided test; 5% significance level used in this example).

Note that the SH-test simultaneously assesses the significance level for each of the topologies $T_x \in \mathcal{M}$. It immediately reduces to a version of the KH-test, modified for the comparison of a priori–selected topology T_1 and a posteriori–selected topology T_{ML}, when attention is restricted to the significance level computed for δ_1 from the distribution of the $\delta_1^{(i)}$ or $\delta_1^{\prime(i)}$. Note, however, that the set $\mathcal{M}$ of all plausible topologies still has to be considered to compute this distribution.

The effect of the new centering procedure introduced in this method is to decrease the significance accorded to the difference δ_x in log-likelihoods between each topology T_x and the ML topology T_{ML} (Shimodaira and Hasegawa, 1999), in comparison with the significance indicated by the corresponding but inappropriate KH-test. Intuitively, this is because the attained value of δ_x should be attributed to two components: one (necessarily positive) being a consequence of the selection of T_{ML} precisely because it has the highest likelihood, and another (of unknown sign) attributable to the difference in the abilities of T_x and T_{ML} to explain the observed data. Whereas the SH-test correctly compares T_x and T_{ML} on the basis of the second component alone, making an appropriate allowance for the first component, the incorrectly applied KH-test assesses both components combined as though they were only the second component. The fact that the first component is necessarily > 0 acts to make the new test more conservative (i.e., less likely to reject the null hypothesis). However, the SH-test correctly uses a one-sided test, and this acts to increase the significance of results.

Is It Possible to Salvage the Results of Incorrect Applications of the KH-Test?

We can find only one possible adjustment that might render some previously published results from incorrect applications of the KH-test useful in the light of the theoretical advances described in this paper. It is straightforward to convert the significance level of a two-sided test to that of a one-sided test: the *P*-value should simply be halved. If the *P*-value obtained from an incorrectly applied KH-test is p, then the *P*-value that would be obtained in the SH-test is necessarily $\geq p/2$. Therefore, if the adjusted value $p/2$ is large enough to indicate no rejection of the null hypothesis (a priori tree T_1), e.g., $p/2 > 0.05$ for a 5% significance level, we can be certain that using the SH-test would give the same conclusion.

However, in all other cases (where the adjusted value $p/2$ is sufficiently small to indicate rejection of the null hypothesis in favor of the ML tree, e.g., $p/2 < 0.05$ for a 5% significance level) we cannot assume that this result would hold under a SH-test, which must give a *P*-value that would exceed $p/2$ by an unknown amount. Note that this will necessarily be the case whenever p indicated rejection of the null hypothesis in the incorrectly applied KH-test (e.g., $p < 0.05$ implies $p/2 < 0.025$), and in some instances when p did not indicate rejection of the null hypothesis (e.g., $0.05 < p < 0.1$ implies $0.025 < p/2 < 0.05$).

In summary, if the *P*-value obtained from an incorrectly applied KH-test is greater than twice the value required to indicate no rejection of the null hypothesis, then that conclusion would hold under the SH-test. If the *P*-value is less than this, we cannot determine from the KH-test result what the result would be for any test making proper allowance for a posteriori selection of hypotheses, and a full reanalysis of the original data by appropriate tests is necessary.

The SOWH-Test: a Correct Parametric Test of Topologies

Parametric statistical testing of hypotheses is becoming increasingly popular in

phylogenetics, based on the same models of sequence evolution being used for making phylogenetic inferences. These tests are generally based on parametric bootstrapping techniques (also known as Monte Carlo simulation), the theory of which is described in more detail by Goldman (1993), Huelsenbeck and Crandall (1997), and Huelsenbeck and Rannala (1997). Whereas in nonparametric bootstrap methods, replicate data sets are created by resampling with replacement from the original data set, in parametric bootstrapping, replicate data sets that conform precisely to the assumptions of a parametric null hypothesis are created by simulating data through the use of those assumptions in conjunction with parameter values estimated under the null hypothesis from the original data set (Goldman, 1993). Subsequent analysis of these data sets by the same methods as used for the original data gives replicate values of any required statistic. These are guaranteed to be drawn from the distribution induced by the null hypothesis, and their distribution therefore is a parametric estimate of the null hypothesis distribution of that statistic.

Parametric tests offer the possibility of more power than nonparametric tests, because the knowledge they can use of the form of the distribution giving rise to the data is unavailable to nonparametric tests. The cost of this power is an increased reliance on the models they assume (which could be inaccurate and thus lead to misinterpretation of the data). Typically, LRTs in phylogenetics (including those performed by using the parametric bootstrap) have been found to be powerful and are quite robust to deviations from the assumed model so long as a reasonable effort is made to use a model sufficiently complex to encompass the major features of the distribution of the data (e.g., Goldman, 1993; Yang et al., 1994, 1995; Hillis et al., 1996; Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Cunningham et al., 1998; Zhang, 1999).

It is possible to create a parametric bootstrap LRT to assess whether an a priori selected topology T_1 is supported by a sequence data set or should be rejected in favor of another topology. The following test is a straightforward application of the parametric bootstrap, yet it appears to be little-known or -used. The only previous (and brief) description we know of is due to Swofford et al. (1996), and the only published application is by Hillis et al. (1996). We refer to it as the SOWH-test, after the authors who originally described it. The hypotheses compared are these:

H_0: T_1 is the true topology

H_A: some other topology is true

and the parametric bootstrap statistical test of these hypothesis is as follows:

Test *posPfud*:

- Calculate the test statistic $\delta \equiv L_{ML} - L_1$.
- Simulate data sets i by parametric bootstrapping, based on the null hypothesis topology T_1 and the ML estimates of any free parameters, $\hat{\theta}_1$, derived for T_1 from the original data set.
- Use T_1 and reestimate free parameters θ_1 to get maximized log-likelihoods $L_1^{(i)}$ under H_0.
- Maximize likelihood over all topologies T_x and their respective parameters θ_x to get log-likelihoods $L_{ML}^{(i)}$.
- Calculate values of $\delta^{(i)} \equiv L_{ML}^{(i)} - L_1^{(i)}$, the set of these giving an estimate of the distribution (under H_0) of δ.
- Test whether the attained value of δ (from the original data) is a plausible sample from the estimated distribution of δ given by the set of the $\delta^{(i)}$ by seeing if it falls below the 95% point (for example) of the ranked list of the $\delta^{(i)}$. Such a one-sided test is appropriate because we know that δ must be >0; in this example, a 5% significance level is being used.

Notice that the test statistic δ is the same as in the KH- and SH-tests. The use of T_{ML} in the test means that the assumption $E[\delta] = 0$ cannot be made, but the use of parametric bootstrapping to generate data conforming to the null hypothesis means that this presents no problem: the repeated analysis of parametric bootstrap data sets guarantees the appropriate statistical properties. (Indeed, $E[\delta]$ may be estimated by $\bar{\delta}^{(i)}$, the mean of the $\delta^{(i)}$.) The fact that the data necessarily conform to the null hypothesis is also the reason that no centering procedure is necessary for this test (Table 1, level 4, option *u*).

This test has a substantial time penalty, however, caused by the need to repeatedly

maximize likelihoods over topologies under the hypothesis H_A. The same penalty exists in the basic SH-test (*posNPfcd*) above but is avoided by using the RELL method (test *posNPncd* above). Although we do not have theoretical results to justify applying all the most useful approximations described above for nonparametric tests, we have a several suggestions for possible ways to reduce the computational burden of the above parametric test.

The first suggestion is to use RELL-like methods applied only to the a priori–specified null hypothesis topology T_1 (Table 1, level 3, option *p*):

Test *posPpud* (approximation under H_0):
- Calculate the test statistic $\delta \equiv L_{ML} - L_1$.
- Simulate data sets i by parametric bootstrapping, based on the null hypothesis topology T_1 and the ML estimates of any free parameters, $\hat{\theta}_1$, derived for T_1 from the original data set.
- Use T_1 and the ML estimates of parameters $\hat{\theta}_1$ to get log-likelihoods $L_1'^{(i)}$ under H_0.
- Maximize likelihood over all topologies and their respective θ_x to get maximized log-likelihoods $L_{ML}^{(i)}$ under H_A.
- Calculate values of $\delta'^{(i)} \equiv L_{ML}^{(i)} - L_1'^{(i)}$.
- Test whether the attained value of δ (from the original data) is a plausible sample from the estimated distribution of δ given by the set of the $\delta'^{(i)}$ by seeing if it falls below the 95% point of the ranked list of the $\delta'^{(i)}$ (one-sided test; 5% significance level used in this example).

This saves on the maximization of parameters under the fixed topology T_1 and results in a small saving in computation time. It does not address the more difficult problem of repeated maximizations over topologies in the alternative hypothesis. As with other tests described above, the approximation under H_0 can be trusted and the significance levels taken at face value. Alternatively, note that necessarily $L_1'^{(i)} \leq L_1^{(i)}$ and so $\delta'^{(i)} \geq \delta^{(i)}$. Therefore, if this test rejects H_0 (the attained δ is too big), then this result is certain (the approximation cannot have changed the result). But if the test does not reject H_0 (the attained δ is sufficiently small), we do not know whether it would have been rejected if the exact $\delta^{(i)}$ had been used in place of the $\delta'^{(i)}$, and the test does not give us a definitive result.

A similar approach applied to the alternative hypothesis will be less effective (e.g., for a test denoted *posPnud*). Although it provides a much greater scope for saving time under H_A, because searches are performed over topologies, using fixed values of T_{ML} and $\hat{\theta}_{ML}$ from the ML analysis (over all trees) of the original data to assess the replicate data set is not sensible. The original T_{ML} and corresponding $\hat{\theta}_{ML}$ will probably be far from the optimal values for replicate data sets (which were simulated using the original T_1 and $\hat{\theta}_1$), so the difference between $L_{ML}^{(i)}$ and its possible approximation $L_{ML}'^{(i)}$ may be large.

However, ML estimates of some parameters of nucleotide substitution models are known to be quite stable over different topologies (e.g., Yang et al., 1994, 1998; Sullivan et al., 1996; Yang, 1997). Examples of such parameters are base frequencies (π_A, π_C, π_G, π_T), the transition/transversion rate ratio κ, and the shape parameter α of the gamma distribution widely used to model among-sites rate variation. Therefore, we think it reasonable to use fixed values of these parameters estimated under H_0 from each bootstrap data set i (i.e., the components of $\hat{\theta}_1^{(i)}$ that are not the lengths of branches of T_1 and are thus common to all topologies T_x) when assessing that data set under H_A. This gives the following test:

Test *posPpud* (approximation under H_A):
- Calculate the test statistic $\delta \equiv L_{ML} - L_1$.
- Simulate data sets i by parametric bootstrapping, based on the null hypothesis topology T_1 and the ML estimates of any free parameters, $\hat{\theta}_1$, derived for T_1 from the original data set.
- Use T_1 and reestimate free parameters θ_1 to get maximized log-likelihoods $L_1^{(i)}$ under H_0 (and respective optimal values of $\hat{\theta}_1^{(i)}$).
- Maximize likelihood over all topologies T_x to get log-likelihoods $L_{ML}'^{(i)}$ under H_A: these maximizations all fix the values of substitution process parameters to be equal to $\hat{\theta}_1^{(i)}$, but the maximization is performed over topologies T_x and their respective branch length parameters.

- Calculate values of $\delta'^{(i)} \equiv L'^{(i)}_{\mathrm{ML}} - L^{(i)}_1$.
- Test whether the attained value of δ (from the original data) is a plausible sample from the estimated distribution of δ given by the set of the $\delta'^{(i)}$ by seeing if it falls below the 95% point of the ranked list of the $\delta'^{(i)}$ (one-sided test; 5% significance level used in this example).

(Note that the two preceding tests both receive the mnemonic *posPpud* because they vary only in the form of the approximation used in their likelihood maximizations [Table 1, level 3]. A more complex naming system that would assign different mnemonics to these tests seems unwarranted in this paper.) Now, some substantial saving of time is made as the substitution process parameter values are fixed during the likelihood optimizations under H_A. The greater problem of optimizing over topologies is still not addressed. For this test, we know that necessarily that $L'^{(i)}_{\mathrm{ML}} \leq L^{(i)}_{\mathrm{ML}}$ and so $\delta'^{(i)} \leq \delta^{(i)}$. Therefore, if this test fails to reject H_0 (attained δnot excessively large relative to the null hypothesis distribution of the $\delta'^{(i)}$), then this result is certain. If this test does reject H_0, this could in principle be a consequence solely of the approximation. However, we expect the effect of this approximation to be small; in the example given below, it is insignificant and has no bearing on the conclusions reached.

If approximations are made under both hypotheses, for example, by some combination of the two *posPpud* tests above (and as in tests *priNPncd*, *priNPncn*, *priNPnca*, *priNPncs*, and *posNPncd* above), it is no longer possible to make general statements about the direction of the bias that they produce in the $\delta^{(i)}$. The precise effects of the combination of such approximations in a posteriori parametric tests require further investigation. Approximations based on assumptions of a normal distribution for δ(Table 1, level 5, options *n*, *a*, and *s*) seem unlikely to be useful in tests designed for hypotheses chosen a posteriori, given that the necessary condition of $\delta \geq 0$ indicates a truncation of the distribution of δ, which precludes normality.

Other SOWH-Like Tests

It is also straightforward to devise a parametric bootstrap test of the following hypotheses, which are akin to those of the fundamental KH-test for a priori–specified trees T_1 and T_2:

H_0: T_1 is the true topology

H_A: T_2 is the true topology

This fundamental version of such a test would be denoted *priPfud* (see Table 1). Denoting the topology with second highest likelihood by T_{ML2}, we could also devise a parametric bootstrap test of the hypotheses:

H_0: T_{ML2} is the true topology

H_A: T_{ML} is the true topology

This test would be based on modifications of *posPfud* by using the test statistic $\delta \equiv L_{\mathrm{ML}} - L_{\mathrm{ML2}}$, data simulated by using T_{ML2} and $\hat{\theta}_{\mathrm{ML2}}$; ML analysis of simulated data sets to find the distinct topologies $T^{(i)}_{\mathrm{ML}}$ and $T^{(i)}_{\mathrm{ML2}}$ which give the greatest and second-greatest likelihoods, respectively; and $\delta^{(i)} \equiv L^{(i)}_{\mathrm{ML}} - L^{(i)}_{\mathrm{ML2}}$. Having introduced the general principles of such tests, we will not go into further details here. We also draw readers' attention to the related parametric bootstrap test of monophyly described by Huelsenbeck et al. (1996), which compares partially constrained topologies (chosen a priori) with the ML topology (chosen a posteriori).

EXAMPLES

HIV-1 Subtypes A, B, D, and E gag *and* pol *Nucleotide Sequences*

Six homologous sequences, each consisting of 2,000 base pairs (bp) from the *gag* and *pol* genes, were selected from isolates of HIV-1 subtypes A (two sequences, A1 and A2), B (one sequence), D (one sequence), and E (two sequences, E1 and E2). The sequences were easily aligned by eye. The conventional phylogeny for these subtypes would group the two subtype A sequences and also the two subtype E sequences—that is, T_1 = ((A1, A2), (B, D), (E1, E2))—for which the optimal log-likelihood is $L_1 = -5{,}073.75$. For our sequences, however, the ML phylogeny indicated that the subtype A sequences did not cluster together; that is, T_{ML} = (A1, (B, D), (A2, (E1, E2))) with $L_{\mathrm{ML}} = -5{,}069.85$. In this example, all ML calculations were performed with the general time reversible model of nucleotide substitution, using a

TABLE 3. Results of statistical tests of topologies for HIV-1 *gag* and *pol* gene nucleotide data set.

Test code	Notes	*P*-value[a]
priNPfcd	KH-test (incorrect application); full optimization; direct estimation of *P*-value	0.38 (0.19)
priNPfcn	KH-test (incorrect application); full optimization; normal approximation for distribution of δ	0.41 (0.20)
priNPncs	KH-test (incorrect application); RELL approximation; stronger normal approximation for distribution of $\delta(k)$	0.38 (0.19)
posNPfcd	SH-test; full optimization; direct estimation of *P*-value	0.26
posPfud	SOWH-test; full optimization; direct estimation of *P*-value	0.002
posPpud	SOWH-test; partial optimization under H_A; direct estimation of *P*-value	0.002

[a] First value is from a two-sided test, as widely used to date; second value, when present, is for the more appropriate one-sided test.

gamma distribution to model rate heterogeneity among sites (REV+Γ; Yang, 1994, 1996, 1997); the parameters θ_x for topology T_x are branch lengths, base frequencies, parameters describing the relative rates of substitution between each nucleotide pair, and the shape parameter α of the gamma distribution.

We illustrate some of the statistical tests described above by investigating whether or not the data provide significant evidence that T_{ML} is to be preferred over T_1. For all the tests we performed, the test statistic $\delta = L_{ML} - L_1 = -5{,}069.85 - (-5{,}073.75) = 3.90$. Because T_{ML} has been selected for testing a posteriori, that is, as a consequence of having the highest likelihood, the KH-test is inappropriate (but was performed for comparative purposes), and the SH- or SOWH-tests should be used. These tests were performed as described above, with 1,000 replicates used whenever parametric or nonparametric bootstraps were performed. The results are summarized in Table 3.

We performed three versions of the KH-test, two using full likelihood optimizations and computing the tests' *P*-values either directly (test *priNPfcd*) or by assuming a normal distribution for δ (test *priNPfcn*), and one using the strongest assumption of normality of the sitewise $\delta(k)$ (test *priNPncs*). In all cases, both a two-sided test and a one-sided test were performed. The two-sided test is inappropriate for this a posteriori test (as indeed is the entire KH-test) but represents the computation performed in the most widely available implementations of the KH-test (PHYLIP, Felsenstein, 1995; PUZZLE, Strimmer and von Haeseler, 1996; Strimmer et al., 1997; and PAUP*, Swofford, 1998). The one-sided test is more suitable and, as described above, at least has the possibility of giving a statistically interpretable result. Indeed, that is the case in this example: the one-sided *P*-values of ~0.2 indicate no rejection of the null hypothesis, and as explained above, this conclusion must necessarily be maintained by the correction inherent in the SH-test. We also note the good agreement between the *P*-values calculated by the three variants of the KH-test.

Our application of the SH-test used full likelihood optimizations (test *posNPfcd*) and permitted the consideration of three topologies as possibly true: T_1, T_{ML}, and the topology (A2, (B, D), (A1, (E1, E2))). We report only the significance level for the test of T_1 against T_{ML}. This test, with its allowance for the a posteriori selection of one topology, must give a higher *P*-value (i.e., a less significant result), and this is confirmed in Table 3. There seems no way to draw any general conclusions about the size of the difference in the *P*-values (0.26 vs. 0.19–0.20 for the KH-tests). Figure 1 shows the distribution of the 1,000 replicate values $\delta_1^{(i)}$ against which the attained value $\delta_1 = 3.90$ is compared. We conclude that the SH-test indicates no significant difference between T_1 and T_{ML}; therefore, we do not reject T_1 in favor of T_{ML} for these data. We also note from Figure 1 that the minimum value of δ_1 that would indicate rejection of the null hypothesis (T_1) in this example is ~8.8 (the value of δ for which the SH-test distribution reaches a cumulative frequency of 0.95).

The results of the SOWH-test are very different. We performed two versions of this test, one using full likelihood optimizations

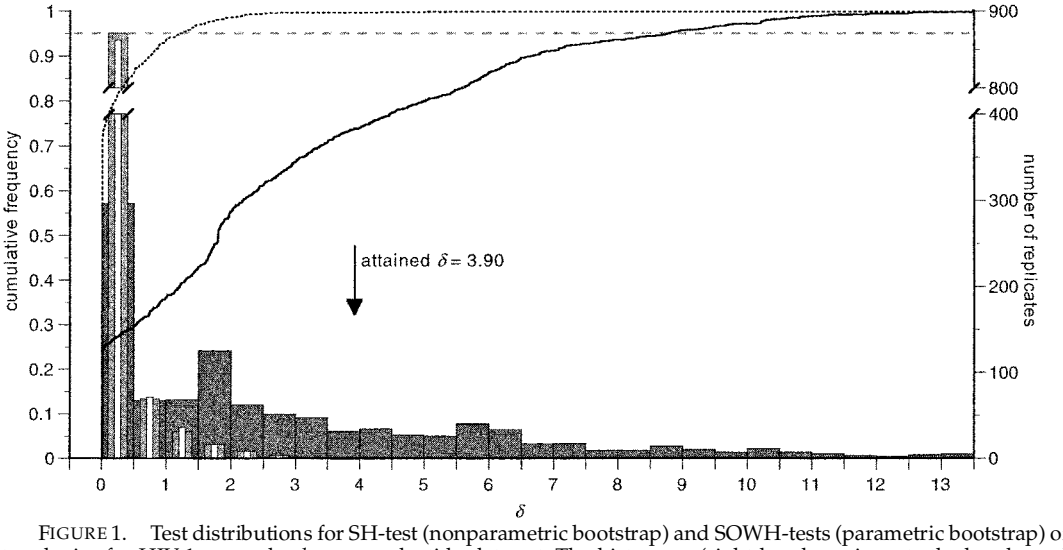

FIGURE 1. Test distributions for SH-test (nonparametric bootstrap) and SOWH-tests (parametric bootstrap) of topologies for HIV-1 *gag* and *pol* gene nucleotide data set. The histogram (right-hand *y*-axis; note the break used on this scale) shows the distribution over 1,000 replicates *i* of the $\delta_1^{(i)}$ (SH-test, code *posNPfcd*; wide dark-gray bars), $\delta^{(i)}$ (SOWH-test, code *posPfud*; narrow white bars), and $\delta'^{(i)}$ (SOWH-test, code *posPpud*; light-gray bars). The curves (left-hand *y*-axis) show the cumulative frequency distributions of the $\delta_1^{(i)}$ (SH-test; solid line) and $\delta^{(i)}$ (SOWH-test *posPfud*; dashed line). The cumulative frequency distribution of the $\delta'^{(i)}$ (SOWH-test *posPpud*) is indistinguishable from that of the $\delta^{(i)}$. The points at which the horizontal line (dashed gray) at a cumulative frequency of 0.95 crosses these curves indicate the values of δ that must be exceeded for a significant result at the 5% level. Given the attained value of $\delta = 3.90$, the SH-test is not significant and does not reject T_1, but the SOWH-tests are highly significant and reject T_1 in favor of T_{ML} (see text for further details).

(test *posPfud*) and one using the approximation described above as test *posPpud* (approximation under H_A). The differences between these two tests were negligible in this example, and both indicated a *P*-value of 0.002 (Fig. 1; Table 3). For these data, this test strongly rejects topology T_1 in favor of T_{ML}. As Figure 1 shows, any observed value of δ exceeding ~1.2 would have resulted in rejection of T_1 at the 95% level. The attained value is 7.7 standard deviations above the mean of the SOWH-test statistic distribution.

One explanation of the difference between the significance levels for the SH-test and the SOWH-test is the different forms of their null hypotheses. In this example the SH-test considers whether the three competing topologies are equally good explanations of the data, whereas the SOWH-test considers whether other topologies are better than the single topology T_1. As a consequence, the SH-test is permitting more a priori possible topologies in its null hypothesis, which will generally lead to more conservative results—an effect of the allowances made for the multiple statistical comparisons being made between the ML and all other permitted hypotheses.

Another factor affecting the significance levels of the different tests may simply be the increase in power expected for a parametric test over a nonparametric test. We also recall the reliance of parametric tests on the models that they assume. Although we may hope for robustness of tests to inaccuracy of models, this has generally been left untested in phylogenetics. To examine whether the REV+Γ model fits the data well in the present example, we performed ML analyses under a variety of nucleotide substitution models to compare those models (Goldman, 1993; Yang et al., 1994). The results of these analyses are shown in Table 4. It is immediately evident that the REV+Γ model fits these data significantly better than any of the other models considered, in agreement with the good performance of both the REV and Γ components reported over a variety of data sets (see Yang, 1994, 1996; Arvestad and Bruno, 1997; and references therein). We conclude that in this example all reasonable steps have been taken to exclude any effects on the SOWH-test attributable to model inaccuracy.

TABLE 4. Maximum likelihood scores for HIV-1 *gag* and *pol* gene data set under various models of nucleotide substitution. Models JC69 (Jukes and Cantor, 1969), K80 (Kimura, 1980), F81 (Felsenstein, 1981), HKY85 (Hasegawa et al., 1985), and REV (Yang, 1994) were implemented, each without (no Γ) and with (+Γ) a gamma distribution to model rate heterogeneity amongst sites (Yang, 1996, 1997). All calculations were performed with the topology referred to in the text as T_{ML}. Numbers given represent the log-likelihood value by which each model is worse than the best value, attained under the REV+Γ model, −5,069.85. Also shown, in parentheses, are the numbers of free parameters in each substitution model. Pairs of nested models can be compared by using a test statistic that is twice the log-likelihood difference between those models, assessed with either a χ^2 distribution (models compared are both of the no 'Γ' form or both of the +'Γ' form) or a $\not{\chi}^2$ distribution (exactly one of the models compared is of the +'Γ' form), degrees of freedom are given by the difference in the numbers of free parameters. For full details of these tests see, for example, Yang et al. (1994) and Goldman and Whelan (2000).

Substitution model	No Γ	+Γ
JC69	395.08 (0)	349.93 (1)
K80	190.28 (1)	131.32 (2)
F81	280.19 (3)	231.43 (4)
HKY85	81.29 (4)	12.79 (5)
REV	65.09 (8)	0 (9)

Mammalian Mitochondrial Protein Amino Acid Sequences

Shimodaira and Hasegawa (1999) illustrated the SH-test with a data set consisting of aligned mt protein sequences, each comprising 3,414 aa, from six mammals: human, harbor seal, cow, rabbit, mouse, and opossum. The grouping (harbor seal, cow) was assumed to be true, which left 15 candidate topologies to be evaluated. Shimodaira and Hasegawa (1999) applied the SH-test to this data set, comparing all 15 candidate topologies simultaneously, and concluded that seven topologies could not be rejected. To illustrate the SOWH-test, we selected (a priori) the topology T_1 = ((human, ((harbor seal, cow), rabbit)), mouse, opossum), called topology $\alpha = 2$ by Shimodaira and Hasegawa (1999), To test against the ML topology, which for these data is T_{ML} = (((human, (harbor seal, cow)), rabbit), mouse, opossum)—topology $\alpha = 1$ of Shimodaira and Hasegawa (1999). We compare our results from the SOWH-test with Shimodaira and Hasegawa's (1999) results for analogous comparisons from KH- and SH-tests. In this example, as in Shimodaira and Hasegawa (1999), all ML calculations were performed using a model of mammalian mt aa replacement described by Yang et al. (1998), with aa frequencies estimated from the data set being analyzed and using a gamma distribution to model rate heterogeneity amongst sites (mtmam+F+Γ; see also Yang, 1997). For this model, the optimal log-likelihoods for these topologies were $L_1 = -21{,}727.26$ and $L_{ML} = -21{,}724.60$; therefore, the test statistic for all the tests of topologies considered below was $\delta = L_{ML} - L_1 = -21{,}727.26 - (-21{,}724.60) = 2.66$. Table 5 summarizes the results of the KH-, SH-, and SOWH-tests for these data. The KH- and SH- test results are taken from Shimodaira and Hasegawa (1999) and were calculated by using RELL approximations and estimating the tests' *P*-values directly (without a normal approximation). The SOWH-test was performed by using full likelihood optimizations.

The *P*-value obtained from a one-sided comparison in the KH-test, as given by Shimodaira and Hasegawa (1999), was 0.36, which suggests that T_1 cannot be rejected in favor of T_{ML}. As explained above, this conclusion must be maintained by the SH-test and we see from Table 5 that this is so ($P = 0.81$). Notice that the difference between the *P*-values for the (one-sided) KH- and SH-tests (0.36 and 0.81, respectively) is considerably greater than in the HIV-1 example (~0.20 and 0.26, respectively).

The SOWH-test again gives very different results. The *P*-value from this test, for 1,000 replicate parametric bootstrap data sets, is estimated to be <0.001—in other words, in none of the 1,000 replicates *i* did the value

TABLE 5. Results of statistical tests of topologies for mammalian mitochondrial protein amino acid data set.

Test code	Notes	*P*-value
priNPncd	KH-test (incorrect application); RELL approximation; direct estimation of *P*-value	0.36[a]
posNPncd	SH-test; RELL approximation; direct estimation of *P*-value	0.81[a]
posPfud	SOWH-test; full optimization; direct estimation of *P*-value	<0.001

[a] From Shimodaira and Hasegawa (1999).

of $\delta^{(i)}$ equal or exceed the value $\delta = 2.66$ observed for the real data. (The attained value $\delta = 2.66$ lies 27.8 standard deviations above the mean of the $\delta^{(i)}$.) Thus, topology T_1 is very strongly rejected in favor of T_{ML}.

As with the HIV-1 example above, there is no obvious single reason for the contrasting results of the SH- and SOWH-tests for these mt protein sequences. The SOWH-test considers only one a priori topology, T_1, and of the 1,000 replicate data sets generated by using T_1 only 7 resulted in topologies different from T_1 when analyzed by ML. Evidently, if this topology, its corresponding parameter values $\hat{\theta}_1$ (as estimated by ML from the original mt protein data set), and the mtmam+F+Γ model of aa replacement were all adequate, then we would expect to retrieve the correct topology from a data set of 3,414 aa with high probability $(1{,}000 - 7)/1{,}000 \approx 0.99$; consequently, our finding that for the original data T_{ML} and not T_1 is optimal seems unreconcilable with the hypothesis that T_1 is true. In contrast, the SH-test has 15 topologies considered equally plausible a priori in its null hypothesis and therefore the significance level assigned to a particular one of these, e.g., T_1, is reduced. The effects of differences between parametric and nonparametric tests and the possibility that the mtmam+F+Γ model is inadequate have not been assessed.

Discussion

We want to emphasize once more that the problems described above with typical applications of the KH-test are very real and will have practical consequences in many applications. We contend that all future applications must use new methods such as the SH-test and the SOWH-test (above). Assessment of the results of published analyses based on incorrect applications of the KH-test must be made with extreme caution. The sole correction to these results that we have been able to derive (see above) will often generate inconclusive results, demanding reanalysis of data.

Evidently, it is vital that researchers think carefully about what phylogenetic hypotheses they wish to test. A priori hypotheses and a posteriori hypotheses can be quite different, as can the statistical distributions required to test them. It serves no scientific purpose to "cheat" and represent an a posteriori hypothesis as an a priori one simply for the expediency of a more readily available or faster test. The SOWH-test, as described above, tests a single a priori hypothesis of topology. If such tests are used repeatedly, to assess the significance of multiple trees, the issue of correcting significance values for multiple tests arises. This might occur with data sets for which large numbers of tree topologies are considered plausible a priori. Bar-Hen and Kishino (in press) describe a novel parametric likelihood-based test for computing simultaneous significance values for multiple topologies.

The SH-test simultaneously compares all members of a set $\mathcal{M}$ of topologies. The inclusion in $\mathcal{M}$ of all a priori possible topologies is important. Even topologies with low bootstrap replicate likelihoods ($L_x^{(i)}$) can readily affect the significance levels of other topologies, because these are based only on variations in likelihoods over bootstrap replicates ($\tilde{L}_x^{(i)} \equiv L_x^{(i)} - \bar{L}_x^{(i)}$). A posteriori selection of topologies for inclusion in or exclusion from $\mathcal{M}$ based on their likelihoods may thus bias all significance levels recorded—analogous to performing multiple comparisons tests on only a subset of a larger number of comparisons, selected (for example) to be the most (or least) significant. Decreasing the number of comparisons performed this way will unjustifiably increase the apparent significance levels of the results. In the HIV-1 example above, if the SH-test (*posNPfcd*) is applied by considering that all 105 topologies for six sequences are possibly true, the *P*-value for T_1 is increased from 0.26 to 0.90. Considering all 105 possible topologies in the SH-test applied in the mt protein sequence example above increases the *P*-value for T_1 from 0.81 to 0.93 (RELL approximation, *posNPncd*); for the topologies called α = 9–15 by Shimodaira and Hasegawa (1999), *P*-values are increased from significant values (<0.05) to nonsignficant values (>0.05). Clearly, the honest choice of a priori hypothesis topologies may be crucial to the conclusions ultimately drawn.

The claim is sometimes made—when phylogenies of different genes are compared, for instance—that no a priori topologies can be constructed. In such cases, however, one can usually recast the hypothesis and its statistical test differently. To use the example above, when comparing the evolutionary histories of different genes, we may restate this as a test of whether the two (or more) trees

are sample estimates of the same phylogeny (Rodrigo et al., 1993).

As illustrated in this paper, the results of parametric tests (e.g., the SOWH-test) and nonparametric tests (e.g., the SH-test) can appear to be very different. The SH-test may often appear to be more conservative than the SOWH-test. As we have explained, this may be due to some or all of the following phenomena: different forms of null hypotheses; increased power of parametric tests; and greater reliance of parametric tests on models of sequence evolution. The relative consequences of these and other effects require further investigation in the future.

Program and Data Availability

Notes on the use of PAUP* (Swofford, 1998) to perform SOWH-tests, and details of the HIV-1 nucleotide sequences (6 sequences, each 2,000 bp) and mammalian mt protein aa sequences (6 sequences, each 3,414 aa) used in the examples above, can be obtained from the authors at http://www.zoo.cam.ac.uk/zoostaff/goldman/tests and 'downstream' Web pages. A computer program, shtests, to perform SH-tests by using the RELL approximation (*posNPncd* above) can be obtained from Andrew Rambaut at http://evolve.zoo.ox.ac.uk/software/shtests. Versions of PHYLIP and PAUP* package programs (Felsenstein, 1995; Swofford, 1998) implementing the SH-test are currently under development (J. Felsenstein, pers. comm.; D. Swofford, pers. comm.).

Acknowledgments

Work by N.G. and A.G.R. on this topic was partially supported by the Isaac Newton Institute for the Mathematical Sciences programme on "Biomolecular Function and Evolution in the Context of the Genome Project" (July–December 1998). N. G. is supported by a Wellcome Trust Fellowship in Biodiversity Research. J.P.A. is supported by a NIH Institutional NRSA Interdisciplinary Training in Genomic Sciences Fellowship and by the University of Washington Center for AIDS Research (CFAR). We are very grateful for the assistance given to us by Hidetoshi Shimodaira, Joe Felsenstein, David Swofford, Hirohisa Kishino, and Andrew Rambaut throughout the preparation of this paper; for prepublication versions of papers provided by Hidetoshi Shimodaira, Thomas Buckley, and Hirohisa Kishino; and for critical readings of draft versions of the paper by Edward Holmes, Ann Oakenfull, Tim Massingham, Martin Embley, and Andrew Rambaut. Andrew Rambaut and Korbinian Strimmer provided all of the 105-topology SH-test *P*-value calculations given in the Discussion.

References

ADACHI, J., AND M. HASEGAWA. 1996. MOLPHY: Programs for molecular phylogenetics based on maximum likelihood, vers. 2.3. Institute of Statistical Mathematics, Tokyo.

ARVESTAD, L., AND W. J. BRUNO. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. J. Mol. Evol. 45:696–703.

BAR-HEN, A., AND H. KISHINO. In press. Comparing the likelihood functions of phylogenetic trees. Ann. Inst. Stat. Math.

CUNNINGHAM, C. W., H. ZHU, AND D. M. HILLIS. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52:978–987.

EFRON, B. 1982. The jackknife, the bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics, volume 38. Society for Industrial and Applied Mathematics, Philadelphia.

EFRON, B., AND R. TIBSHIRANI. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1:54–77.

EFRON, B., E. HALLORAN, AND S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13434.

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. 1995. PHYLIP (Phylogenetic inference package), version 3.57. Univ. of Washington, Seattle.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

GOLDMAN, N., AND S. WHELAN. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. 17:975–978.

HALL, P., AND S. R. WILSON. 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47:757–762.

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.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1988. Phylogenetic inference from DNA sequence data. Pages 1–13 *in* Statistical theory and data analysis II (K. Mutusita, ed.). Elsevier, Amsterdam.

HILLIS, D. M., B. K. MABLE, AND C. MORITZ. 1996. Applications of molecular systematics: the state of the field and a look to the future. Pages 515–543 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

HOPE, A. C. A. 1968. A simplified Monte Carlo significance test procedure. J. R. Statist. Soc. B 30:582–598.

HUELSENBECK, J. P., AND K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28:437–466.

HUELSENBECK, J. P., AND B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232.

HUELSENBECK, J. P., D. M. HILLIS, AND R. NIELSEN. 1996. A likelihood-ratio test of monophyly. Syst. Biol. 45:546–558.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

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. 31:151–160.

MARRIOTT, F. H. C. 1979. Barnard's Monte Carlo tests: how many simulations? Appl. Statist. 28:75–77.

RODRIGO, A. G., M. KELLY-BORGES, P. R. BERGQUIST, AND P. L. BERGQUIST. 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. N.Z. J. Bot. 31:257–268.

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

STRIMMER, K., AND A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

STRIMMER, K., N. GOLDMAN, AND A. VON HAESELER. 1997. Bayesian probabilities and quartet puzzling. Mol. Biol. Evol. 14:210–211.

SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. 42:308–312.

SWOFFORD, D. L. 1998. PAUP* 4.00: *Phylogenetic analysis using parsimony (and other methods). Sinauer, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244.

WESTFALL, P. H., AND S. S. YOUNG. 1993. Resampling-based multiple testing: Examples and methods for *p*-value adjustment. John Wiley & Sons, New York.

YANG, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

YANG, Z. 1996. Among-site variation and its impact on phylogenetic analysis. TREE 11:367–372.

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

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. 44:384–399.

YANG, Z., R. NIELSEN, AND M. HASEGAWA. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15:1600–1611.

ZHANG, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. 16:868–875.

Received 9 November 1999; accepted 17 December 1999
Associate Editor: R. Olmstead