Syst. Biol. 52(6):781–793, 2003
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390258949

Least and Most Powerful Phylogenetic Tests to Elucidate the Origin of the Seed Plants in the Presence of Conflicting Signals under Misspecified Models

STÉPHANE ARIS-BROSOU

Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695-7566, USA; E-mail: stephane@statgen.ncsu.edu

Abstract.—Several tests of molecular phylogenies have been proposed over the last decades, but most of them lead to strikingly different *P*-values. I propose that such discrepancies are principally due to different forms of null hypotheses. To support this hypothesis, two new tests are described. Both consider the composite null hypothesis that all the topologies are equidistant from the true but unknown topology. This composite hypothesis can either be reduced to the simple hypothesis at the least favorable distribution (frequentist significance test [FST]) or to the maximum likelihood topology (frequentist hypothesis test [FHT]). In both cases, the reduced null hypothesis is tested against each topology included in the analysis. The tests proposed have an information-theoretic justification, and the distribution of their test statistic is estimated by a nonparametric bootstrap, adjusting *P*-values for multiple comparisons. I applied the new tests to the reanalysis of two chloroplast genes, *psaA* and *psbB*, and compared the results with those of previously described tests. As expected, the FST and the FHT behaved approximately like the Shimodaira–Hasegawa test and the bootstrap, respectively. Although the tests give overconfidence in a wrong tree when an overly simple nucleotide substitution model is assumed, more complex models incorporating heterogeneity among codon positions resolve some conflicts. To further investigate the influence of the null hypothesis, a power study was conducted. Simulations showed that FST and the Shimodaira–Hasegawa test are the least powerful and FHT is the most powerful across the parameter space. Although the size of all the tests is affected by misspecification, the two new tests appear more robust against misspecification of the model of evolution and consistently supported the hypothesis that the Gnetales are nested within gymnosperms. [Approximately unbiased test; bootstrap proportion; hypothesis test; *P*-value adjustment; Shimodaira–Hasegawa test; significance test.]

It is now accepted that estimating phylogenies is a problem of statistical inference. However, because of sampling errors and because models of molecular evolution at best only approximate reality, the estimated tree topology, although optimal with respect to some measures, may not be correct. Among the model-based approaches developed to assess the confidence of an estimated topology, the bootstrap probability (BP; Felsenstein, 1985), which can be used to test hypotheses (Shao and Tu, 1996), and the Kishino–Hasegawa (KH) test (Kishino and Hasegawa, 1989) have been extensively used. However, the BP is not accurate (e.g., Hillis and Bull, 1993; Newton, 1996; for an interpretation, see Felsenstein and Kishino, 1993), and because of the inclusion of the maximum likelihood (ML) tree among the tested topologies, the KH test is subject to selection bias (Shimodaira and Hasegawa, 1999; Goldman et al., 2000).

These problems were corrected, respectively, by implementing a double bootstrap procedure accommodating the geometry of the sample space (Efron et al., 1996) and by taking the selection bias into account (the SH test: Shimodaira and Hasegawa, 1999). However, Goldman et al. (2000) and Strimmer and Rambaut (2002) pointed out that the *P*-values of the SH test at a given α level increase with the number of topologies included in the analysis. Tests can be made less conservative by weighting the test statistic of the SH test (WSH test: Shimodaira and Hasegawa, 1999; Buckley et al., 2001) or by using an approximately unbiased (AU) test (Shimodaira, 2002), which was derived as a faster and more accurate approximation to the BP than the double bootstrap (Efron et al., 1996).

Besides the sensitivity of the SH test to the number of topologies included in the analysis, an important issue is that each test leads to different *P*-values and consequently to different confidence sets of topologies (Shimodaira and Hasegawa, 1999; Goldman et al., 2000; Whelan et al., 2001; Shimodaira, 2002; Strimmer and Rambaut, 2002). Shimodaira (2002) showed that selection procedures can be biased, but the statistical properties of the different tests are generally not well understood and their respective power has never been investigated thoroughly.

Other selection procedures have been described and include the parametric bootstrap (Waterman, 1995:375; Huelsenbeck et al., 1996; Swofford et al., 1996; Huelsenbeck and Crandall, 1997; Goldman et al., 2000), information criteria (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1990; Hasegawa et al., 1991; Cao et al., 2000; Strimmer and Rambaut, 2002), and Bayesian approaches (Yang and Rannala, 1997; Larget and Simon, 1999; Huelsenbeck and Imennov, 2002; Aris-Brosou, 2003). These selection procedures can be very sensitive to the specification of the model of evolution, which is problematic (e.g., Shimodaira, 2002), but the impact of model specification on the outcome of classical approaches (e.g., SH, WSH, AU, BP) is not well studied (but see Steel et al., 1993).

The objective of the present article is twofold. First, I suggest that the difference among previous tests of molecular phylogenies principally is due to different forms of hypotheses tested. Second, I show that the form of the null hypothesis dramatically affects the power of these tests. To support these assertions, two new nonparametric tests are proposed: the frequentist

significance test (FST) and the frequentist hypothesis test (FHT). The objectives of these two tests differ: FST selects topologies that are close to the unknown topology (where "closeness" is determined by the α level of the test and by the least favorable distribution), whereas FHT aims at selecting the correct topology. The null and alternative hypotheses are constructed to match those of the SH test and the BP, respectively. A theoretical justification to these tests is provided and their limitations are shown. The main simulation result is that FST and FHT have power curves close to those of the SH test and the BP, respectively. Further simulations suggested that the size of the new tests, FST and FHT, is less sensitive to specification of the substitution model than existing tree topology selection procedures such as the BP, the SH, the WSH, or the AU test. These simulation results are in stark contrast with the extreme sensitivity to specification of the substitution model exhibited by all the tests, including FST and FHT, when real data are analyzed. These results suggest that most Markov models used in phylogenetics, homogeneous in time and space, do not always describe molecular evolution appropriately.

Materials and Methods

Significance and Hypothesis Tests

Following the notation used by Goldman et al. (2000), let T_i be the topology of the ith prespecified tree. For a given data set X, the topology can be estimated by ML. One peculiarity is that each T_i has its own log-likelihood function, $\ell(\theta_i, T_i \mid X) = \ln\{p(X \mid \theta_i, T_i)\}$, where θ_i is a vector of nuisance parameters, typically the branch lengths (conditional on T_i) and the parameters of the substitution model. Therefore, a given topology T_i is not a parameter in the usual sense but rather is similar to a family of distributions with parameters θ_i (Yang et al., 1995). When estimating a topology under the ML criterion, $\ell(\theta_i, T_i \mid X)$ is maximized with respect to θ_i to obtain $\ell(\hat{\theta}_i, T_i \mid X)$. Ideally, this procedure is carried out for each possible topology, although in practice some heuristics are used (e.g., Swofford et al., 1996). The topology with the largest likelihood, $\ell(\hat{\theta}_{ML}, T_{ML} \mid X) = \max_i \ell(\hat{\theta}_i, T_i \mid X)$, is taken as the estimate of the true topology.

Related to the issue of topology estimation is that of hypothesis testing. Because of the nature of topologies, the corresponding tests concern distributions and not parameters as when testing for the molecular clock (Felsenstein, 1981) or for the presence of sites under positive selection (Yang et al., 2000). Rather than selecting the best topology, we want to select out of k prespecified topologies the set of those that are the closest to the true and unknown distribution (topology) h at a prespecified significance level. This objective naturally leads to use the Kullback–Leibler (KL) distance (Kullback and Leibler, 1951; see also Kishino and Hasegawa, 1989) to test whether the k topologies are equidistant from the distribution h:

$$\mathrm{H}_0 : E_h\{\ell(\theta_1, T_1 \mid X)\} = \cdots = E_h\{\ell(\theta_k, T_k \mid X)\} \quad (1)$$

for all $i = 1, \ldots, k$, where θ_i are the unknown nuisance parameters. This composite null hypothesis is tested against the class of alternatives:

$$\mathrm{H}_\mathrm{A}: \mathrm{H}_0 \text{ does not hold for } T_i \text{ with } E_h\{\ell(\theta_i, T_i \mid X)\} < E_h\{\ell(\theta_j, T_j \mid X)\} \quad (2)$$

for some $i \neq j = 1, \ldots, k$. The distributions (topologies T_i) for which H_0 is not rejected are included in the set of topologies closer to h at the prespecified level α.

The difficulty with computing the expectations in Equations 1 and 2 is that the distribution h and the nuisance parameters θ_i are unknown. The basic idea to get around this difficulty is to reduce the composite null hypothesis to a simple one (Lehmann, 1986) and consider the nuisance parameters at their ML estimates under their respective topologies. The Bayesian solution to this problem of testing topologies is probably less ad hoc (see below and Discussion section). In a frequentist framework, reducing H_0 to a simple hypothesis can be done in two different ways and leads to two distinct null hypotheses and hence to two different tests, the FST and the FHT.

FST.—The reduction of H_0 to a simple hypothesis is done by considering a weighted average of the topologies in H_0 over the tree space T: Equation 1 is replaced by the simple null hypothesis $\mathrm{H}_{0,\tau}$ that the logarithm of the probability density of the data X is given by

$$h_{0,\tau}(x) = \int_\mathrm{T} \ell(\theta, T \mid X)\, d\tau(T), \quad (3)$$

where $\tau(T)$ is a probability distribution over the space T of the possible topologies (Lehmann, 1986:104). The choice of τ reflects H_0 in that it must convey no information with respect to the different topologies. This argument is essentially Bayesian: when no prior knowledge about the different topologies T_i is available, these topologies are assigned an equal weight by choosing the distribution τ to be uniform. Note however that this is not a fully Bayesian argument because Equation 3 does not integrate over the nuisance parameters θ. An ad hoc but common treatment of the nuisance parameters is to maximize this least favorable distribution (see Shimodaira and Hasegawa, 1999) $h_{0,\tau}$ by taking the average of the estimated log-likelihoods $\ell(\hat{\theta}_i, T_i \mid X)$ over the topology space. With n species, there are $\prod_{i=3}^{n}(2i-5)$ unrooted topologies (e.g., Swofford et al., 1996), which is equal to $2^{n-2}\Gamma(n-3/2)/\sqrt{\pi}$. Using Striling's approximation ($\lim_{m\to\infty}\{\Gamma(m+1)\} = (m/e)^m\sqrt{2m\pi}$), this is about $2^{n-2}[(n-5/2)/e]^{n-5/2}\sqrt{2n-5}$ for large n, so that the number of topologies grows much faster than exponentially. For large n, averaging over the whole topology space may therefore be impractical, and the average is taken only over the set of the k prespecified topologies,

$E_{i\in[1,k]}\{\ell(\hat{\theta}_i, T_i \mid X)\}$:

$$h_{0,\tau} = \sum_{i=1}^{k} \ell(\hat{\theta}_i, T_i \mid X)/k. \quad (4)$$

This reduced null hypothesis $H_{0,\tau}$ corresponds to an average (network) of trees and is equivalent to the null hypothesis of the SH test. The limitation of $H_{0,\tau}$ to the set of the k tested topologies, implicit in the SH test, explains why the size of tests at the least favorable distribution depends on how many topologies are included in the analysis. Because of the reduction of the composite hypothesis H_0 to the simple hypothesis $H_{0,\tau}$, which includes all the tested T_i, multiple comparisons may not be as relevant as for the hypothesis test presently described.

FHT.—In the way H_0 was reduced to $H_{0,\tau}$, no value was attached to the possibility that one of the tested topologies might be correct. In the context of ML, the a posteriori specified ML topology, T_{ML}, is the most likely to be the true topology. For the FST, τ was chosen to be uniform to reflect our lack of prior knowledge about the different topologies. To account for our knowledge that T_{ML} is most probable, a peculiar weight is assigned to this topology in $H_{0,\tau}$:

$$h_{0,\tau} = \ell(\hat{\theta}_{ML}, T_{ML} \mid X). \quad (5)$$

This formulation of the null hypothesis is equivalent to that of the bootstrap hypothesis test (Shao and Tu, 1996:177). Unlike FST, FHT is by construction insensitive to the number of trees included in the analysis because only pairs of topologies are compared. However, because of this pairwise structure, the P-values of FHT must be corrected for multiple comparisons.

P-value adjustments for multiple comparisons.—When comparing several topologies, multiple tests are performed and the resulting probability of rejecting a particular null hypothesis is larger than the prespecified significance level α. This risk of false discovery, also called the false discovery rate (FDR), is defined as the probability of rejecting at least one topology T_i given that the null hypothesis H_0 is true:

$$\text{FDR} = \Pr(\text{reject at least one } T_i \mid H_0). \quad (6)$$

The FDR is corrected by making tests more conservative and FDR is controlled if

$$\text{FDR} \leq \alpha. \quad (7)$$

FDR is usually controlled by adjusting P-values. One such adjustment is the Bonferroni correction: when performing k tests, a partial null hypothesis T_i is rejected when its P-value is less than α/k. The Bonferroni correction, as used in phylogenetics by Bar-Hen and Kishino (2000), controls the FDR at level α (e.g., Westfall and Young, 1993:44). However, for highly correlated data, as is the case with molecular sequences, the FDR may not be controlled by the Bonferroni correction. This correction is conservative for data with light-tailed sampling distributions but can become liberal for heavy-tailed distributions (Westfall and Young, 1993:44). By taking the distributional characteristics of the data into account, resampling techniques produce smaller adjusted P-values and increase the power of the correction. Adjusted P-values are computed as the smallest significance level for which a tree T_i is still rejected at level α over the resampled data sets. This procedure controls the FDR at level α, at least approximately (Westfall and Young, 1993:53).

The test statistics.—The KL distance (Kullback and Leibler, 1951) measures the distance between the generating model f of a process and an approximating model g:

$$d(f, g) = \int f(x \mid \theta) \log \frac{f(x \mid \theta)}{g(x \mid \theta)} dx, \quad (8)$$

i.e., the average with respect to the generating model of the logarithmic difference between the generating and the approximating model:

$$d(f, g) = E_f[\log f(x \mid \theta)] - E_f[\log g(x \mid \theta)]. \quad (9)$$

This quantity can be approximated by the relative estimated distance between f and g:

$$\hat{d}(f, g) = E_f[\log f(x \mid \hat{\theta})] - E_f[\log g(x \mid \hat{\theta})]. \quad (10)$$

Equation 10 is the approximation implicitly used when defining the composite null hypothesis in Equation 1. The generating model f is generally unknown, but when comparing two approximating models g_0 and g_1, the constants $E_f[\log f(x \mid \hat{\theta})]$ cancel out,

$$\hat{d}(f, g_0) - \hat{d}(f, g_1) = E_f[\log g_1(x \mid \hat{\theta})] - E_f[\log g_0(x \mid \hat{\theta})], \quad (11)$$

and the relative expected difference $E_f[\hat{d}(f, g_0) - \hat{d}(f, g_1)]$ can be approximated by

$$E_f[\hat{d}(f, g_0) - \hat{d}(f, g_1)] \approx \ell\left(\hat{\theta}_{g_1} \mid X\right) - \ell\left(\hat{\theta}_{g_0} \mid X\right). \quad (12)$$

Burnham and Anderson (1998:239–247) gave more details to derive Equation 12, herein after used as a test statistic t. To do so, the distributions under the null and alternative hypotheses are substituted for those of g_0 and g_1. Thus, according to Equation 12, $t_i = \ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}$, where $h_{0,\tau}$ is either $\sum_i^k \ell(\hat{\theta}_i, T_i \mid X)/k$ or $\max_i \ell(\hat{\theta}_i, T_i \mid X)$, respectively, for FST and FHT. This test looks like a likelihood ratio test in the case of the FHT, but the resemblance is not as clear in the case of the FST. A similar situation appears in the Bayesian version of the FST, which considers an average of log probabilities (Aris-Brosou, 2003).

Following Bar-Hen and Kishino (2000), the asymptotic distribution of the test statistics is easily derived. Given the sampling distribution Y, $\{t_i - E_Y[t_i]\}/\mathrm{var}(t_i)$ approximately follows a standard normal distribution. Here however the distribution of t_i is directly estimated by bootstrap, which includes in a second-level bootstrap, adjusting P-values for multiple comparisons.

The algorithm. —Let v be a parameter of some distribution. When testing the null hypothesis H_0: $v = v_0$ against the alternative H_A : $v \neq v_0$, resampling must be done in a way that reflects H_0 (Hall and Wilson, 1991). Let $\hat{v}$ be an estimator of the unknown quantity v, and $\hat{v}^*$ be the value of $\hat{v}$ computed from the bootstrapped samples. Testing H_0 against H_A is based on the unknown distribution of $\hat{v} - v_0$ under H_0, estimated by the distribution of $\hat{v}^* - \hat{v}$.

This centering step is applied to the computation of the P-values of both tests, FST and FHT. As shown above, the test statistic $\hat{v}$, defined a priori, is $\ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}$ (for simplicity's sake, I do not consider pivotal quantities for the moment), and with the same notation as above, $v_0 = 0$. The distribution of $\{\ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}\} - \{0\}$ is then estimated from the bootstrapped data by that of $\{\ell(\hat{\theta}_i^*, T_i \mid X^*) - h_{0,\tau}^*\} - \{\ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}\}$. Maximizing $\ell(\theta_i, T_i \mid X^*)$ with respect to θ_i for each bootstrapped data set is demanding, so that the RELL approximation (Kishino et al., 1990) is used. The value of $h_{0,\tau}^*$ from the bootstrapped samples is $\sum_i^k \ell(\hat{\theta}_i^*, T_i \mid X^*)/k$ for the FST and $\max_i \ell(\hat{\theta}_i^*, T_i \mid X^*)$ for the FHT. Pivotal quantities are obtained by standardizing the bootstrap distribution of $(\hat{v}^* - \hat{v})$ by the scale $\hat{\sigma}^*$ (variance of the log-likelihood distribution) to estimate that of $(\hat{v} - v_0)/\hat{\sigma}$ under the null hypothesis. The algorithm computing the P-values of both FST and FTH is adapted from that of Westfall and Young (1993:47).

1. Compute the test statistics $t_i = \{\ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}\}/\hat{\sigma}_i$ for each topology T_i ($i = 1, \ldots, k$) and initialize the counting variables $C_i = 0$. The scale $\hat{\sigma}_i$ is estimated by the standard error of the distribution of sitewise loglikelihoods (Yang, 1997).
2. Generate B bootstrapped pseudo data sets of sitewise log likelihoods (RELL approximation with, e.g., $B = 10{,}000$ replicates) to obtain the first two moments $\ell(\hat{\theta}_i^*, T_i \mid X^*)$ and $\hat{\sigma}_i^*$ of the empirical distributions of sitewise log likelihoods.
3. Compute the resampled statistic s_i^* as $(\{\ell(\hat{\theta}_i^*, T_i \mid X^*) - h_{0,\tau}^*\} - \{\ell(\hat{\theta}_i, T_i \mid X) - h_{0,\tau}\})/\hat{\sigma}_i^*$ for each T_i. When genes are combined, the resampling procedure is stratified within each partition of the original data set (e.g., Yoder and Yang, 2000).
4. Find the smallest s_j^* over the k tested trees. For each T_i, compare this minimum resampled statistic with the original test statistic t_i. If $\min_{1 \leq j \leq k}(s_j^*) \leq t_i$, then increment C_i by one.
5. Repeat steps 2–4 N times (e.g., $N = 1{,}000$). The estimated adjusted P-value is approximated by C_i/N.

Data Sets

To compare FST and FHT with existing tests and to examine the effect of specification of the substitution model on the confidence sets obtained by inverting the tests, I reanalyzed two highly conserved chloroplast photosystem genes, *psaA* and *psbB*. These genes were originally analyzed by Sanderson et al. (2000) to elucidate the origin of the seed plants. The alignments, comprising 3,795 nucleotides for the two genes, were obtained from Sanderson (2002).

Seed plants are composed of five extant lineages subdivided into three groups: (1) the angiosperms (Ag), (2) a gymnosperm group including conifers, *Gingko*, and cycads (Gy) and (3) the Gnetales (Gn) (Donoghue, 1994). All three possible evolutionary scenarios have been proposed in the literature (see Sanderson et al., 2000). In the first scenario (scenario A), Gnetales and angiosperms are sister groups: (Gy(Gn, Ag)). This scenario is supported by cladistic analyses based on morphological characters. Depending on the genes investigated, molecular analyses support either scenario B, where the Gnetales diverged earlier than the other lineages, (Gn(Gy, Ag)), or most generally support scenario C, where the Gnetales are nested within the gymnosperms. These three scenarios are classically referred to as the Anthophyte, the Gnetale, and the Gymnosperm hypotheses, respectively.

The *psaA* and *psbB* genes exhibit striking conflict among codon positions, where first and second positions strongly support scenario C, but third positions strongly support scenario B (Sanderson et al., 2000). Recombination and horizontal gene transfer can be ruled out as an explanation for this pattern, so that the different gene partitions must have evolved under different processes (Huelsenbeck and Bull, 1996). The original authors indicated that base composition heterogeneity and saturation at the third codon positions were likely to cause ML methods to converge to an attract tree, whereas substantial among-site rate variation led to unresolved and poorly supported results using LogDet distance methods (Sanderson et al., 2000). Accommodating among-site rate variation is more important than accommodating the other parameter. To this effect, Steel et al. (1994) recommended that the LogDet method be used with invariant sites (LogDet + I).

Of the species originally analyzed by Sanderson et al. (2000), I considered only the 19 taxa common to both genes. Given the codon positions of the two genes, four data partitions were distinguished: (1) the first and second codon positions (CP12) of *psaA* (1,510 bp), (2) the third codon positions (CP3) of *psaA* (755 bp), (3) CP12 of *psbB* (1,020 bp), and (4) CP3 of *psbB* (510 bp). Among the different possibilities, two types of analyses were performed. First, separate analyses on these four partitions were conducted under two contrasted substitution models: JC69 (Jukes and Cantor, 1969) and REV + Γ (Tavaré, 1986; Yang, 1994). This latter model is also known as the general time reversible model (GTR) + Γ. Second, combined analyses were performed without or with partitioned likelihoods, where branch lengths across partitions are proportional but all the other parameters are distinct (Yang, 1996). Two substitution models again were assumed, either JC69 or REV + Γ. These two substitution models were chosen a priori because the objective here is not to select which one explains the data better (e.g., Huelsenbeck and Crandall, 1997)

to optimize the bias–variance tradeoff (Burnham and Anderson, 1998:21) but rather to show the effect of too simplistic a model of molecular evolution (bias) on tree topology selection procedures.

Selection of the Topologies Included in the Analysis

Because it is not clear how many and which topologies should be included in an analysis comparing different branching processes (Buckley et al., 2002), I adopted the following selection procedure. Topology searches were performed in a Bayesian framework with `Bambe` 1.02 (Larget and Simon, 1999) under the JC69 and HKY85 + Γ (Hasegawa et al., 1985) nucleotide substitution models. Nuisance parameters (branch lengths and for HKY85 + Γ the base frequencies π, transition to transversion rate ratio κ, and shape parameter α of the Γ distribution modeling among-site rate variation) were marginalized over noninformative prior distributions as described by Larget and Simon (1999). Each Bayesian analysis was performed on the concatenated sequences of *psaA* and *psbB* (no partitions). Four independent chains starting from different initial values were run under each substitution model to check convergence, first by verifying homogeneity of the posterior estimates across the four chains and then by monitoring time series plots of the parameters integrated over the chain and the likelihood sampled from the posterior distribution (e.g., Aris-Brosou and Yang, 2002). Each Markov chain was 10^{14} generations long, and states were sampled every 100 steps (thinning). The first 10^7 steps were discarded as a burn-in to allow the chain to reach stationarity. Posterior probabilities of tree topologies were then computed with the program `Summarize` from `Bambe` (Larget and Simon, 1999). Topologies sampled at stationarity under each substitution model with a frequency >1% were included in the analysis.

Simulations and Power Analysis

Simulations were performed to compare the power of the new tests (FST and FHT) with those of previously described procedures (SH, WSH, BP, and AU). The different simulations also were used to compare the sensitivity of the different tests to the assumed substitution model. Because the Markov models only approximate the actual process of molecular evolution, the generating model used for the simulations was chosen to be complicated, a procedure advocated in the statistical literature (e.g., Burnham and Anderson, 1998:121). Four classes of data sets, corresponding to the four gene partitions (*psaA* CP12 and CP3 and *psbB* CP12 and CP3), were simulated under the REV + Γ nucleotide substitution model. Model parameters were set to the ML estimates obtained under REV + Γ from the independent analyses of the original four gene partitions. For each of these four sets of simulations, 500 data sets were generated using `Evolver` (Yang, 1997).

The simulated data sets were analyzed under two contrasted substitution models: JC69 and the generating model, REV + Γ. Sitewise log-likelihood values were computed with a program based on `Baseml` (Yang, 1997), modified to include parts of the code of `Consel` (Shimodaira and Hasegawa, 2001), used with default settings, to compute the P-values of the WSH and AU tests.

To compare trees and order them on a simple scale in the power analysis, a distance should be defined, if possible based on asymptotically sufficient statistics. Because in a simulation study the generating model is known, the KL distance is more appropriate than other distances such as that of Robinson and Foulds (1981). The KL distance $d(f, g)$ between the generating model f and the approximating model g is defined in Equations 8 and 9. Herein after, I use the relative estimated distance between f and g:

$$\begin{aligned} D_{KL} &= E[\log f(x \mid \hat{\theta})] - E[\log g(x \mid \hat{\theta})] \\ &= E[\ell(\hat{\theta}_f, T_f \mid X)] - E[\ell(\hat{\theta}_g, T_g \mid X)]. \end{aligned} \quad (13)$$

Expectations are taken with respect to the empirical sampling distribution, i.e., over the 500 replicates. The power of the different tests is presented as a function of this distance. Power at level $\alpha = 5\%$ was estimated from the simulations as the proportion of P-values less than α. Coverage probability is closely related to power because it is the probability that the generating model be included in the confidence set 1 – Power($D_{KL} = 0$). The objective of a test is usually that its coverage probability is greater than $1 - \alpha$.

RESULTS

Bayesian Selection of Topologies

Depending on the model of evolution assumed during the Bayesian analyses, the Markov chains converged to different stationary distributions over the topology space. Under JC69, only two topologies were sampled at stationarity: $p(T_1 \mid X) = 0.71$ and $p(T_2 \mid X) = 0.29$. These topologies, labeled as in Figure 1, both support scenario B. Under a more complicated substitution model (HKY85 + Γ), the number of parameters is increased so that the variance of the estimates is also increased. As a result, the Markov chain Monte Carlo algorithm is then expected to sample more topologies at stationarity. However, the chains converged to a different set of topologies than obtained under the simpler JC69 model. Nine topologies were sampled with a frequency >1%. Numbered T_3 to T_{11} (Fig. 1), their respective posterior probabilities were estimated as $p(T_3 \mid X) = 0.23$, $p(T_4 \mid X) = 0.19$, $p(T_5 \mid X) = 0.11$, $p(T_6 \mid X) = 0.10$, $p(T_7 \mid X) = 0.06$, $p(T_8 \mid X) = 0.05$, $p(T_9 \mid X) = 0.04$, $p(T_{10} \mid X) = 0.01$, and $p(T_{11} \mid X) = 0.01$. Unlike analyses under JC69, all the analyses under HKY85 + Γ support scenario C. Specification of the model of evolution has an important effect on (Bayesian) selection procedures, yet no analysis showed any support for scenario A. To include this

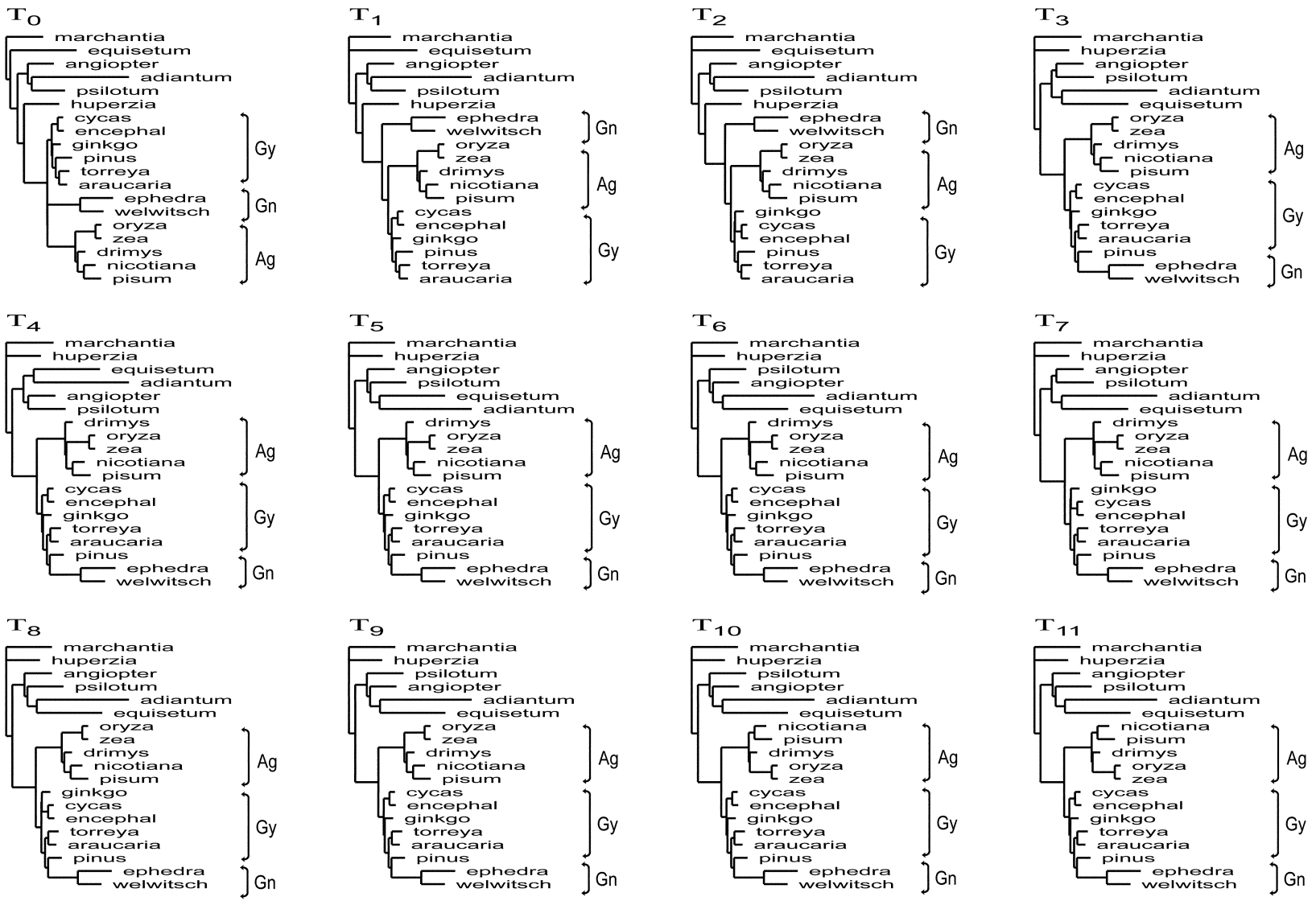

FIGURE 1. The 12 trees included in the analysis. Branch lengths were estimated by ML under the REV + Γ substitution model. Three lineages of seed plants are indicated: Ag = Angiosperms; Gn = Gnetales; Gy = Gymnosperms. The relative positions of these lineages sustain one of the three biological scenarios: A (Anthophyte: T_0), B (Gnetale: T_1 and T_2), or C (Gymnosperm: T_3–T_{11}).

scenario in the power analyses below, topology T_0 (Fig. 1) was randomly selected.

Sensitivity of Phylogenetic Tests to Specification of the Substitution Model

Separate analyses.—Under a simple substitution model (JC69), each partition strongly rejects one or two scenarios (α level preset at 5%), but different codon positions reject different sets of scenarios (Table 1). The partition grouping first and second codon positions strongly rejects scenarios A and B, but third codon positions strongly reject scenario C. This result is consistent with that of Sanderson et al. (2000), who obtained similar results under a more complicated model (HKY85 + Γ). Because of the absence of recombination among codon positions, only one scenario should be correct. In this case, an ideal test should either reject the same scenarios or, if the model is terribly misspecified, fail to reject any scenario at a given level. This is not the result obtained, so that all tests are overconfident in choosing the possibly wrong topologies.

Under a more complex model such as REV + Γ, the selection procedures consistently chose scenario C: Scenarios A and B (T_0–T_2) are rejected by all the tests for all codon partitions, with the exception of the third codon positions of *psbB* (Table 2). This partition still exhibits conflicting results with respect to the other partitions and among different tests. None of the three scenarios tested can be rejected by the FST and the SH and AU tests, but BP and the FHT both give strong support to scenarios A and B. This result may be caused by particular lineage effects at third codon position, such as heterogeneous π, not taken into account by the time-homogeneous REV + Γ substitution model. However, in this case, where REV + Γ fits the data poorly, the selection procedures at the least favorable distribution (FST and SH test) and the AU test appear safe to use because they all fail to reject all the possible scenarios.

For this specific data set, the use of more complex substitution models dramatically affected confidence in a given tree or in a given scenario. However, consistent behavior of the tests is apparent. Regardless of the substitution model, FST and the SH test on the one hand and FHT, the AU test, and the BP on the other hand led to the construction of almost the same confidence sets. Nevertheless, some differences exist, and for the data sets examined here, FST is the most conservative test, and FHT is the most liberal.

Combined analyses.—The joint influence of specification of the substitution model and data partitioning

TABLE 1. *P*-values for the separate analysis of the four gene partitions under JC69. Nonsignificant values ($\alpha = 5\%$) are underlined.

	psaA CP12					*psbB* CP12					*psaA* CP3					*psbB* CP3				
Tree[a]	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT
T_0 (A)	0.008	0.000	0.006	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.284	0.999	0.005	0.002	0.000	0.079	0.999	0.002	0.000	0.000
T_1 (B)	0.031	0.000	0.027	0.009	0.000	0.001	0.000	0.000	0.000	0.000	0.836	0.999	0.460	0.440	0.373	0.844	0.999	0.476	0.480	0.430
T_2 (B)	0.029	0.000	0.012	0.005	0.000	0.002	0.000	0.000	0.000	0.000	—	—	0.583	0.557	—	—	—	0.539	0.519	—
T_3 (C)	0.725	0.999	0.324	0.162	0.000	0.581	0.999	0.188	0.053	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_4 (C)	0.414	0.999	0.052	0.004	0.000	0.636	0.999	0.089	0.046	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_5 (C)	0.342	0.999	0.003	0.000	0.000	0.878	0.999	0.616	0.318	0.373	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_6 (C)	0.603	0.999	0.032	0.014	0.000	0.497	0.999	0.070	0.013	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_7 (C)	0.303	0.994	0.009	0.001	0.000	0.747	0.999	0.460	0.157	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_8 (C)	0.529	0.999	0.095	0.042	0.000	—	—	0.692	0.399	—	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_9 (C)	—	—	0.928	0.656	—	0.450	0.999	0.075	0.009	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_{10} (C)	0.668	0.999	0.293	0.104	0.000	0.292	0.999	0.001	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
T_{11} (C)	0.384	0.999	0.025	0.002	0.000	0.672	0.999	0.005	0.006	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000

[a]Letters in parentheses are the scenarios supported : A = Anthophyte hypothesis; B = Gnetale hypothesis; C = Gymnosperm hypothesis.

TABLE 2. *P*-values for the separate analysis of the four gene partitions under REV + Γ. Nonsignificant values ($\alpha = 5\%$) are underlined.

	psaA CP12					*psbB* CP12					*psaA* CP3					*psbB* CP3				
Tree[a]	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT
T_0 (A)	0.017	0.000	0.008	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.028	0.000	0.014	0.001	0.000	0.719	0.993	0.281	0.084	0.109
T_1 (B)	0.041	0.000	0.032	0.016	0.000	0.001	0.000	0.000	0.000	0.000	0.041	0.000	0.035	0.008	0.000	—	—	0.712	0.339	—
T_2 (B)	0.041	0.000	0.032	0.006	0.000	0.002	0.000	0.000	0.000	0.000	0.036	0.000	0.013	0.003	0.000	0.804	0.999	0.602	0.368	0.396
T_3 (C)	0.551	0.999	0.283	0.058	0.007	0.542	0.999	0.155	0.050	0.000	—	—	0.732	0.354	—	0.218	0.111	0.088	0.002	0.000
T_4 (C)	0.481	0.997	0.137	0.011	0.000	0.656	0.999	0.152	0.065	0.001	0.889	0.999	0.609	0.168	0.390	0.221	0.117	0.162	0.005	0.000
T_5 (C)	0.470	0.997	0.028	0.004	0.000	—	—	0.745	0.398	—	0.698	0.999	0.396	0.080	0.117	0.278	0.266	0.157	0.019	0.000
T_6 (C)	0.824	0.999	0.143	0.080	0.281	0.510	0.999	0.057	0.009	0.000	0.670	0.999	0.271	0.042	0.086	0.197	0.074	0.093	0.001	0.000
T_7 (C)	0.445	0.990	0.077	0.012	0.000	0.726	0.999	0.536	0.183	0.007	0.836	0.999	0.415	0.126	0.329	0.183	0.012	0.070	0.009	0.000
T_8 (C)	0.511	0.998	0.127	0.029	0.000	0.853	0.999	0.541	0.260	0.167	0.746	0.999	0.393	0.125	0.174	0.223	0.050	0.125	0.019	0.000
T_9 (C)	—	—	0.818	0.459	—	0.404	0.999	0.040	0.004	0.000	0.766	0.999	0.343	0.070	0.166	0.196	0.069	0.031	0.001	0.000
T_{10} (C)	0.868	0.999	0.720	0.299	0.394	0.368	0.999	0.004	0.001	0.000	0.572	0.984	0.095	0.006	0.033	0.322	0.348	0.192	0.034	0.000
T_{11} (C)	0.507	0.998	0.154	0.026	0.000	0.744	0.999	0.092	0.030	0.022	0.600	0.993	0.160	0.016	0.051	0.413	0.671	0.419	0.119	0.001

[a]Letters in parentheses are the scenarios supported: A = Anthophyte hypothesis; B = Gnetale hypothesis; C = Gymnosperm hypothesis.

TABLE 3. *P*-values for the combined analysis under different substitution models. Nonsignificant values ($\alpha = 5\%$) are underlined.

	JC69, 1 partition					JC69, 4 partitions					REV + Γ, 1 partition					REV + Γ, 4 partitions				
Tree[a]	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT	SH	FST	AU	BP	FHT
T_0 (A)	0.115	0.999	0.000	0.000	0.000	0.148	0.000	0.000	0.000	0.000	0.008	0.000	0.004	0.000	0.000	0.001	0.000	0.000	0.000	0.000
T_1 (B)	—	—	0.586	0.566	—	—	—	0.636	0.366	—	0.037	0.000	0.027	0.007	0.000	0.003	0.000	0.001	0.000	0.000
T_2 (B)	0.828	0.999	0.445	0.434	0.402	0.838	0.999	0.573	0.337	0.483	0.034	0.000	0.019	0.005	0.000	0.003	0.000	0.001	0.000	0.000
T_3 (C)	0.002	0.060	0.003	0.000	0.000	0.466	0.999	0.470	0.172	0.000	—	—	0.756	0.316	—	—	—	0.693	0.320	—
T_4 (C)	0.000	0.000	0.000	0.000	0.000	0.187	0.039	0.040	0.002	0.000	0.918	0.999	0.570	0.136	0.401	0.812	0.999	0.377	0.068	0.169
T_5 (C)	0.000	0.000	0.001	0.000	0.000	0.123	0.000	0.035	0.001	0.000	0.727	0.992	0.304	0.046	0.135	0.703	0.999	0.309	0.045	0.051
T_6 (C)	0.000	0.000	0.001	0.000	0.000	0.125	0.000	0.041	0.000	0.000	0.746	0.995	0.337	0.046	0.166	0.770	0.999	0.336	0.054	0.145
T_7 (C)	0.000	0.000	0.000	0.000	0.000	0.205	0.126	0.080	0.009	0.000	0.813	0.999	0.301	0.077	0.245	0.702	0.999	0.174	0.036	0.036
T_8 (C)	0.001	0.000	0.001	0.000	0.000	0.358	0.986	0.281	0.076	0.000	0.710	0.991	0.265	0.065	0.113	0.765	0.999	0.310	0.099	0.113
T_9 (C)	0.001	0.000	0.000	0.000	0.000	0.288	0.860	0.179	0.034	0.000	0.815	0.999	0.496	0.117	0.262	0.919	0.999	0.661	0.290	0.501
T_{10} (C)	0.000	0.000	0.000	0.000	0.000	0.131	0.000	0.018	0.001	0.000	0.767	0.995	0.408	0.102	0.207	0.746	0.999	0.252	0.048	0.118
T_{11} (C)	0.000	0.000	0.000	0.000	0.000	0.128	0.000	0.012	0.001	0.000	0.748	0.995	0.389	0.083	0.163	0.683	0.999	0.272	0.040	0.036

[a]Letters in parentheses are the scenarios supported: A = Anthophyte hypothesis; B = Gnetale hypothesis; C = Gymnosperm hypothesis.

(see Yang, 1996) on tests of molecular phylogenies is shown in Table 3. Concatenated sequences under JC69 without taking into account genes and codon position partitions led the SH test and the FST to reject scenario C, consistently with results from the separate analyses. Scenario A is also rejected by the AU test, the BP, and the FHT. However, with the exception of FHT, all the tests fail to reject scenario C for partitioned likelihood analyses, although scenario B is still the most likely.

When a more complex substitution model is considered, the most likely hypothesis is scenario C. All the tests fail to reject scenario B under a simple model (JC69, with the two genes analyzed as one partition or four partitions), but they all reject scenarios A and B under REV + Γ (Table 3). The improvement provided by REV + Γ over JC69 is dramatic ($\delta = 2[(-24104.42) - (-26270.17)] = 4331.50$; $P \ll 0.01$) and, as suggested earlier, is due to several factors such as rate heterogeneities among sites and lineages (Sanderson et al., 2000). Jointly, these results suggest that specifying more realistic models of evolution and accommodating heterogeneity among the data can be important for resolving conflicting results.

Power Analysis of Tree Selection Procedures

Because the reduced null hypotheses under FHT and FST are similar to that of the BP and the SH test, respectively, FHT and BP on the one hand and FST and the SH test on the other hand are expected to have similar power. However, the analysis of the seed plant data showed that the tests examined behave differently in their details, such that FST and the SH test do not in general have identical *P*-values. Results of the power analysis at the 5% level (Fig. 2) suggest that the tests studied here approximately behave according to the two expected patterns: hypothesis tests (BP and FHT) are most powerful, and significance tests (SH, WSH, and FST) are the least powerful (which is why $h_{0,\tau}$ is "least favorable" under these tests). Results from the AU test, a hypothesis test, appear to be intermediate.

The tests also have different sensitivities to the specification of the substitution model (Fig. 2). Under correct specification, the power curves of the significance tests are confounded, and the nine topologies corresponding to scenario C are not significantly different, whichever partition or simulation condition is considered (Figs. 2a–d). But when the substitution model used for the analysis is simpler than the generating model (JC69 vs. REV + Γ), the SH and WSH tests become apparently more powerful than the FST (Figs. 2e, 2f), which results in putting overconfidence in a smaller set of trees. The effect of misspecification is actually to decrease the size of the tests, causing them to reject more hypotheses than they should at a given α level. The size of the hypothesis tests is similarly affected when the substitution model is not correctly specified. The size of the AU test seems to be the most sensitive to specification of the substitution model, whereas that of FHT appears to be the least sensitive (e.g., Fig. 2a vs. Fig. 2e and Fig. 3).

Interestingly, the relationship between misspecification and the size of the tests, or their power as a proxy, is complex. Figure 3 shows that test size can be decreased (inflated power) by misspecification (WSH, AU, and BP for simulations under *psbB* CP12: Figs. 3a, 3b) but can also be increased at high power and decreased at low power (SH, WSH, FST, BP, and AU for simulations under *psbB* CP3: Figs. 3c, 3d). The exact conditions leading to each of these results are beyond the scope of this paper and are not investigated here. Yet, both FST and FHT appear to be the tests least sensitive to specification of the substitution model; their power curves and the resulting confidence set of tree topologies are little affected by a change of the assumed substitution model (e.g., Fig. 2a vs. Fig. 2e and Fig. 3).

Note that the FDR is not controlled for the FHT (power $> \alpha$ for $D_{KL} = 0$). Previously published tests such as BP, KH, or SH are known to lack control of the FDR, in particular when the true configuration is at the least favorable distribution (Shimodaira, 2002). However, this is not the situation simulated here, and the underlying reason for this lack of control is not clear, although the procedure by Westfall and Young (1993)

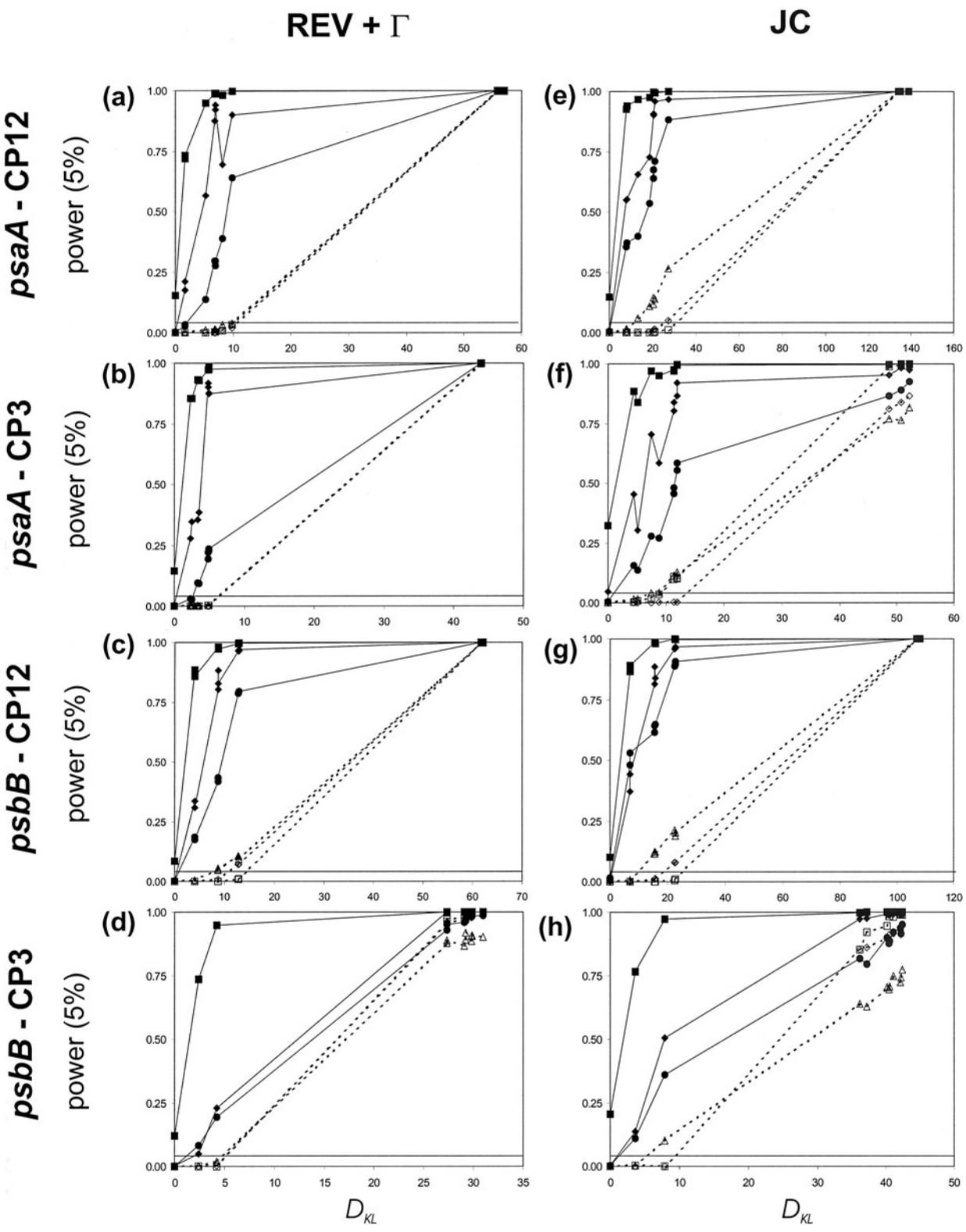

FIGURE 2. Power analysis at the nominal level $\alpha = 5\%$ of the different tree selection procedures. The significance tests (broken lines) are: SH (◇), WSH (△), FST (□); the hypothesis tests (solid lines) are RELL (♦), FHT (■), AU (●). The abscissa, D_{KL}, represents the expected KL distance on a log scale. Data were simulated under REV + Γ for the ML estimates of each partition of the seed plant data set. Analyses were performed under either REV + Γ (a–d) or JC69 (e–f).

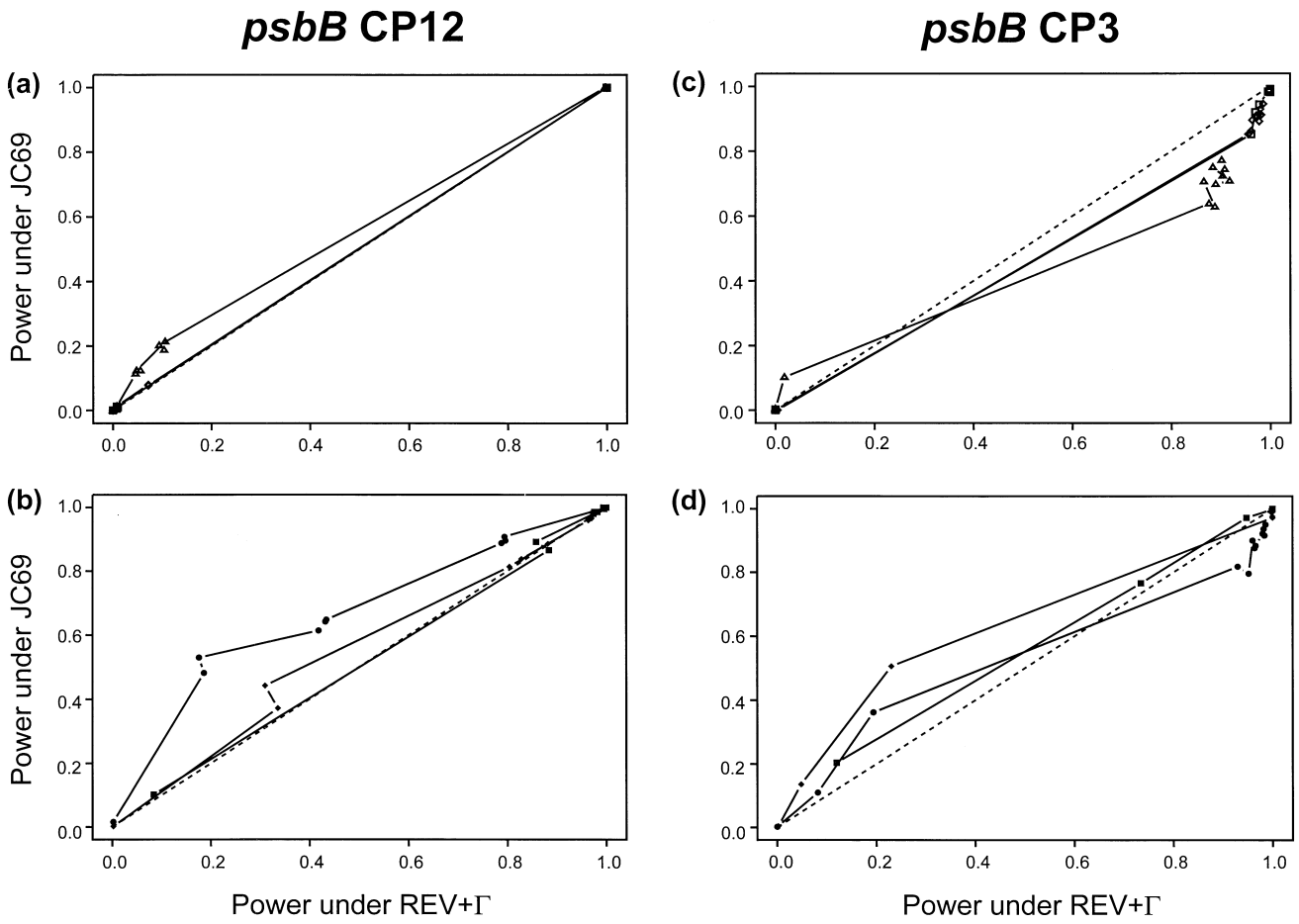

FIGURE 3. Power under the misspecified model of evolution (JC69) as a function of power under the correctly specified model (REV + Γ) for *psbB* CP12 (a, b) and psbB CP3 (c, d) at the nominal level $\alpha = 5\%$ for the tree selection procedures. (a, c) Significance tests: SH (◇), WSH (△), FST (□). (b, d) Hypothesis tests: RELL (♦), FHT (■), AU (●). The broken lines represent the bisector line (no bias).

is known to control FDR only approximately. All the other tests have power at $D_{KL} = 0$ less than the nominal level ($\alpha = 5\%$). Indeed, the distribution of the P-values of, for example, the WSH test for the generating tree is extremely skewed to the right (not shown), with an average at 0.95 and no values <0.48. This excellent coverage occurs in particular when the true configuration is not at the least favorable distribution, i.e., when a few trees are much better than the other ones (Shimodaira, 2002), which is the situation simulated here.

DISCUSSION AND CONCLUSIONS

Because the tree topology is not a proper statistical parameter (Yang et al., 1995), deciding whether some prespecified topologies are significantly different is difficult. The phylogenetic literature already contains a number of methods for assessing phylogenies. The choice is all the more confusing because these methods generally give different P-values (Goldman et al., 2000; Whelan et al., 2001). Goldman et al. (2000) suggested that these different results may be due to different forms of the null hypotheses and/or to the larger power of parametric tests and their stricter dependence on a substitution model.

The two tests presented here are both based on information theoretic arguments and only differ by how their common composite null hypothesis is reduced to their respective simple null hypotheses. All other specificities (estimation of the distribution of their test statistics under H_0 by nonparametric bootstrap or P-value adjustments for multiple comparisons) are identical in both tests. The results obtained here, in particular with respect to the power of these tests, suggest that the exact form of the null hypothesis explains an important part of the difference between different tests. Indeed, the tests described are either very conservative (FST) or very powerful (FHT). This distinction is also supported by a comparison of tests similarly constructed in a Bayesian framework (Aris-Brosou, 2003).

Specifically, the reduction of the composite null hypothesis (that the k prespecified topologies are equidistant from the true topology) to a simple hypothesis at the least favorable distribution leads to tests such as the FST or the SH test, which are least powerful (Lehmann, 1986:104). Because H_0 at the least favorable distribution

is tested against each topology, these tests are distinct from pure significance tests as described by Steel et al. (1995) under a parsimony approach and by Bar-Hen and Kishino (2000) under a likelihood approach. The aim of the latter tests is to test whether there is enough information in the data to reconstruct a phylogenetic tree, whereas the FST and the SH test aim at constructing a confidence set of topologies close to the true topology at the least favorable distribution for a given significance level. Because estimation of the least favorable distribution depends on which topologies are considered, both the FST and the SH test are sensitive to the number of trees included in the analysis. However, although both tests have equivalent null hypotheses and correct for selection bias, they differ in their reliance on a different bootstrapping approach. The FST corrects the statistics s_i^* to conform to the null hypothesis (Hall and Wilson, 1991), whereas the SH test corrects the resampling strategy to conform to the null hypothesis (see Tibshirani, 1992; Nick Goldman, pers. com., 2002). Although the FST intrinsically adjusts P-values for multiple comparisons, its power is similar to that of the SH and WSH tests, at least when the substitution model is correctly specified. Thus, the P-value adjustments, which result in an increased computational burden, may not be justified for these tests. The significance level of tests at the least favorable distribution is indeed implicitly adjusted (e.g., Shimodaira, 1998).

Alternatively, when the composite null hypothesis is reduced to the density under T_{ML}, the test obtained, the FHT, is equivalent to the BP and aims at deciding which topology is correct. Posterior probabilities and the Bayesian version of the FHT also answer the same question (Aris-Brosou, 2003). The power gained over the BP highlights the importance of correcting for multiple comparisons in the case of paired hypothesis tests (Westfall and Young, 1993). However, the quest for tests that are more powerful (e.g., Lehmann, 1986:72) may be questioned, in particular when greater power is obtained at the expense of enlarging the rejection region for trees unduly close to the null hypothesis (e.g., Perlman and Wu, 1999). Phylogeneticists are faced by multifaceted questions for which the pairwise structure of such hypothesis tests may not be adequate.

Taking the geometry of the sample space into account may help obtain more accurate assessments (Efron et al., 1996; Shimodaira, 2002). Although the AU test proved here to be somewhat sensitive to model specification in some cases (Fig. 3), it turned out to be much safer than any other hypothesis test. As noted previously (Buckley, 2002), the relationship between power and misspecification is complex. The simulations carried out here show that when the data are generated and analyzed under homogeneous Markov models, the size of the tests can be affected by model specification.

More generally, a possible difficulty with frequentist approaches to comparing topologies is their treatment of nuisance parameters and, as shown here, their dependence on an estimated log likelihood, namely $\ell(\hat{\theta}_i, T_i \mid X)$, where the uncertainty about the nuisance parameters θ is disregarded. By marginalizing nuisance parameters to actually compare $p(X \mid T_i) = \int_\Theta p(X \mid \theta_i, T_i)\ p(\theta \mid T_i)\ d\theta$ for different T_i (the same prior distribution is set for the different θ_i), the Bayesian approach offers a more natural solution to the problem of topology, independent of θ (e.g., Aris-Brosou, 2003). On the other hand, computing $p(X \mid T_i)$ is challenging, and the estimator based on the harmonic mean of the likelihoods (or $e^{\ell(\hat{\theta}_i, T_i \mid X)}$) sampled from the posterior distribution, although attractive for its relative computational ease, may be unstable (e.g., Raftery, 1996; Aris-Brosou, 2003). Alternative estimators (Chib and Jeliazkov, 2001) or measures such as the deviance information criterion (Spiegelhalter et al., 2002) should be evaluated. However, integrating θ out makes selection procedures more dependent on a specific model of molecular evolution, because the model itself is not integrated out. This may explain the greater sensitivity of Bayes topology selection procedures to the assumed model (e.g., Aris-Brosou, 2003). Bayesian model averaging (Hoeting et al., 1999) may be an attractive solution, but it demands that realistic models be built.

The dependence on simple models of evolution appears as a major difficulty in phylogenetic inferences, because (1) liberal selection procedures may give overconfidence in a small set of topologies and (2) misspecification of the substitution model may decrease the size of the test. Although these problems may not appear important for simulated data (e.g., Fig. 2), such a reduction of confidence sets may be problematic when the assumed model of evolution is so wrong that the method becomes inconsistent or "positively misleading" and converges to an attract topology. This risk has been emphasized by previous studies (e.g., Steel et al., 1993; Huelsenbeck et al., 1996; Buckley, 2002). Here in particular, the choice of too simple a model to estimate the origin of the seed plants led with high confidence to a conclusion similar to that supported by morphological data, but this conclusion stands in sharp contrast with that suggested by the use of more complex models of evolution. In the case of *psaA* and *psbB*, partitioning the likelihood (Yang, 1996) over data that proved heterogeneous (high among-site rate variation, different transition to transversion rate ratios and base frequencies; see Sanderson et al., 2000) led to more consistent results across genes and codon partitions. While heterogeneity across partitions may result in incorrect estimations (Bull et al., 1993; Steel et al., 1993), partitioning the likelihood by keeping branch lengths proportional across partitions led here to consistent estimation. This result, although intellectually satisfying, may suggest that partitioning the likelihood improved on merely combining data in the case of the seed plants studied here, which may be a general result (Pupko et al., 2002). Such heterogeneous models, recently implemented in a Bayesian framework (Ronquist and Huelsenbeck, 2003), are likely to facilitate the analyses of large data sets and improve on their efficiency. However, factors other than heterogeneity over the data exist, as suggested by the unresolved discrepancy at the third codon positions of *psbB*. One general possibility is heterogeneity over lineages (see

Yang and Roberts, 1995; Huelsenbeck and Bull, 1996), such as variable rates (κ, α) or nucleotide compositions, resulting in branch attractions when homogeneous models are used to estimate the phylogeny (e.g., Steel et al., 1993).

In order to avoid putting overconfidence in a small and possibly spurious set of topologies based on the results of any of the tests discussed, more effort should probably be directed toward improving analysis schemes and current models of evolution (Huelsenbeck and Bull, 1996; Ronquist and Huelsenbeck, 2003). In the meantime, conservative approaches where the null hypothesis is taken at the least favorable distribution (Shimodaira and Hasegawa, 1999) may be safer (Shimodaira, 2002; Aris-Brosou, 2003), even if they are the least powerful.

ACKNOWLEDGMENTS

I am indebted to Nick Goldman, who pointed out the correct way to compute test statistics during the resampling step, and to Hirohisa Kishino, who suggested Figure 3. I also thank David Balding, Joseph Bielawski, Lounès Chikhi, Elizabeth Thompson, Jeffrey Thorne, and Ziheng Yang for invaluable discussions and Peter Lockhart, Chris Simon, Peter Waddell, and two anonymous reviewers for constructive comments. This work was funded by a Biotechnological and Biological Sciences Research Council grant to Ziheng Yang, National Science Foundation grant DEB-0120635 to Jeffrey Thorne, and a Japanese Science & Technology Corporation grant to Hirohisa Kishino.

REFERENCES

ARIS-BROSOU, S. 2003. How Bayes tests of molecular phylogenies compare with frequentist approaches. Bioinformatics 19:618–624.

ARIS-BROSOU, S., AND Z. YANG. 2002. The effects of models of rate evolution on estimation of divergence dates with a special reference to the metazoan 18S rRNA phylogeny. Syst. Biol. 51:703–714.

BAR-HEN, A., AND H. KISHINO. 2000. Comparing the likelihood functions of phylogenetic trees. Ann. Inst. Stat. Math. 52:43–56.

BUCKLEY, T. R. 2002. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. Syst. Biol. 51:509–523.

BUCKLEY, T. R., P. ARENSBURGER, C. SIMON, AND G. K. CHAMBERS. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.

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

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD, AND P. J. WADDELL. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

BURNHAM, K. P., AND D. R. ANDERSON. 1998. Model selection and inference. A practical information-theoretic approach. Springer, New York.

CAO, Y., M. FUJIWARA, M. NIKAIDO, N. OKADA, AND M. HASEGAWA. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene 259:149–158.

CHIB, S., AND I. JELIAZKOV. 2001. Marginal likelihood from the Metropolis–Hastings output. J. Am. Stat. Assoc. 96:270–281.

DONOGHUE, M. J. 1994. Progress and prospects in reconstructing plant phylogeny. Ann. Mo. Bot. Gard. 81:405–418.

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., AND H. KISHINO. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42:193–200.

GOLDMAN, N., J. P. ANDERSON, AND A. G. RODRIGO. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

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 estimation of the hominoid tree for mitochondrial DNA sequences. Evolution 43:672–677.

HASEGAWA, M., H. KISHINO, AND N. SAITOU. 1991. On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. 32:443–445.

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.

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

HOETING, J. A., D. MADIGAN, A. E. RAFTERY, AND C. T. VOLINSKY. 1999. Bayesian model averaging: A tutorial. Stat. Sci. 14:382–417.

HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.

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., D. M. HILLIS, AND R. NIELSEN. 1996. A likelihood-ratio test of monophyly. Syst. Biol. 45:544–556.

HUELSENBECK, J. P., AND N. S. IMENNOV. 2002. Geographic origin of human mitochondrial DNA: Accommodating phylogenetic uncertainty and model comparison. Syst. Biol. 51:155–165.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–32 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

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., AND M. HASEGAWA. 1990. Converting distance to time: Application to human evolution. Methods Enzymol. 183:550–570.

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

KULLBACK, S., AND R. A. LEIBLER. 1951. On information and sufficiency. Ann. Math. Stat. 22:79–86.

LARGET, B., AND D. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

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

NEWTON, M. A. 1996. Bootstrapping phylogenies: Large deviations and dispersion effects. Biometrika 83:315–328.

PERLMAN, M. D., AND L. WU. 1999. The Emperor's new tests. Stat. Sci. 14:355–381.

PUPKO, T., D. HUCHON, Y. CAO, N. OKADA, AND M. HASEGAWA. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.

RAFTERY, A. E. 1996. Hypothesis testing and model selection Pages 163–187 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, Boca Raton, Florida

ROBINSON, D. F., AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

RONQUIST, F., AND J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

SANDERSON, M. J., M. F. WOJCIECHOWSKI, J. HU, T. S. KHAN, AND S. G. BRADY. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17:782–797.

SHAO, J., AND D. TU. 1996. The jackknife and the bootstrap. Springer, New York.

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

SHIMODAIRA, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

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

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

SPIEGELHALTER, D. J., N. G. BEST, B. P. CARLIN, AND A. M. VAN DER LINDEN. 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B 64:583–639.

STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1995. A frequency-dependent significance test for parsimony. Mol. Phylogenet. Evol. 4:64–71.

STEEL, M. A., L. A. SZÉKELY, AND M. D. HENDY. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comput. Biol. 1:153–163.

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

SWOFFORD, D. L., G. J. OLSEN, P. G. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

TIBSHIRANI, R. 1992. Bootstrap hypothesis testing. Biometrics 48:969–970.

WATERMAN, M. S. 1995. Introduction to computational biology. Chapman and Hall, London.

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

WHELAN, S., P. LIO, AND N. GOLDMAN. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 17:262–272.

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. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

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

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, N. GOLDMAN, AND A. M. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

YANG, Z., AND D. ROBERTS. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451–458.

YODER, A. D., AND Z. YANG. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

First submitted 26 November 2002; reviews returned 6 May 2003; final acceptance 30 June 2003
Associate Editor: Peter Lockhart