

Selecting the Best-Fit Model of Nucleotide Substitution

DAVID POSADA AND KEITH A. CRANDALL

Department of Zoology, Brigham Young University, Provo, Utah 84602-5255, USA; E-mail: dp47@email.byu.edu
(author for correspondence), keith_crandall@byu.edu

Abstract.—Despite the relevant role of models of nucleotide substitution in phylogenetics, choosing among different models remains a problem. Several statistical methods for selecting the model that best fits the data at hand have been proposed, but their absolute and relative performance has not yet been characterized. In this study, we compare under various conditions the performance of different hierarchical and dynamic likelihood ratio tests, and of Akaike and Bayesian information methods, for selecting best-fit models of nucleotide substitution. We specifically examine the role of the topology used to estimate the likelihood of the different models and the importance of the order in which hypotheses are tested. We do this by simulating DNA sequences under a known model of nucleotide substitution and recording how often this true model is recovered by the different methods. Our results suggest that model selection is reasonably accurate and indicate that some likelihood ratio test methods perform overall better than the Akaike or Bayesian information criteria. The tree used to estimate the likelihood scores does not influence model selection unless it is a randomly chosen tree. The order in which hypotheses are tested, and the complexity of the initial model in the sequence of tests, influence model selection in some cases. Model fitting in phylogenetics has been suggested for many years, yet many authors still arbitrarily choose their models, often using the default models implemented in standard computer programs for phylogenetic estimation. We show here that a best-fit model can be readily identified. Consequently, given the relevance of models, model fitting should be routine in any phylogenetic analysis that uses models of evolution. [AIC; BIC; dynamic LRT; hierarchical LRT; likelihood ratio tests; model selection; substitution models.]

Phylogenetic reconstruction has been regarded as a problem of statistical inference since the pioneering work of Edwards and Cavalli-Sforza (1964). Because statistical inferences cannot be drawn in the absence of a probability model, the use of a model of nucleotide substitution—a model of evolution—becomes necessary when using DNA sequences to estimate phylogenetic relationships among organisms. Models of evolution, sets of assumptions about the process of nucleotide substitution (Fig. 1), are used in phylogenetic analyses to describe the different probabilities of change from one nucleotide to another, with the aim of correcting for unseen changes along the phylogeny. Whereas maximum parsimony implicitly assumes a model of evolution (Farris, 1973; Felsenstein, 1973; Yang, 1996a; Steel and Penny, 2000), distance and maximum likelihood methods estimate parameters according to an explicit model of evolution. However, whereas distance methods estimate only a single parameter (substitutions per site) given the model, maximum likelihood can estimate all the relevant parameters of the substitution model.

Although for the last 30 years an array of models of increasing complexity regarding nucleotide substitution have been described

(see Swofford et al., 1996), choosing among models remains a major problem in phylogenetic reconstruction (Cunningham et al., 1998). As is well established, the use of one model of evolution or another may change the results of an analysis (Leitner et al., 1997; Sullivan and Swofford, 1997; Cunningham et al., 1998; Kelsey et al., 1999). Especially, estimates of branch length or bootstrap support can be severely affected (Yang et al., 1994; Buckley et al., 2000). In general, phylogenetic methods may be less accurate (recover an incorrect tree more often) or may be inconsistent (converge to an incorrect tree with increased amounts of data) when the wrong model of evolution is assumed (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Penny et al., 1994; Bruno and Halpern, 1999). Because the performance of a method is maximized when its assumptions are satisfied, some indication of the fit of the data to the phylogenetic model is necessary (Huelsenbeck, 1995). Indeed, model selection is not important just because of its consequences in phylogenetic analysis, but because the characterization of the evolutionary process at the sequence level is itself a legitimate pursuit. Moreover, models of evolution are especially critical for estimating substitution

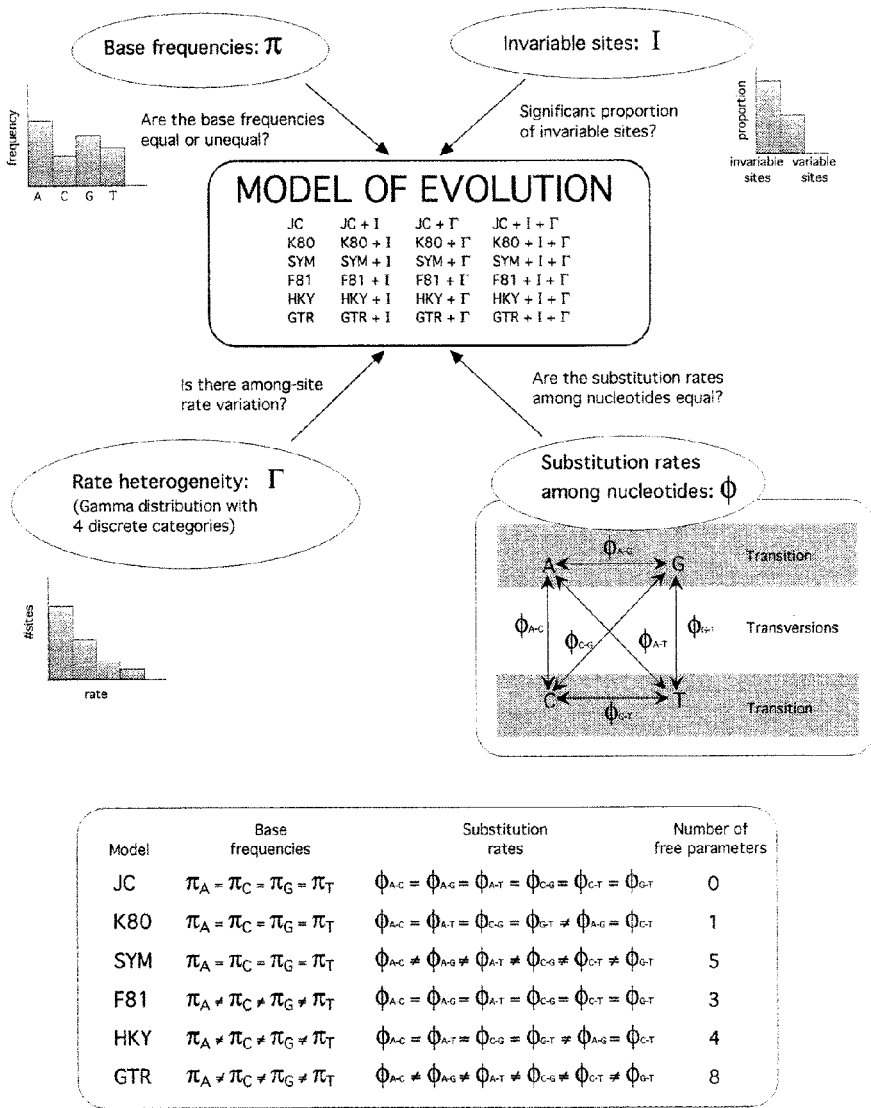


FIGURE 1. A comparison of models of nucleotide substitution. Model selection methods selected the best-fit model for the data set at hand among 24 possible models. See Table 1 footnote for explanation of acronyms for methods.

parameters or for hypothesis testing (Tamura, 1992; Wakeley, 1994; Adachi and Hasegawa, 1995; Yang et al., 1995; Zhang, 1999).

Unfortunately, and despite these conclusions, the unjustified use of models of evolution is still a common practice in phylogenetic studies. In a quick examination of 13 issues of *Systematic Biology* (March 1997 to March 2000), 30 empirical articles used models of evolution in the analyses, of which the model of nucleotide substitution

implemented was statistically justified in only 6 of those studies (20%). If the model of evolution may influence the results of the analysis, then the use of a particular model should be justified. An a priori attractive selection procedure would be the arbitrary use of complex, parameter-rich models, but this approach has several disadvantages: (1) a large number of parameters need to be estimated, making the analysis computationally difficult and requiring a large amount of time, and (2) as more parameters need to

be estimated, more error is included in each estimate (Huelsenbeck and Crandall, 1997). Ideally, we would like to incorporate as much complexity as needed.

The best-fit model of evolution for a particular data set can be selected through statistical testing. Statistical tests of models of nucleotide substitution are of two types: some tests are designed to compare two different models, others to test the overall adequacy of a particular model. In this study we are interested only in the first class of tests, that is, how to select for the best-fit model for the specific data set at hand for a given set of alternative models. The likelihood ratio test statistic (LRT) has been suggested for comparing two models of evolution (Felsenstein, 1981, 1988; Goldman, 1993). Several LRTs can be performed hierarchically (η LRT) to select the simplest model that best explains the data among a set of possible models (Yang et al., 1994; Frati et al., 1997; Huelsenbeck and Crandall, 1997; Sullivan et al., 1997; Posada and Crandall, 1998). Rzhetsky and Nei (1995) developed several tests, using linear invariants to assess the applicability of a particular model to the data. They tested whether the deviation from the expected invariant would be significant if the evaluated model

were true. Although these tests do not require the use of an initial phylogeny, and they are independent of evolutionary time, they are model-specific, and currently can be applied to a small set of substitution models. A different approach for model selection is the simultaneous comparison of all competing models through the Akaike information criterion (AIC) (Akaike, 1974) or the Bayesian information criterion (BIC) (Schwarz, 1974). The use of LRTs is more extensive in the phylogenetic literature, however, examples of the use of the AIC to select the best-fit model of nucleotide substitution can be found in Hasegawa et al. (1990a,b), Tamura (1994), and Muse (1999). Morozov et al. (2000) applied the BIC to compare different models of protein evolution.

Although several statistical procedures exist for model selection, the absolute accuracy and relative performance of these procedures are unknown. In this study, we compare the performance of different LRTs with the AIC and BIC model selection procedures under various conditions. We do this by simulating DNA sequences under a known model of nucleotide substitution and recording how often this true model is recovered by the different model-selecting strategies (Fig. 2).

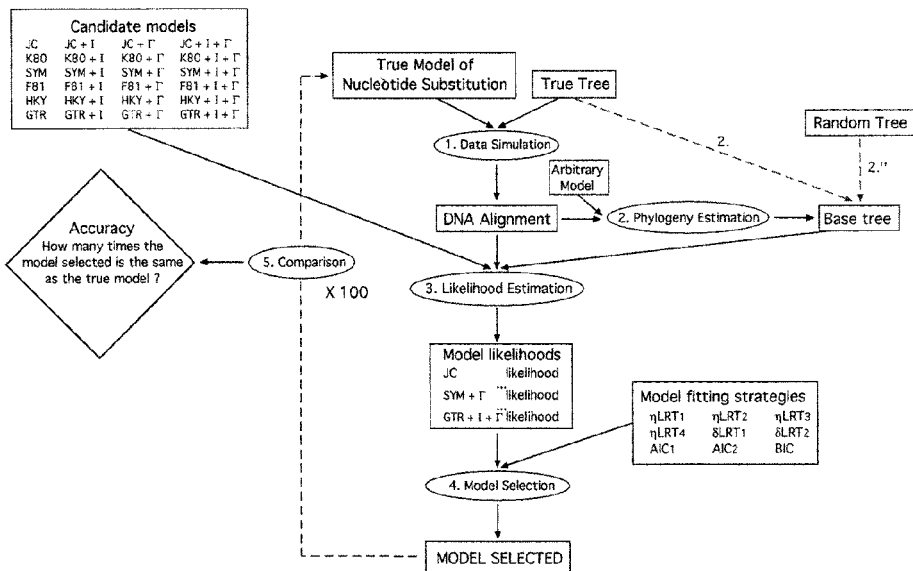


FIGURE 2. Study design. Sequences were simulated under a known model of substitution. The times that a model selection method recovered the true model out of 100 replicates was used as the measure of a method's accuracy. 2'' and 2''' indicate alternative paths to step 2 used in part of the simulations. See text for explanation of model fitting strategies.

METHODS

Data Simulation

Simulations were conducted in several steps (Table 1). An initial, global simulation was carried out to explore the effect of different broad conditions in model selection (Simulation I). Given the results from this simulation, additional simulations were conducted to explore a more restricted but different parameter space. A 20-taxon clocklike tree (Fig. 3a) was used as the model tree in most of the simulations (Simulations I and IV and part of Simulation II). Other clocklike trees, with 10, 50, and 100 taxa were used in part of Simulation II, and a nonclock tree was used in Simulation III (Fig. 3b). Clocklike trees were simulated by a birth–death process with complete taxon sampling (Yang and Rannala, 1997) using the program PAML 2.0g (Yang, 1997a). The birth–death process is a continuous-time process in which the probability that a speciation event occurs along a lineage during an infinitesimal time interval Δt is $\lambda \Delta t$, the probability that an extinction occurs is $\mu \Delta t$, and the probability that two or more events occur is of order $O(\Delta t)$ (Rannala and Yang, 1996). Parameters λ and μ are the branching and extinction rates per lin-

age, respectively. The values used to simulate the trees were $\lambda = 0.1$, $\mu = 0.1$, and sampling fraction = 1.0. We parameterized the rate of substitution as the tree height (m), which is the expected number of substitutions per site for a single lineage from the root to the tip of the tree. To study the influence of the substitution rate in model selection we used various tree heights (0.01, 0.10, 0.20, 0.50, and 0.75) (Table 1). The nonclock tree (Fig. 3b) was obtained by arbitrarily changing the length of some branches of the tree represented in Figure 3a. DNA sequences were simulated over the generated trees by using the program SeqGen 1.1 (Rambaut and Grassly, 1997) according to different models of DNA substitution (Table 2). When appropriate, a gamma shape parameter (α) (Yang, 1993; Yang, 1994a; Yang, 1996b) of 0.5 (0.2 and 0.05 in Simulation IV) was used to model rate variation among sites.

Likelihood Estimation and Base Tree

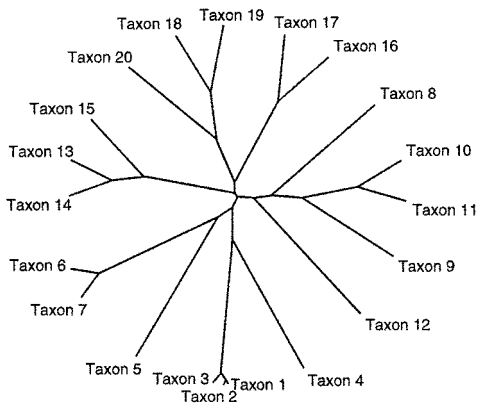
The likelihood of a tree is calculated as the probability of observing the data if the tree is true, under a given model of nucleotide substitution. To estimate the relative fit of different models to a given data set, we can

TABLE 1. Simulations scheme. One hundred data sets were simulated for each set of conditions. True tree is the tree upon which sequences were evolved. Tree height is the expected number of substitutions per site from the root to the tip. Ntaxa is the number of taxa. Nchar is the total number of characters. True model is the model of nucleotide substitution upon which sequences were evolved. Alpha (α) is the shape of the Γ distribution. Ncat is the number of discrete categories for the Γ distribution. Base tree is the tree estimated from the data and on which parameters and likelihood scores were estimated.

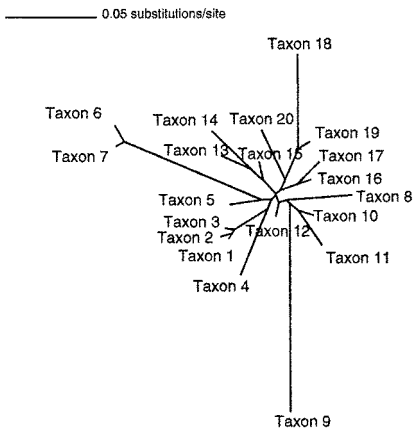
Simulation	True tree	Tree height	Ntaxa	Nchar	True model ^a	α	Ncat	Base tree ^b
I	Clock	0.10	20	100	1–6	0.5	4	1–6, R, T
	Clock	0.10	20	500	1–6	0.5	4	1–6, R, T
	Clock	0.10	20	1,000	1–8	0.5	4	1–6, R, T
II	Clock	0.10	10	500	1–6	0.5	4	3
	Clock	0.10	10	1,000	6	0.5	4	3
	Clock	0.10	50	500	1–6	0.5	4	3
	Clock	0.10	100	500	1–6	0.5	4	3
	Clock	0.01	20	500	1–6	0.5	4	3
	Clock	0.20	20	500	1–6	0.5	4	3
	Clock	0.50	20	500	1–6	0.5	4	3
	Clock	0.75	20	500	1–6	0.5	4	3
	III	Nonclock	—	20	1,000	1–6	0.5	4
IV	Clock	0.10	20	1,000	1–6	0.2	4	1
	Clock	0.10	20	1,000	1–6	0.2	8	1
	Clock	0.10	20	1,000	1–6	0.05	8	1

^aTrue models: JC(1), JC + Γ (2), HKY(3), HKY + Γ (4), GTR(5) GTR + Γ (6). JC: Jukes and Cantor (1969) model; F81: Felsenstein (1981) model; HKY: Hasagawa et al. (1985) model; GTR (also called REV): general time reversible model (Tavaré, 1986). Γ represents the discrete gamma distribution with four rate categories.

^bThe base tree is a NJ tree estimated according to the specified model (1–6)^r, a random tree (R), or the true tree (T).



(a) clock-like tree



(b) non clock-like tree

FIGURE 3. True trees used in the simulations. (a) This clock tree was simulated by a birth–death process with birth rate (λ) = 0.1, death rate (μ) = 0.1, and sampling fraction = 1.0. The height of the tree is 0.10. (b) The non-clock tree was obtained by arbitrarily altering the branch lengths in tree (a). The scale of the branch lengths is indicated.

contrast the likelihoods obtained for a tree estimated from the data (hereafter called the base tree) under the different models compared. The base tree ideally would be the true tree. However, for real data sets, the true tree is unknown and needs to be estimated. To quantify the potential effect of the base tree on model selection, we used eight different base trees: the model tree (= true tree), six neighbor-joining (Saitou and Nei, 1987) (NJ) trees calculated under different arbitrary models of evolution, and a random tree (Table 1). For each set of simulated data and base tree, 24 likelihood scores, corre-

sponding to 24 different models of evolution (Fig. 1), were calculated in PAUP* (Swofford, 1998).

Model Selection Strategies

These likelihood scores were used to select the best-fit model of evolution for each data set and conditions, using nine different methods that could be grouped in four classes:

Hierarchical Likelihood Ratio Tests.—In traditional statistical theory, a widely accepted statistic for testing the goodness of fit of models is the LRT statistic,

$$\delta = 2(\ln L_1 - \ln L_0),$$

where L_1 is the maximum likelihood under the more parameter-rich, complex model (alternative hypothesis) and L_0 is the maximum likelihood under the less parameter-rich simple model (null hypothesis). The value of this statistic is always ≥ 0 , because the likelihood under the more complex model will always be equal or bigger than the likelihood under the simpler model. Simply put, the superfluous parameters in the complex model provide a better explanation of the stochastic variation in the data than the simpler model does, even if the simple model is the true one. When the models compared are nested (the null hypothesis is a special case of the alternative hypothesis) and the null hypothesis is correct, this statistic is asymptotically distributed as χ^2 with q degrees of freedom, where q is the difference in number of free parameters between the two models; equivalently, q is the number of restrictions on the parameters of the alternative hypothesis required to derive the particular case of the null hypothesis (Kendall and Stuart, 1979; Goldman, 1993). To preserve the nesting of the models, the likelihood scores are estimated on the same tree topology. Goldman (1993) questioned the appropriateness of the χ^2 approximation of the LRT statistic when comparing models of evolution, but Yang et al.'s (1995) simulation study suggested that the χ^2 approximation is acceptable in most cases. However, the χ^2 distribution may not be appropriate when the null model is equivalent to fixing some parameters at the boundary of the parameter space of the alternative model. An example of this situation is the rate homogeneity among sites test, where the

TABLE 2. Parameter values used in the simulations. $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ describes base frequencies at equilibrium and has three free parameters, because of the constraint that $\sum \pi_x = 1$. $\varphi = (\varphi_{A-C}, \varphi_{A-G}, \varphi_{A-T}, \varphi_{C-G}, \varphi_{C-T}, \varphi_{G-T})$ describes the substitution rates among bases; it has five free parameters because substitution rates are expressed relative to φ_{G-T} which is set up equal to 1. The parameter κ describes the transition/transversion ratio, a specific constraint on φ , $\kappa = (\varphi_{A-G} = \varphi_{C-T} / \varphi_{A-C} = \varphi_{A-T} = \varphi_{C-G} = \varphi_{G-T})$. The parameter α is the shape parameter of the gamma distribution (Γ), which was simulated with four discrete categories.

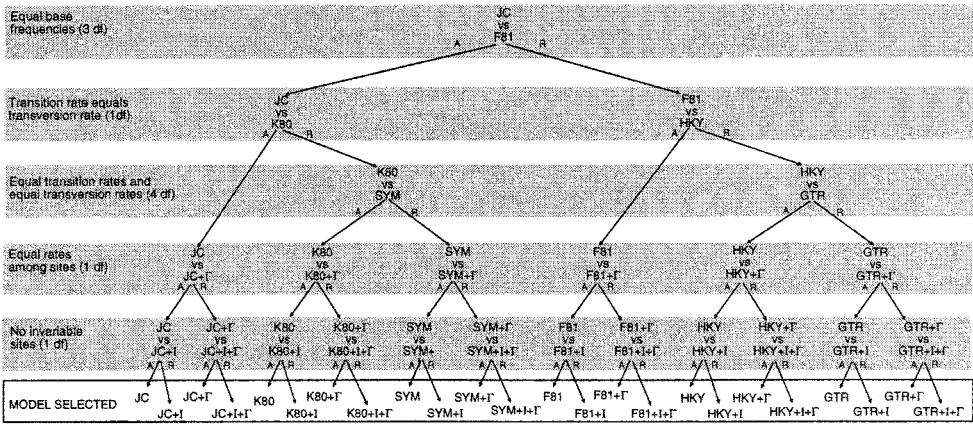
Parameters	JC	JC + Γ	HKY	HKY + Γ	GRT	GTR + Γ
π_A	0.25	0.25	0.35	0.35	0.35	0.35
π_C	0.25	0.25	0.15	0.15	0.15	0.15
π_G	0.25	0.25	0.25	0.25	0.25	0.25
π_T	0.25	0.25	0.25	0.25	0.25	0.25
κ	—	—	2	2	—	—
φ_{A-C}	—	—	—	—	2	2
φ_{A-G}	—	—	—	—	4	4
φ_{A-T}	—	—	—	—	1.8	1.8
φ_{C-G}	—	—	—	—	1.4	1.4
φ_{C-T}	—	—	—	—	6	6
φ_{G-T}	—	—	—	—	1	1
α	—	0.5	—	0.5	—	0.5

See Table 1 footnote for explanation of abbreviations.

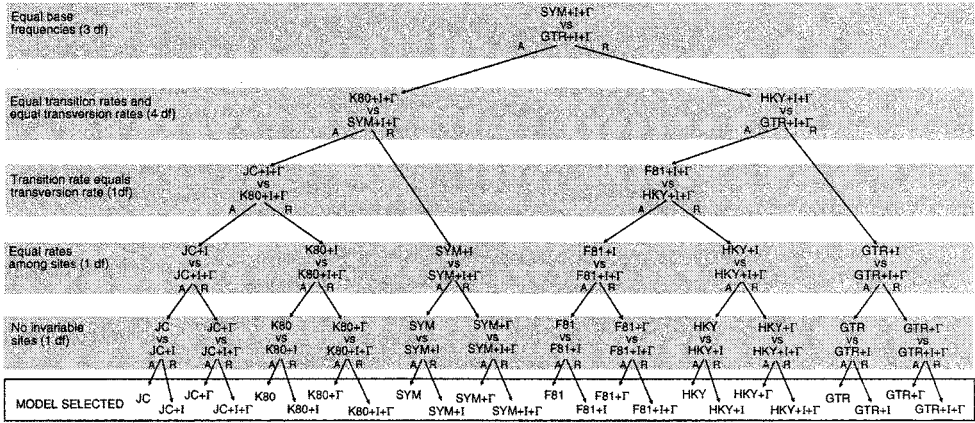
null hypothesis (rate homogeneity) is a special case of the gamma-distribution model (rate heterogeneity), with shape parameter equal to infinity (Yang, 1996c). Whelan and Goldman (1999) concluded that for comparisons of rate variation across sites and nucleotide frequencies estimated as the observed base frequencies, the observed distribution of the LRT statistic was significantly different from the χ^2 distribution. To solve this problem, Ota et al. (2000) and Goldman and Whelan (2000) have suggested instead the use of a mixed χ^2 (or $\bar{\chi}^2$) distribution, consisting of 50% χ_0^2 and 50% χ_1^2 , when a parameter in the null model is fixed at the boundary of its parameter space. On the other side, the difference in likelihood when comparing models may be very large, and the inaccuracy of the χ^2 approximation might not change the results of the tests in these cases. To study this question, here we used both the standard and mixed χ^2 for the “boundary LRTs,” and the standard χ^2 for the other LRTs.

When we compare two different nested models through a LRT, we are actually testing hypotheses about our data. The hypotheses tested are those represented by the difference in the assumptions among the models compared. Several hypotheses can be tested hierarchically to select the best-fit model for the data set at hand (Fрати et al., 1997; Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998). It is to our advantage to test one hypothesis at a time: Are the base frequencies equal? Is there a tran-

sition/transversion bias? Are all transition rates equal? Are there invariable sites, or is there rate homogeneity among sites? and so on. For example, to test the equal base frequencies hypothesis, we could do a LRT comparing JC with F81 (see Table 1 for identification of models), models differing only in the fact that F81 allows for unequal base frequencies (alternative hypothesis), whereas JC assumes equal base frequencies (null hypothesis). However, to test this hypothesis, we could also compare JC + Γ with F81 + Γ , or K80 + I with HKY + I, or SYM with GTR. Which model comparison is used to compare which hypotheses depends on the starting model of the hierarchy and on the order in which different hypotheses are applied. For example, we could start with the simple JC or with the most-complex GTR + I + Γ . In the same way, we could perform first a test for equal base frequencies and later a test for rate heterogeneity among sites, or vice versa. Given that the choice of the best-fit model has been suggested to be affected by the parameter addition sequence (Cunningham et al., 1998), the model selection process might also be dependent on the order in which the LRTs are performed. To test whether the order in which hypotheses are tested influences model selection, we used four different hierarchies to perform the LRTs (η LRT₁ through η LRT₄) (Fig. 4). Indeed, many different hierarchies might be possible, and we also devised a dynamic LRT procedure (δ LRT) (Fig. 5),



(a) ηLRT_1 (JC) $\pi \cdot \kappa \cdot \phi \cdot \Gamma \cdot I$



(b) ηLRT_2 (GTR+I+Γ) $\pi \cdot \phi \cdot \kappa \cdot \Gamma \cdot I$

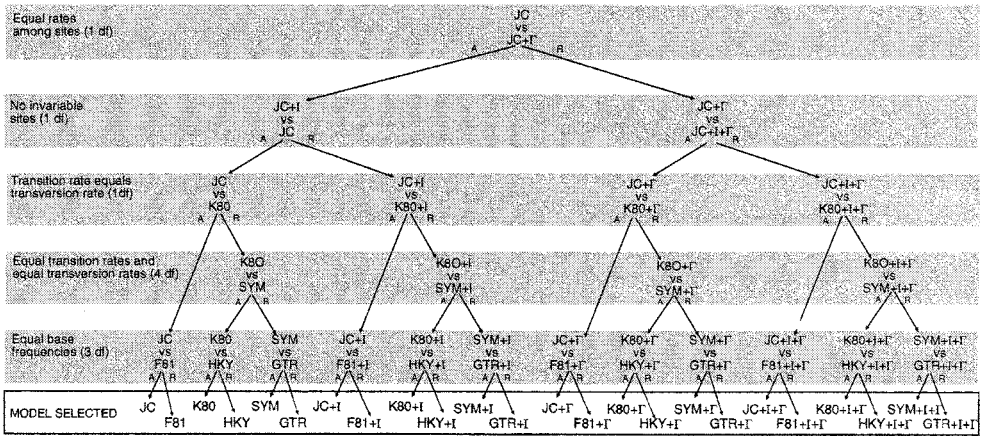
FIGURE 4. Hierarchical LRTs. LRTs are used to compare two different models at a time. The simpler model represents the null hypothesis. A model is accepted (A) or rejected (R) and the next LRT in the corresponding path, A or R, is performed until a final model is selected. Several starting models (in parentheses) and several orders of parameter additions were tried (panels a–d). (Continued)

explained below. To adjust for the inflation of type I error (rejection of the null model when it is true) when performing multiple LRTs, we applied a standard Bonferroni correction. Because four or five LTRs were carried out in each case, the individual alpha level was set to 0.01 to preserve on average a family alpha level of 0.05. The inflation of type II error (failure to reject the null model when it is false) was not corrected, because there is no obvious procedure to adjust for this kind of error when performing multiple LRTs. Although this is not the most satisfactory solution, it

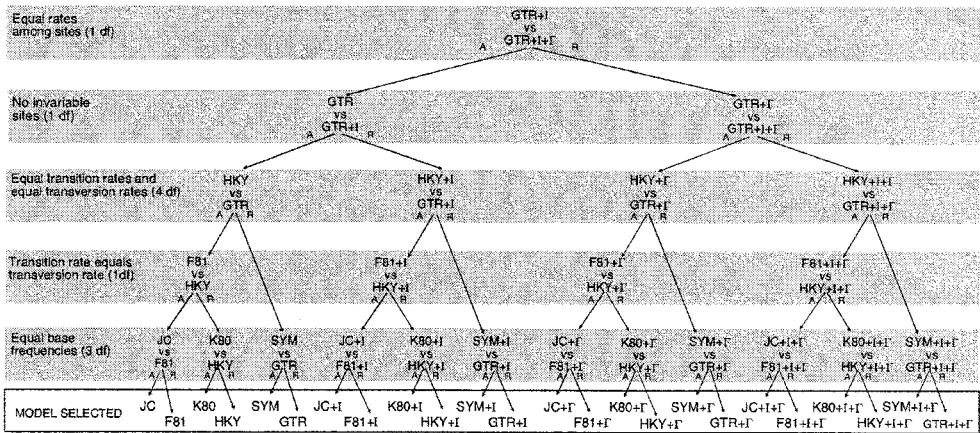
should not influence the model selection procedure, *P*-values for the LRTs being typically very small.

Dynamic Likelihood Ratio Tests.—An alternative to the use of a predefined ηLRT is to let the data set itself determine the order in which the hypotheses are tested; that is, the hierarchy used does not have to be the same for different data sets. The algorithms we suggest (δLRT_1 and δLRT_2) are as follows:

1. Start with a simple JC (δLRT_1) or a complex GTR + I + Γ (δLRT_2) model and



(c) ηLRT_3 (JC) $\Gamma \cdot I \cdot \kappa \cdot \phi \cdot \pi$



(d) ηLRT_4 (GTR+I+Γ) $\Gamma \cdot I \cdot \phi \cdot \kappa \cdot \pi$

FIGURE 4. (Continued).

calculate its likelihood. This is the current model.

2. Calculate the likelihood of the alternative (δLRT_1) or null (δLRT_2) models that differ by one assumption, and perform the corresponding nested LRTs.
3. δLRT_1 : If any hypothesis or hypotheses are rejected, the alternative model corresponding to the LRT with smallest associated P -value becomes the current model.

In the case of several equally smallest P -values, select the alternative model with the best likelihood.

δLRT_2 : If any hypothesis or hypotheses are not rejected, the null model corresponding to the LRT with greatest associated P -value becomes the current model. In the case of several equally greatest P -values, select the null model with the best likelihood.

4. Repeat steps 2 and 3 until the algorithm converges.

The alternative paths the algorithm can generate can be represented graphically (Fig. 5). Regarding multiple significance, it

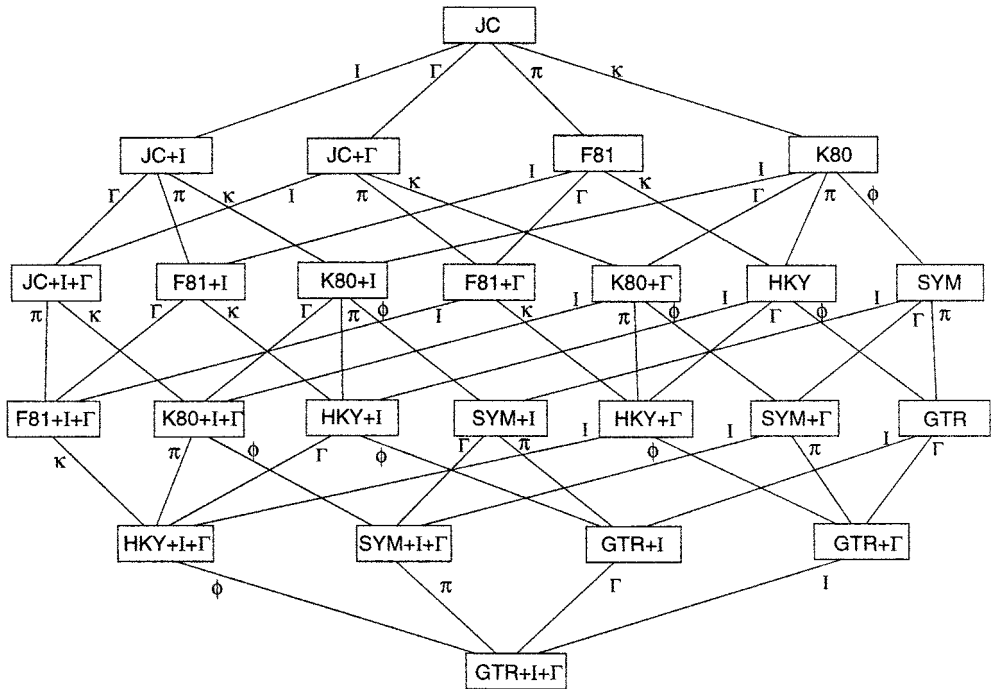


FIGURE 5. Dynamic LRTs. Starting with the simplest (JC) or the most complex model (GTR + I + Γ), LRTs are performed among the current model and the alternative model that maximizes the difference in likelihood. π : base frequencies. κ : transition/transversion bias. ϕ : substitution rates among nucleotides. Γ : rate heterogeneity among sites. I: proportion of invariable sites.

is not clear how to apply a correction consistently in this case. The number of tests performed may vary in each case. In addition, several tests are performed, but only some of them are actually considered. We decided to use an individual alpha value of 0.01 in all tests. In any case, the P -values obtained are generally so small that the different possible corrections for the type I error inflation should not change the final outcome.

Akaike Information Criterion.—The AIC (Akaike, 1974) is an asymptotically unbiased estimator of the Kullback–Leibler information quantity (Kullback and Leibler, 1951). The smaller the AIC, the better the fit of the model to the data (approximately equivalent to minimizing the expected Kullback–Leibler distance between the true model and the estimated model). An advantage of the AIC is that it can also be used to compare both nested and nonnested models. Because the AIC penalizes for the increasing number of parameters in the model, it is taking into account not only the goodness of fit but also the variance of the parameter estimates. It is

computed as

$$AIC_i = -2 \ln L_i + 2N_i,$$

where N_i is the number of free parameters in the i th model and L_i is the maximum-likelihood value of the data under the i th model. We also tried to empirically “tune” the penalty of the AIC by running several simulations and finding which penalty would increase model selection accuracy in those simulations. We use the name AIC_1 for the standard definition with a penalty of 2; while AIC_2 is the empirically tuned AIC ($AIC_{2i} = -2 \ln L_i + 5N_i$).

Bayesian Information Criterion.—The BIC (Schwarz, 1974) provides an approximate solution to the natural log of the Bayes factor, especially when sample sizes are large and competing hypotheses are nested (Kass and Wasserman, 1994). The Bayes factor measures the relative support the data set gives to different models, but its computation often involves difficult integrals and

an approximation becomes convenient. As with the AIC, the BIC can also be used to compare nested and nonnested models. Its definition is

$$\text{BIC}_i = -2 \ln L_i + N_i \ln n,$$

where n is the sample size (sequence length). The smaller the BIC, the better the fit of the model to the data. Because in real data analysis, the natural log of n is usually >2 , the BIC should tend to choose simpler models than does the AIC₁ but more complex models than the AIC₂.

RESULTS

We recorded the number of times a model selection method chose any model as the best-fit model out of 100 replicates. The accuracy of the different model selection strategies was defined as the number of times a method recovered the correct model of evolution out of the 100 replicates, that is, the probability of recovering the true model under which the data were simulated. The absolute and relative accuracy of the different methods varied across the different simulated conditions: base tree and number of characters (Simulation I), number of taxa and tree height (Simulation II), molecular clock (Simulation III), and rate heterogeneity among sites (Simulations I and IV).

Base Tree

The first step in the model selection procedure is the estimation of a base tree (including branch lengths). The use of different models of evolution for estimating the NJ, base trees did not affect model selection accuracy (Fig. 6). The use of the true tree as the base tree only increased model selection performance, compared with the use of NJ base trees, when the number of characters was small (100). When the base model was a random tree, then relative to the use of a NJ as the base tree, model selection accuracy slightly increased for 100 characters for some methods, but decreased vastly for 500 and 1,000 characters. However, this decrease in accuracy was mainly due to the overestimation of rate variation. The substitutional pattern and the base frequencies were correctly identified only 10–20%

less frequently than when a NJ base tree was used as the base tree. When the true model did not include rate variation, the model selected by using a random tree as the base tree was identical to the true model in the substitutional pattern and in the base frequencies assumption, but included rate heterogeneity among sites as modeled by the gamma distribution (i.e., the selected model was model + Γ instead of just model; data not shown). When the true model included rate variation (+ Γ), the model selected was identical to the true model except that a significant proportion of invariable sites was also included (i.e., the selected model was model + I + Γ instead of just model + Γ).

Number of Characters

Increasing the number of characters rapidly improved the performance of most model selection methods. When the true model was JC (Fig. 7), accuracy values were ~95% for all methods, except for the AIC₁, which selected the true model 70% of the time. When the true model was HKY (Fig. 8), accuracy was ~50–80% for 100 characters and increased to 75–100% with 500 and 1,000 characters. However, the ηLRT_3 selected the true model only 10% of the time, being extremely biased towards models that are more complex (selecting GTR when the true model was HKY). In addition, the AIC₁, with 500 and 1,000 characters, recovered the true model only 75% of the time. When the true model was GTR (Fig. 9), accuracy increased from 10% to 80% and 95% with 100, 500, and 1,000 characters, respectively. The AIC₁ performed better than the rest with 100 characters, but the opposite was true with 500 or 1,000 characters. With rate heterogeneity (Figs. 7–9), the patterns were similar but with a decrease in accuracy. This decrease was particularly true for ηLRT_2 and ηLRT_4 when the number of characters simulated were low (100), and for the AICs and BIC when the true model was GTR + Γ .

Number of Taxa

Adding taxa also increased, in general, the accuracy of the different methods (Table 3). In the absence of rate variation, most methods performed quite well (>90%) with 10 taxa when the true model was JC or HKY, or with

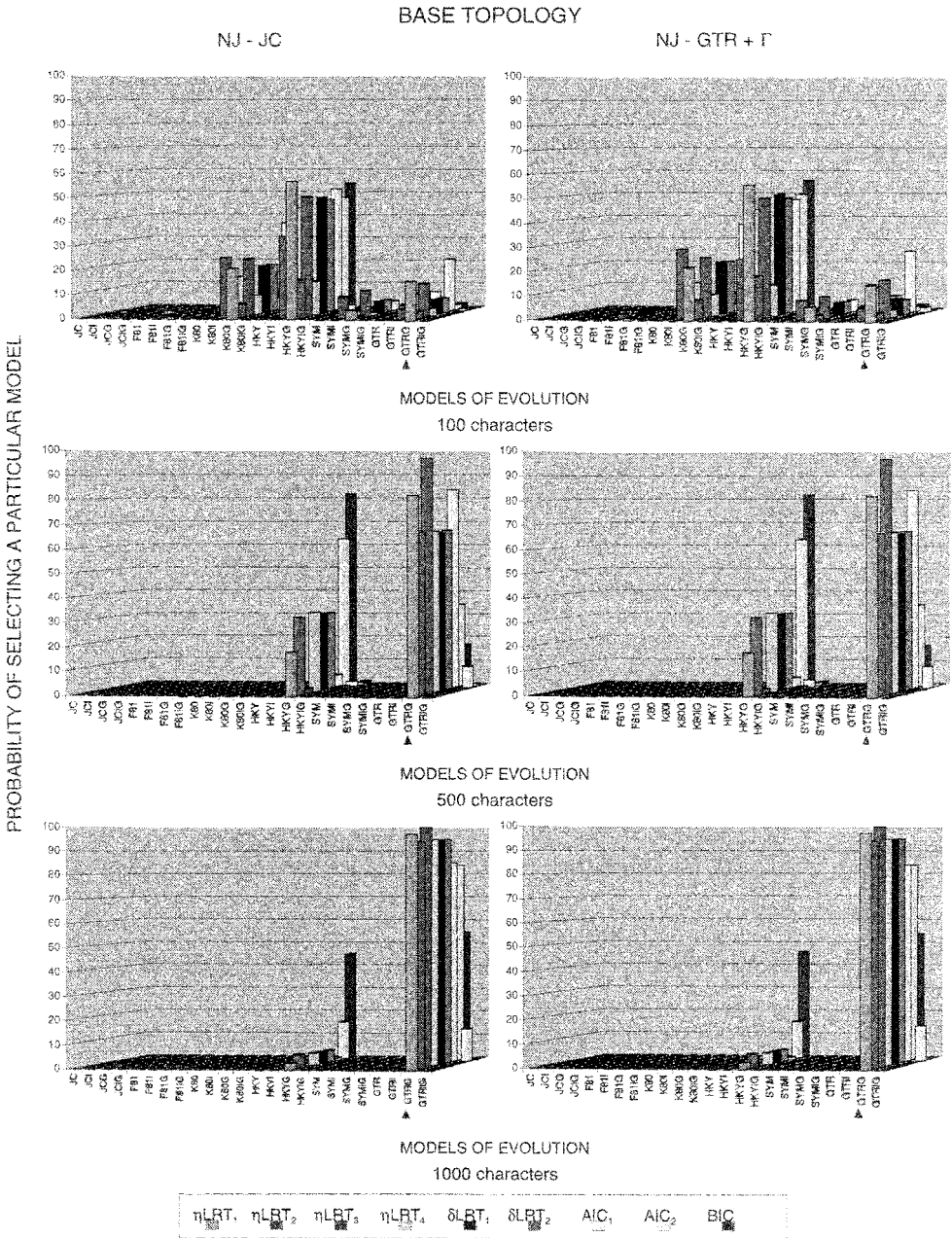


FIGURE 6. Effect of the base tree in model selection. The y-axis represents the number of times a model was selected as the best-fit model (out of 100 replicates). The x-axis represents the model of nucleotide substitution selected (GTRIG corresponds to GTR + I + Γ model, and so on). The true model is identified by the black triangle below the x-axis. Different methods for model selection are represented on the z-axis (left to right on the legend is front to back on the z-axis). Data were simulated for 20 taxa and a tree height of 0.10 according to the GTR + Γ model with parameter values as in Table 2. (Continued)

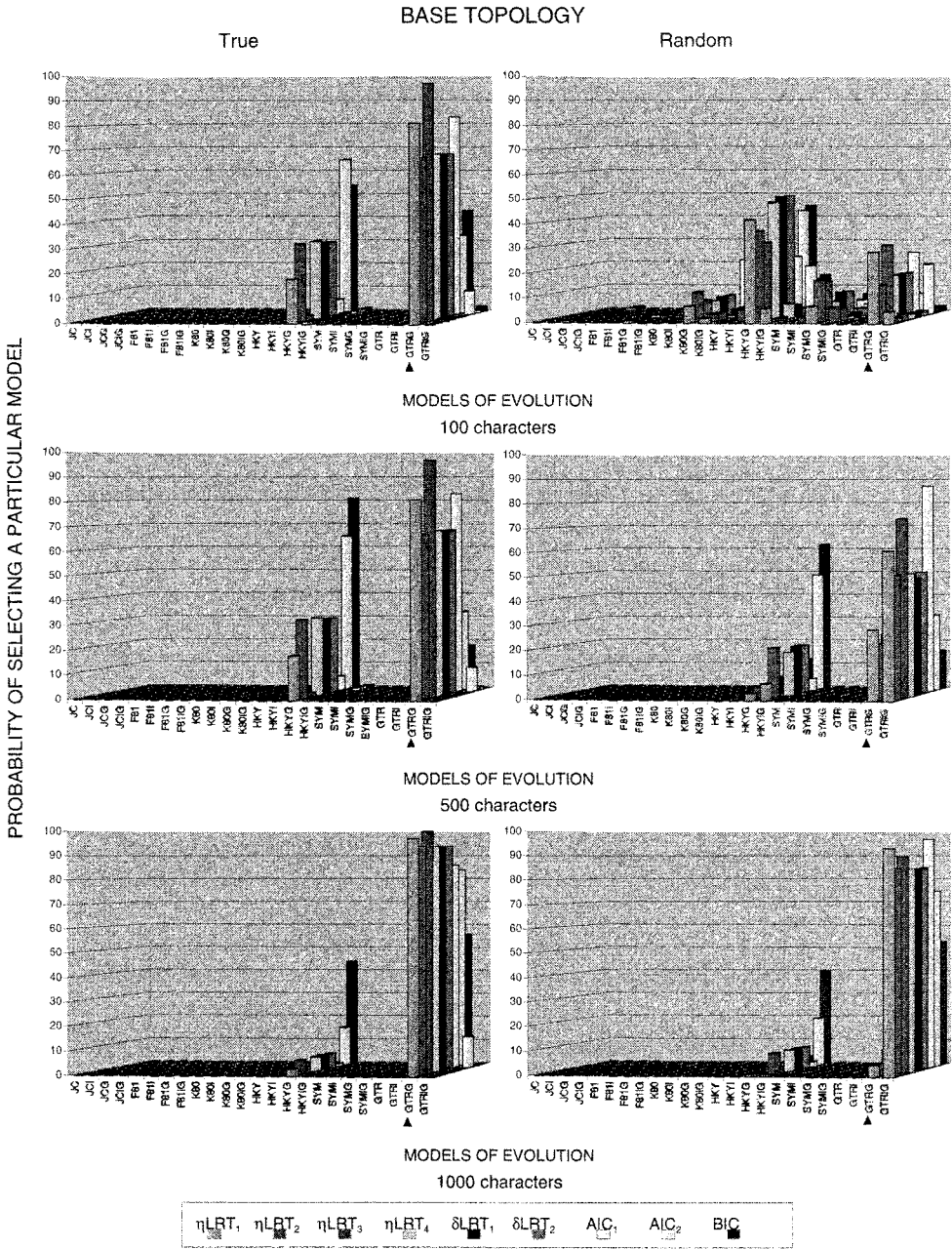


FIGURE 6. (Continued).

20 taxa when the true model was GTR. In the presence of rate variation, the performance patterns were more complex, and the η LRT₁ and η LRT₃ accuracy decreased with more taxa when the true model was HKY + Γ . For the most complex model, simulated, GTR + Γ , and with 500 characters, 50 taxa were needed to obtain accuracies >90%.

Tree Height

The effect of tree height depended on the complexity of the true model (Table 4). When the true model was JC or HKY, accuracy was not very affected by the different tree heights explored. The η LRT₃ method, however, showed a dramatic decrease in accuracy with increasing tree height, because of its

TABLE 3. Effect of the number of taxa on model selection. Alignments with 500 characters were simulated under different topologies with a tree height of 0.10, generated by a birth–death process. Sequences were simulated with parameter values as in Table 2. The base tree was a NJ-HKY tree. Ntaxa is the number of taxa.

True model	Ntaxa	ηLRT_1	ηLRT_2	ηLRT_3	ηLRT_4	δLRT_1	δLRT_2	AIC ₁	AIC ₂	BIC
JC	10	98	98	98	97	98	98	64	96	98
	20	97	97	97	96	97	96	70	93	96
	50	97	95	97	95	97	95	64	95	98
	100	94	96	94	94	94	93	67	95	98
HKY	10	100	100	37	100	100	100	87	100	100
	20	95	95	12	95	95	95	78	99	99
	50	99	99	0	100	99	99	86	98	99
GTR	100	98	97	0	98	98	98	80	100	100
	10	59	60	60	60	59	59	71	38	16
	20	91	89	97	91	91	91	84	63	50
JC + Γ	50	99	100	97	100	99	99	88	98	97
	100	100	100	90	100	100	100	81	100	100
	10	98	98	98	97	98	98	64	96	98
HKY + Γ	20	93	91	93	88	93	91	52	91	94
	50	92	93	94	93	94	92	61	90	95
	100	94	97	97	97	97	96	72	95	99
GTR + Γ	10	97	0	54	0	22	32	27	31	31
	20	95	98	4	99	99	99	80	100	100
	50	83	96	0	95	95	95	78	96	97
GTR + Γ	100	70	99	0	99	99	99	89	99	99
	10	52	11	67	11	30	29	52	8	1
	20	82	67	97	67	66	67	84	35	17
	50	93	94	95	95	95	95	86	74	54
	100	99	99	98	99	99	99	93	98	98

the presence or absence of a molecular clock, did not affect model selection accuracy for the 1,000-character data sets simulated (Table 5). The extreme bias of the ηLRT_3 for more complex models (selecting GTR when the true model was HKY, and selecting GTR + Γ when the true model was HKY + Γ) was constant in the presence or absence of a molecular clock.

Rate Variation Among Sites

Almost invariably, throughout all the simulations, the presence of rate variation among sites reduced accuracy, especially for low numbers of characters or taxa or for small tree heights. When the true model was JC and 1,000 characters were simulated, the amount of rate variation did not affect model selection (Table 6). When the true model was HKY, the ηLRT_3 performed poorly in the presence of any rate variation (again, bias for more complex models), whereas the ηLRT_1 and ηLRT_2 methods showed a decrease in accuracy with extreme values of rate variation ($\alpha = 0.05$). In this case, the decrease in accuracy due to rate variation resulted from a trend to infer more complex substitutional

patterns (data not shown). When the true model was GTR + Γ , most methods showed lower accuracies with extreme rate variation. In that case, the decrease in accuracy was due to the inference of less complex substitutional patterns (data not shown).

LRT Distribution and Mixed χ^2

Results from the use of a mixed χ^2 distribution instead of a standard χ^2 distribution to approximate the LRT *P*-values were undistinguishable (data not shown).

DISCUSSION

Accuracy of Model Selection

The results of this simulation study suggest that model selection procedures perform quite well under different conditions. Given a modest number of characters (500) and taxa (20), most model selection procedures are highly accurate in most conditions. The LRT methods performed better than the AIC or BIC methods, although ηLRT_3 performed very badly when the models were of medium complexity (HKY). The differences among the LRTs methods were small, and the hierarchical and dynamic approaches seemed

TABLE 4. Effect of tree height on model selection. Alignments with 20 taxa and 500 characters were simulated under different topologies generated by a birth–death process. Sequences were simulated with parameter values as in Table 2. The base tree was a NJ-HKY tree. Tree height is the probability of substitution per site from the root to the tip.

True model	Tree height	η LRT ₁	η LRT ₂	η LRT ₃	η LRT ₄	δ LRT ₁	δ LRT ₂	AIC ₁	AIC ₂	BIC
JC	0.01	99	98	99	97	99	98	66	98	99
	0.10	97	97	97	96	97	96	70	93	96
	0.20	97	95	97	97	97	96	64	97	98
	0.50	97	96	97	96	97	96	62	96	96
	0.75	98	99	98	98	98	98	60	97	99
HKY	0.01	99	99	93	99	99	99	86	99	99
	0.10	95	95	12	95	95	95	78	99	99
	0.20	99	99	0	99	99	99	86	99	100
	0.50	96	98	0	98	96	96	83	96	96
	0.75	95	95	0	96	95	95	74	95	96
GTR	0.01	12	13	11	12	12	12	34	1	0
	0.10	91	89	97	91	91	91	84	63	50
	0.20	99	99	98	99	99	99	83	98	92
	0.50	99	100	99	100	99	99	81	97	99
	0.75	98	98	99	98	98	98	81	97	98
JC + Γ	0.01	27	0	27	0	13	11	16	16	13
	0.10	93	91	93	88	93	91	52	91	94
	0.20	96	96	96	96	96	96	66	94	96
	0.50	92	92	92	92	92	92	60	93	94
	0.75	98	96	98	97	98	97	71	97	99
HKY + Γ	0.01	74	0	67	0	16	17	24	27	20
	0.10	95	98	4	99	99	99	80	100	100
	0.20	81	97	0	97	97	97	70	96	99
	0.50	65	100	0	100	100	100	81	100	100
	0.75	48	98	0	98	98	98	87	98	99
GTR + Γ	0.01	6	0	0	0	1	2	9	0	0
	0.10	82	67	97	67	66	67	84	35	17
	0.20	91	77	97	78	78	78	84	47	30
	0.50	90	86	95	86	86	86	91	60	40
	0.75	90	89	82	90	90	90	88	56	45

to perform equally well, although the dynamic approaches were more stable in their accuracy patterns. The “empirically tuned” AIC₂ performed better than the original AIC₁.

Influence of the Base Tree

The initial tree topology used to estimate the likelihood scores for the different models (the base tree) did not affect model selection as long it was not a random tree. For as few

TABLE 5. Effect of the presence or absence of a molecular clock on model selection. Alignments with 20 taxa and 1,000 characters were simulated under the conditions for the trees shown in Figure 3. Sequences were simulated with parameter values as in Table 2. The base tree was a NJ-JC tree.

True model	Molecular clock	η LRT ₁	η LRT ₂	η LRT ₃	η LRT ₄	δ LRT ₁	δ LRT ₂	AIC ₁	AIC ₂	BIC
JC	Clock	95	95	95	95	95	94	63	97	98
	Nonclock	96	92	96	93	96	92	55	95	98
HKY	Clock	98	98	1	98	98	98	81	97	100
	Nonclock	95	96	0	96	95	95	83	98	99
GTR	Clock	98	98	98	98	98	98	87	97	94
	Nonclock	100	100	100	100	100	100	91	98	93
JC + Γ	Clock	98	97	97	96	97	96	65	94	98
	Nonclock	97	96	97	97	97	96	61	97	98
HKY + Γ	Clock	98	99	0	98	99	99	82	98	100
	Nonclock	92	96	0	96	96	96	82	96	97
GTR + Γ	Clock	97	94	100	95	95	95	85	84	55
	Nonclock	98	92	99	92	92	92	88	81	50

TABLE 6. Effect of rate variation among sites on model selection. Alignments with 20 taxa and 1,000 characters were simulated under the conditions for the tree shown in Figure 3a. The tree height was 0.10. Sequences were simulated with base frequencies and substitution parameters as in Table 2. The base tree was a NJ-JC tree. Alpha (α) is the shape of the Γ distribution Ncat is the number of discrete categories for the Γ distribution. $\alpha = \infty$ effectively means no rate variation has been observed among sites.

True model	α	Ncat	η LRT ₁	η LRT ₂	η LRT ₃	η LRT ₄	δ LRT ₁	δ LRT ₂	AIC ₁	AIC ₂	BC
JC + Γ	∞	—	95	95	95	95	95	94	63	97	98
	0.5	4	98	97	97	96	97	96	65	94	98
	0.2	4	96	94	96	96	96	95	64	95	96
	0.2	8	95	94	96	93	96	95	63	97	98
	0.05	8	97	70	98	95	98	96	70	97	99
HKY + Γ	∞	—	98	98	1	98	98	98	81	97	100
	0.5	4	98	99	0	98	99	99	82	98	100
	0.2	4	81	98	0	99	99	99	74	98	99
	0.2	8	77	98	0	98	98	98	73	97	98
	0.05	8	67	54	7	96	97	96	88	100	100
GTR + Γ	∞	—	98	98	98	98	98	98	87	97	94
	0.5	4	97	94	100	95	95	95	85	84	55
	0.2	4	92	90	100	94	94	94	91	73	45
	0.2	8	87	79	100	79	79	79	92	50	17
	0.05	8	70	64	93	70	70	70	92	39	13

as 100 characters, the use of the true tree as the base tree increased accuracy. This is explained by the fact that NJ trees are worse estimates of the true tree with 100 characters than with 500 or 1,000 characters. At the same time, the amount of information is less in the 100 character data sets, and the estimation of the different parameters in a reliable topology becomes more relevant. With increased amounts of data, the use of NJ trees as base trees resulted in the same accuracy as the use of the true tree. A simple NJ-JC tree worked as well as a NJ-GTR tree in all cases. It has been found previously that accurate estimates of substitution parameters can be obtained even with an incorrect phylogeny (Yang, 1994b), although the tree used in these simulations had very short internal branches (Sullivan et al., 1996). As shown previously by Sullivan et al. (1996), the use of random trees can lead to an overestimation of rate heterogeneity.

Adding or Removing Parameters?

An open debate in statistics is whether model selection procedures should start with a simple model to which parameters might be added (bottom-up) or with a complex model from which parameters might be removed (top-down). For example, to select the best-fit model for the data at hand, we could start with the simple JC model, and test whether the addition of one parameter

improves significantly the likelihood of the model (e.g., JC vs. JC + Γ). If this is the case, the parameter is included in the model, whereas if the likelihood does not improve significantly, the parameter is not added. On the other hand, we could start with the complex GTR + I + Γ and test whether the removal of one parameter (e.g., GTR + I + Γ vs. GTR + I) decreases the likelihood significantly. If the likelihood does not decrease significantly, the parameter is removed from the model. In the context of selection of models of nucleotide substitution, both approaches have been used (e.g., Sullivan and Swofford, 1997; Keylsey et al., 1999). In this simulation, the bottom-up approaches (η LRT₁ and η LRT₃) performed better than the top-down ones (η LRT₂ and η LRT₄) for small tree heights or number of taxa but showed some biases towards wrong models when the true model was HKY or HKY + Γ . This indicates that starting with the simplest or most complex model may influence model selection, although not in a consistent manner. With δ LRTs, however, the complexity of the starting model did not change the accuracy of model selection.

The Order of Tests of Hypotheses

For both the bottom-up or top-down approaches, different parameters of the model may be added or removed in different orders.

In other words, different hypotheses can be tested earlier or later in the hierarchy of LRTs. The order in which parameters are added or removed determines which hypotheses are tested in the presence of which parameters. For example, the κ hypothesis can be tested by comparing JC and K80, with no additional free parameters, or by comparing F81 versus HKY, both of which contain parameters π . If the presence of additional parameters does not affect the outcome of the LRTs, we expect model selection to be independent of the order in which the different hypotheses are tested. Whelan and Goldman (1999) and Goldman and Whelan (2000) suggested this to be the case, but their LRTs were performed by assuming the true model was the null hypothesis. This is not the situation here, where the null hypothesis will be the true model only in a few of the LRTs performed, nor is it the case with real data, because the true model is unknown. On the contrary, Zhang (1999) suggested that LRTs of the transition/transversion bias or rate variation are affected by the presence of other parameters. For example, in Zhang's simulations the failure to take into account unequal base frequencies led to the rejection of the null hypothesis of no transition bias much more often than expected. Cunningham et al. (1998), using an empirically generated phylogeny, observed that the choice of the best-fit models was affected by the order of addition of parameters. However, their conclusion would be the opposite if they had corrected for type I error in the tests presented in their Table 1. In these simulations, the patterns of accuracy of ηLRT_1 and ηLRT_2 , and of ηLRT_3 and ηLRT_4 , were almost identical most of the time, but the ηLRT_3 clearly showed a bias toward complex models not present in ηLRT_4 . This suggests that the order in which parameters are added or removed to or from a model may have some effect on model selection under some circumstances. In general, apparently testing first for the base frequencies and the substitution pattern before testing for rate heterogeneity was slightly more effective, unless rate heterogeneity was strong. Perhaps the inclusion of rate heterogeneity might also account for a large portion of the variation (information) in the data. If included at the beginning of the sequence of tests, especially when starting from simple models, this variation would make it more difficult to effectively test other hypotheses.

However, this effect was not significant, because the performance of the hierarchical and dynamic LRTs was more often similar than not.

The χ^2 Distribution

Although the standard χ^2 distribution may be significantly different from the true LRT distribution in the case of boundary LRTs, the P -values obtained in LRTs of evolutionary hypotheses are often so small that this bias does not affect the results. Indeed, the appropriate χ^2 distribution should be used in each case (Goldman and Whelan, 2000; Ota et al., 2000).

The Importance of Models

The relevance of models of evolution to phylogenetic estimation has been extensively discussed in the literature. In general, phylogenetic methods perform worse when the model of evolution assumed is incorrect (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995; Bruno and Halpern, 1999). When substitution rates vary among lineages, the use of an appropriate model is of utmost importance for obtaining a correct tree topology (Takezaki and Gojobori, 1999; Philippe and Germot, 2000). Cases where the use of wrong models increases phylogenetic performance (see Yang, 1997b; Xia, 2000; Posada and Crandall, 2001) are exceptional and rather represent a bias towards the true tree associated with violated assumptions (Bruno and Halpern, 1999). However, the relationship between the fit of the model to the data and the ability of the model to correctly predict topology is not straightforward. Topology estimation by methods such as maximum likelihood are relatively robust to the model used (Fukami-Kobayashi and Tateno, 1991; Gaut and Lewis, 1995; Yang et al., 1995). The evaluation of reliability of the estimated trees depends critically on the model; false or simple models tend to suggest that a tree is significantly supported when it cannot be (Yang et al., 1994; Buckley et al., 2000).

The use of appropriate models is especially critical for parameter estimation and, consequently, to understand the evolutionary process. When a relatively simple model of substitution is assumed, the transition/transversion ratio, branch lengths, and sequence divergence are underestimated,

whereas the shape parameter of the gamma distribution is overestimated (Tamura, 1992; Wakeley, 1994; Yang et al., 1994, 1995; Adachi and Hasegawa, 1995; Buckley et al., 2000). Moreover, the outcome of different tests of evolutionary hypotheses (e.g., molecular clock) may depend on the model of evolution assumed (Zhang, 1999).

A researcher should then adopt the statistical model-fitting approach (*sensu* Huelsenbeck, 1995) and select among different models the one that best fits its data. However, is this selection reliable?

Caveats and Conclusions

Different model selection methods work well with simulated data sets. What is the relevance of this result to real data sets? The result obtained here pertain to a perfect fit between model and data, and real data rarely fit models perfectly. A future avenue of research might explore more realistic models, as nonreversible models or codon models for coding sequences. However, we have chosen conservative and meaningful parameter values in an attempt to mimic, as much as possible, empirical data sets. We suggest that if model selection procedures are able to recognize some features of the process of evolution in simulated data sets, these same methods can be expected to recognize these same features in real data sets, selecting the more appropriate, although still imperfect, models of evolution. Moreover, the parameter values used here were conservative. For example, more biased base frequencies (e.g., 0.5:0.1:0.1:0.3, instead of the 0.35:0.15:0.25:0.25 used here) would be expected to increase accuracy values for all model selection methods, because the differences among the models would become more evident.

Indeed, any model of nucleotide substitution is necessarily a simplification of the actual evolutionary process. Even the best-fit model is far from the true model underlying the evolution of the sequences under study. However, the statistical selection of the model of evolution used in the analysis is, first, philosophically necessary (for justification of the use of a particular model), and second, should provide equal or better estimates. Model selection should be a standard procedure in phylogenetic studies. A program facilitating this task, Modeltest

(Posada and Crandall, 1998), can be downloaded at no charge from http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm.

ACKNOWLEDGMENTS

Many thanks to David Swofford for his discussions on model selection. This work was supported by a BYU Graduate Studies Award (D.P.), the Alfred P. Sloan Foundation, and grant NIH R01-HD 34350-01A1 (K.A.C.).

REFERENCES

- ADACHI, J., AND M. HASEGAWA. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: Heterogeneity among amino acid sites. *J. Mol. Evol.* 40:622-628.
- AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* 19:716-723.
- BRUNO, W. J., AND A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* 16:564-566.
- BUCKLEY, T. R., C. SIMON, AND G. K. CHAMBERS. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: The effects of model assumptions on estimates of topology, edge lengths, and bootstrap support. *Syst. Biol.* 50:67-86.
- 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.
- EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. Pages 67-76 in *Phenetic and phylogenetic classification* (J. McNeill, ed.). Systematics Association Publication, London.
- FARRIS, J. S. 1973. A probability model for inferring evolutionary trees. *Syst. Zool.* 22:250-256.
- FELSENSTEIN, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22:240-249.
- FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401-410.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376.
- FELSENSTEIN, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22:521-565.
- FRATI, F., C. SIMON, J. SULLIVAN, AND D. L. SWOFFORD. 1997. Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in Collembola. *J. Mol. Evol.* 44:145-158.
- FUKAMI-KOBAYASHI, K., AND Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J. Mol. Evol.* 32:79-91.
- GAUT, B. S., AND P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* 12:152-162.
- 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.
- HASEGAWA, M. 1990a. Mitochondrial DNA evolution in primates: Transition rate has been extremely low in the lemur. *J. Mol. Evol.* 31:113–121.
- HASEGAWA, M. 1990b. Phylogeny and molecular evolution in primates. *Jpn. J. Genet.* 65:243–266.
- HASEGAWA, M., K. KISHINO, AND T. YANO. 1985. Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160–174.
- HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17–48.
- 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 D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247–264.
- JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 in *Mammalian protein metabolism* (H. M. Munro, ed.). Academic Press, New York.
- KASS, R. E., AND L. WASSERMAN. 1994. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Department of Statistics, Carnegie Mellon University. Pittsburgh, Pennsylvania. 16.
- KELSEY, C. R., K. A. CRANDALL, AND A. F. VOEVODIN. 1999. Different models, different trees: The geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* 13:336–347.
- KENDALL, M., AND STUART. 1979. The advanced theory of statistics. Charles Griffin, London.
- KULLBACK, S., AND R. A. LEIBLER. 1951. On information and sufficiency. *An. Math. Stat.* 22:79–86.
- MOROZOV, P., T. SITNIKOVA, G. CHURCHILL, F. J. AYALA, AND A. RZHETSKY. 2000. A new method for characterizing replacement rate variation in molecular sequences: Application of the Fourier and Wavelet models to *Drosophila* and mammalian proteins. *Genetics* 154:381–395.
- MUSE, S. 1999. Modeling the molecular evolution of HIV sequences. Pages 122–152 in *The evolution of HIV* (K. A. Crandall, ed.). Johns Hopkins Univ. Press, Baltimore.
- OTA, R., P. J. WADDELL, M. HASEGAWA, H. SHIMODAIRA, AND H. KISHINO. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* 17:798–803.
- PHILIPPE, H., AND A. GERMOT. 2000. Phylogeny of Eukaryotes based in ribosomal RNA: Long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* 17:830–834.
- POSADA, D., AND K. A. CRANDALL. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- POSADA, D., AND K. A. CRANDALL. 2001. Simple (wrong) models for complex trees: Empirical bias. *Mol. Biol. Evol.* 18:271–275.
- RAMBAUT, A., AND N. C. GRASSLY. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13:235–238.
- RANNALA, B., AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43:304–311.
- RZHETSKY, A., AND M. NEI. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12:131–151.
- SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
- SCHWARZ, G. 1974. Estimating the dimension of a model. *Ann. Stat.* 6:461–464.
- STEEL, M., AND D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17:839–850.
- 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.
- SULLIVAN, J., AND D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J. Mammal. Evol.* 4:77–86.
- 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 Associates, Sunderland MA.
- SWOFFORD, D. L. 1998. PAUP* Phylogenetic analysis using parsimony and other methods. 4.0 beta. Sinauer Associates, Sunderland, Massachusetts.
- TAKEZAKI, N., AND T. GOJOBORI. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol. Biol. Evol.* 16:590–601.
- TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C content biases. *Mol. Biol. Evol.* 9:678–687.
- TAMURA, K. 1994. Model selection in the estimation of the number of nucleotide substitutions. *Mol. Biol. Evol.* 11:154–157.
- TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Pages 57–86 in *Some mathematical questions in biology—DNA sequence analysis* (R. M. Míura, ed.). Am. Math. Soc., Providence, RI.
- WAKELEY, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11:436–442.
- WHELAN, S., AND N. GOLDMAN, 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 16:1292–1299.
- XIA, X. 2000. Phylogenetic relationships among horse-shoe crab species: Effect of substitution models in phylogenetic analysis. *Syst. Biol.* 49:87–100.
- YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401.
- YANG, Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39:306–314.
- YANG, Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* 43:329–342.
- YANG, Z. 1996a. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42:294–307.
- YANG, Z. 1996b. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11:367–372.

- YANG, Z. 1996c. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42:587–596.
- YANG, Z. 1997a. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13:555–556.
- YANG, Z. 1997b. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* 14:105–108.
- 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., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14:717–724.
- ZHANG, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* 16:868–875.

Received 7 April 2000; accepted 13 June 2000
Associate Editor: C. Simon