

The Effect of Branch Length Variation on the Selection of Models of Molecular Evolution

David Posada

Department of Zoology, Brigham Young University, Provo, UT 84602-5255, USA

Received: 2 October 2000 / Accepted: 4 January 2001

Abstract. Models of sequence evolution play an important role in molecular evolutionary studies. The use of inappropriate models of evolution may bias the results of the analysis and lead to erroneous conclusions. Several procedures for selecting the best-fit model of evolution for the data at hand have been proposed, like the likelihood ratio test (LRT) and the Akaike (AIC) and Bayesian (BIC) information criteria. The relative performance of these model-selecting algorithms has not yet been studied under a range of different model trees. In this study, the influence of branch length variation upon model selection is characterized. This is done by simulating sequence alignments under a known model of nucleotide substitution, and recording how often this true model is recovered by different model-fitting strategies. Results of this study agree with previous simulations and suggest that model selection is reasonably accurate. However, different model selection methods showed distinct levels of accuracy. Some LRT approaches showed better performance than the AIC or BIC information criteria. Within the LRTs, model selection is affected by the complexity of the initial model selected for the comparisons, and only slightly by the order in which different parameters are added to the model. A specific hierarchy of LRTs, which starts from a simple model of evolution, performed overall better than other possible LRT hierarchies, or than the AIC or BIC.

Key words: Nucleotide substitution models — Model selection — Likelihood ratio test — Hierarchical likeli-

hood ratio tests — Akaike information criterion — Bayesian information criterion — Mixed χ^2 — Branch length variation — Phylogenetics

Introduction

Models of nucleotide substitution—hereafter models of evolution—play a relevant role in molecular evolutionary studies (Liò and Goldman 1998; Steel and Penny 2000). Models of evolution are commonly used to describe sequence evolution through the estimation of parameters such as sequence divergence, base frequencies, transition/transversion ratios, synonymous/nonsynonymous substitutions, divergence times, etc., and for the estimation of phylogenetic trees. The accurate estimation of these parameters may depend on the model of evolution assumed. For example, when a simple model of evolution is used, transition/transversion ratios and branch lengths may be underestimated (Adachi and Hasegawa 1995; Tamura 1992; Wakeley 1994; Yang 1994a; Yang et al. 1994; Yang et al. 1995). Indeed, the use of correct models may be crucial to statistical tests of evolutionary hypotheses (Zhang 1999). Moreover, the use of a particular model of evolution may change the results of a phylogenetic analysis (Cunningham et al. 1998; Kelsey et al. 1999; Leitner et al. 1997; Sullivan and Swofford 1997)—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 model of evolution assumed is incorrect (Bruno and Halpern 1999; Felsenstein 1978; Huelsenbeck and Hillis 1993; Penny et

al. 1994). Indeed, model selection is not important just because of its consequences in sequence analysis, but because the characterization of the evolutionary process is itself a legitimate pursuit.

To study the evolutionary process acting at the molecular level, and because models of evolution may influence the results of the analysis, the selection of a particular model of evolution for the analysis of a particular data set should be justified. A common strategy for model selection is the arbitrary use of complex, parameter-rich, models. However, this approach has several disadvantages. As a large number of parameters need to be estimated, the analysis becomes computationally difficult and requires a large amount of time, and more error is included in each estimate. Ideally, models should incorporate as much complexity (parameters) as needed. Several statistical procedures have been adapted to sequence data for choosing among alternative models of evolution. Among these, the likelihood ratio tests (LRTs) (Felsenstein 1981; Felsenstein 1988; Goldman 1993a; Goldman 1993b), and the Akaike information criterion (AIC) (Akaike 1974), are the most commonly used in sequence analysis (Corneli and Ward 2000; Hasegawa 1990a; Hasegawa 1990b; Muse 1999; Tamura 1994; Wang et al. 2000). Other model selection strategies, like the Bayesian information criterion (BIC) (Schwarz 1974), might also be suitable for the selection of a model of evolution (Morozov et al. 2000).

The absolute accuracy and relative performance of several statistical procedures for model selection has recently been evaluated under different conditions (Posada and Crandall 2001a). The study showed that the tree used for the estimation of the different parameters and likelihood of a particular model of evolution does not affect model selection as far as this tree is a reasonable estimate of the phylogeny (i.e., not a random tree). The study also showed how the parameter addition sequence and the starting model in the comparisons might have an effect on the selection of models of evolution. However, only a few topologies were used in the study to simulate the data, and the effect of branch length variation in model selection was not taken into account.

In this study, the performance of different hierarchies of LRTs, and the AIC and BIC model selection procedures is compared under different branch lengths in four-taxon trees. This is accomplished 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.

Methods

Data Simulation and Models

Nucleotide sequences were simulated under four-taxon trees for which the lengths for two sets of branches were varied independently (Fig. 1).

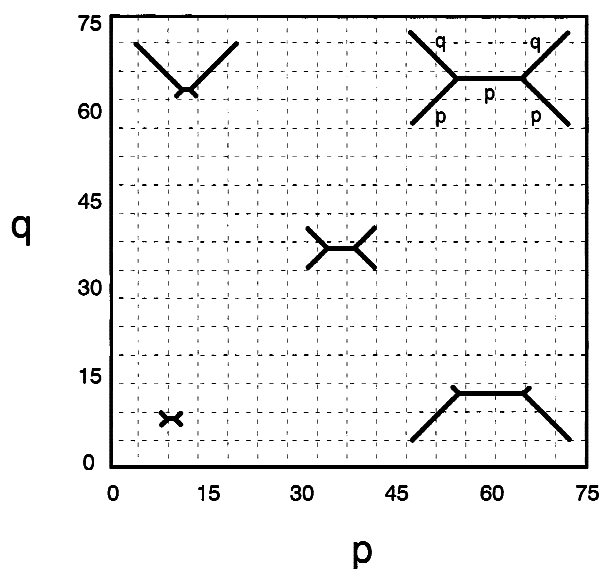


Fig. 1. Parameter space and trees used in the simulations. The three-branch parameter (p) and the two-branch parameter (q) varied from 0.01 to 0.75 in increments of 0.05. The grid is a 16×6 array in which each square represents 100 simulations.

The models of nucleotide substitution used in the simulations (the true models) were the Jukes–Cantor model (Jukes and Cantor 1969) (JC), the Hasegawa–Kishino–Yano model (Hasegawa et al. 1985) with rate variation among sites (HKY + Γ) and the general time-reversible model (Tavarè 1986) with rate variation among sites (GTR + Γ) (Table 1). The values of the parameters were arbitrarily chosen to fit into a range of values commonly observed in real data sets. The rate variation among sites was incorporated using the discrete gamma distribution with four rate categories (Γ) (Yang 1993; Yang 1994b; Yang 1996a). Four different sequence lengths were simulated in the case of the GTR + Γ model (100, 500, 1000 and 3000 characters), while for the JC and HKY + Γ models, only 1000 characters were simulated. For each set of conditions, 100 replicate data sets were simulated using the program Seq-Gen 1.1 (Rambaut and Grassly 1997).

Likelihood Estimation

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, the likelihood obtained for a fixed tree may be contrasted under the alternative models. The tree estimated from the data, and used to estimate the parameters and likelihood of the models compared will be referred hereafter as the *base tree*. Because the base tree does not affect the model-selection procedure as long as it is an estimate of the phylogeny and not a random tree (Posada and Crandall 2001a), a neighbor-joining tree (Saitou and Nei 1987) estimated under the JC model of evolution was used to estimate the model parameters and the likelihood of the models. For each simulated data and base tree, twenty-four likelihood scores, corresponding to twenty-four different models of evolution (Figure 1), were calculated in PAUP* (Swofford 1998).

Model Selection Strategies

The estimated likelihood scores were used to select the best-fit model of evolution for each data set using three different strategies and nine variations of those.

Table 1. Simulation parameter values

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

^a $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ describes equilibrium base frequencies, and has three free parameters, because of the constraint that the sum is one. $\varphi = (\varphi_{A-C}, \varphi_{A-G}, \varphi_{A-T}, \varphi_{C-G}, \varphi_{C-T}, \varphi_{G-T})$ describes the substitution rates among bases, and has five free parameters, because of the constraint $\varphi_{X-Y} = \varphi_{X-Y}/\varphi_{G-T}$. 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 4 discrete categories.

Hierarchical likelihood ratio tests (η LRTs). The LRT statistic is used extensively to compare the quality of fit of two different models:

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

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).

When the compared models are nested, that is, 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 (Kendall and Stuart 1979). The appropriateness of the χ^2 approximation of the LRT statistic when comparing models of evolution has often been debated (Goldman 1993a; Goldman 1993b; Whelan and Goldman 1999; Yang 1996b; Yang et al. 1995). Moreover, the simple χ^2 is expected only if the null model corresponds to fixing some parameters in the alternative model to values inside the parameter space. When the null model corresponds to fixing one parameter at the boundary of its range in the alternative model, a mixed χ^2 (or $\bar{\chi}^2$) distribution, consisting of 50% χ_0^2 and 50% χ_1^2 , should be used (Self and Liang 1987). Recent simulation studies using models of evolution reinforce this result (Goldman and Whelan 2000; Ota et al. 2000). In more complicated cases the null distribution of the LRT statistic might not be known. The effect of using a simple χ^2 for all LRTs, the more common way LRTs have been used in the past, is also evaluated here.

Likelihood ratio tests can be performed in a hierarchical manner to estimate the best-fit model for a particular data set (Fрати et al. 1997; Huelsenbeck and Crandall 1997; Posada and Crandall 1998; Posada and Crandall 2001a; Sullivan et al. 1997; Yang et al. 1994). It has been suggested that the choice of the best-fit model is affected by the order of parameter addition (Cunningham et al. 1998; Posada and Crandall 2001a). To explore this question, four different hierarchies of LRTs with different sequence of parameter addition/removal have been used (η LRT₁- η LRT₄) (Table 2 and Fig. 2), where η LRT₁ and η LRT₃ start in a simple model (JC) and η LRT₂ and η LRT₄ start in a complex model (GTR + I + Γ).

Traditional statistical tests such as LRTs are designed to reject the

Table 2. Model selection strategies

| Method ^a | Parameter ^b addition | Starting model | ρ_1 ^c | ρ_2 |
|---------------------------|---|-------------------------|-----------------------|----------|
| LRT ₁ | $\pi \cdot \kappa \cdot \varphi \cdot I \cdot \Gamma$ | JC | — | — |
| LRT ₂ | $\pi \cdot \varphi \cdot \kappa \cdot I \cdot \Gamma$ | GTR + I + Γ | — | — |
| LRT ₃ | $\Gamma \cdot I \cdot \kappa \cdot \varphi \cdot \pi$ | JC | — | — |
| LRT ₄ | $\Gamma \cdot I \cdot \varphi \cdot \kappa \cdot \pi$ | GTR + I + Γ | — | — |
| δ LRT ₁ | dynamic | JC | — | — |
| δ LRT ₂ | dynamic | GTR + I + Γ | — | — |
| AIC ₁ | — | simultaneous comparison | 2 | 2 |
| AIC ₂ | — | simultaneous comparison | 2 | 5 |
| BIC | — | simultaneous comparison | — | — |

^a η LRT: hierarchical likelihood ratio test. δ LRT: dynamical likelihood ratio test. AIC: Akaike information criterion. BIC: Bayesian information criterion. $AIC = \rho_1 \times \ln \text{likelihood} + \rho_2 \times \text{number of free parameters}$. $BIC = 2 \times \ln \text{likelihood} + \ln \text{sample size (number of characters)} \times \text{number of free parameters}$.

^b π : base frequencies. κ : transition/transversion bias. φ : substitution rates among nucleotides. Γ : rate heterogeneity among sites. I : proportion of invariable sites.

^c ρ_n represents the penalty value.

null hypothesis, not to prove it. Because of this, type I error (reject the null hypothesis when it is true) is perceived to be much more serious than type II error (fail to reject the null hypothesis when it is wrong). To adjust for the inflation of type I error when performing multiple LRTs, a standard Bonferroni correction was applied—because 4 or 5 LRTs were carried out in each case, the individual alpha level was set to 0.01 in order to preserve on average a family alpha level of 0.05. Because the null hypothesis can be wrong in many ways, the type II error is in general unknown. No attempt to correct for type II error was made, as there is not an obvious procedure to correct for this kind of error when performing multiple LRTs. Type I and II errors were also estimated from the performance results for the η LRT strategies.

Dynamical Likelihood Ratio Tests. An alternative to the use of a predefined hierarchy of LRTs is to let the data itself determine the order in which the hypotheses are tested. In this way, the hierarchy used does not have to be the same for different data sets. The algorithms suggested (δ LRT₁ and δ LRT₂) are as follows:

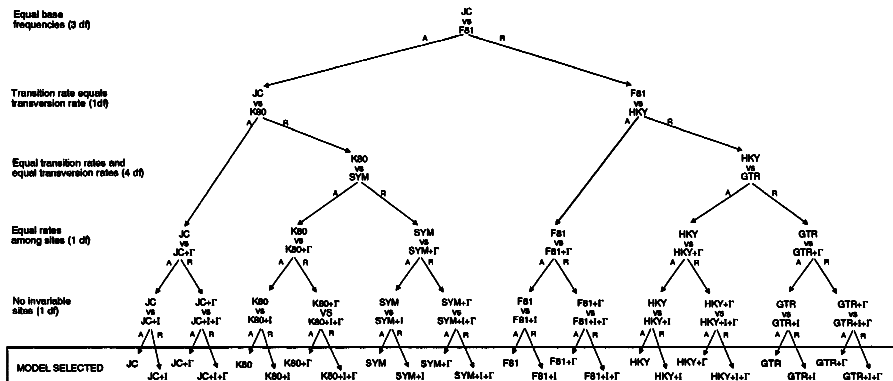
Bottom-up (δ LRT₁)

1. Start with the JC model and calculate its likelihood. This is the current model.
2. Calculate the likelihood of the alternative models differing by one assumption and perform the corresponding nested LRTs.
3. If any hypothesis or hypotheses are rejected, the alternative model corresponding to the LRT with the 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.
4. Repeat steps 2 and 3 until no hypothesis can be rejected. The current model is the selected model.

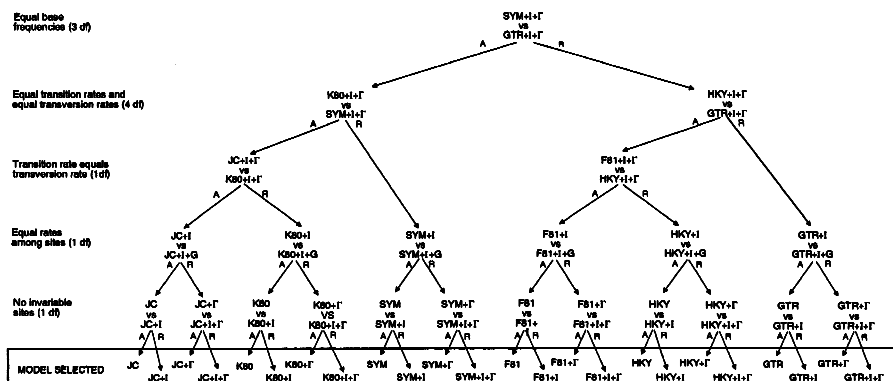
Top-down (δ LRT₂)

1. Start with the GTR + I + Γ model and calculate its likelihood. This is the current model.
2. Calculate the likelihood of the null models differing by one assumption and perform the corresponding nested LRTs.
3. If any hypothesis or hypotheses are not rejected, the null model corresponding to the LRT with biggest associated P-value becomes the current model. In the case of several equally biggest P-values, select the null model with the best likelihood.

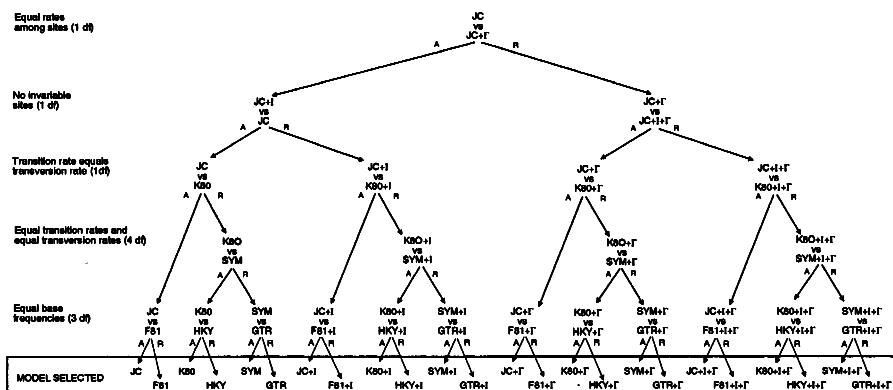
A ηLRT_1 (JC) $\pi \cdot \kappa \cdot \phi \cdot \Gamma \cdot I$



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



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



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

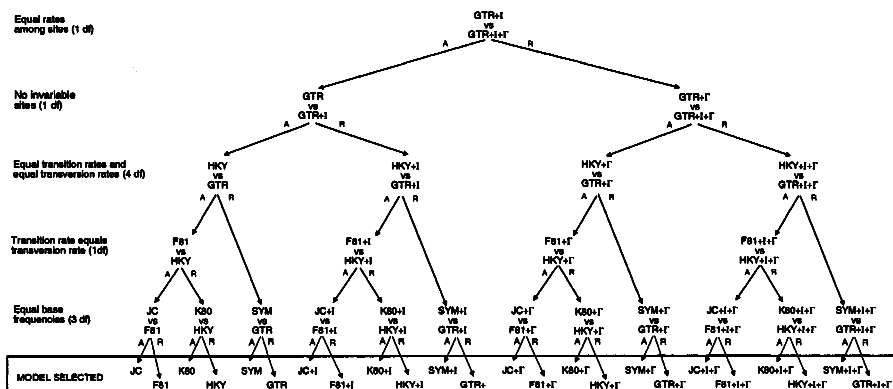


Fig. 2. Hierarchical likelihood ratio tests. Likelihood ratio tests are used to compare two different models at a time. The simpler model represents the null hypothesis. A model is accepted (A) or rejected (R) and the next LRT in the corresponding path is performed until a final model is selected. Several starting models (in parentheses) and several orders of parameter additions were evaluated (A–D: ηLRT_1 to ηLRT_4). The models of nucleotide substitution are: JC (Jukes and Cantor 1969), K80 (Kimura 1980), SYM (Zharkikh 1994), F81 (Felsenstein 1981), HKY (Hasegawa et al. 1985), and GTR (Tavaré 1986). Γ : rate heterogeneity among sites; I : proportion of invariable sites; df : degrees of freedom.

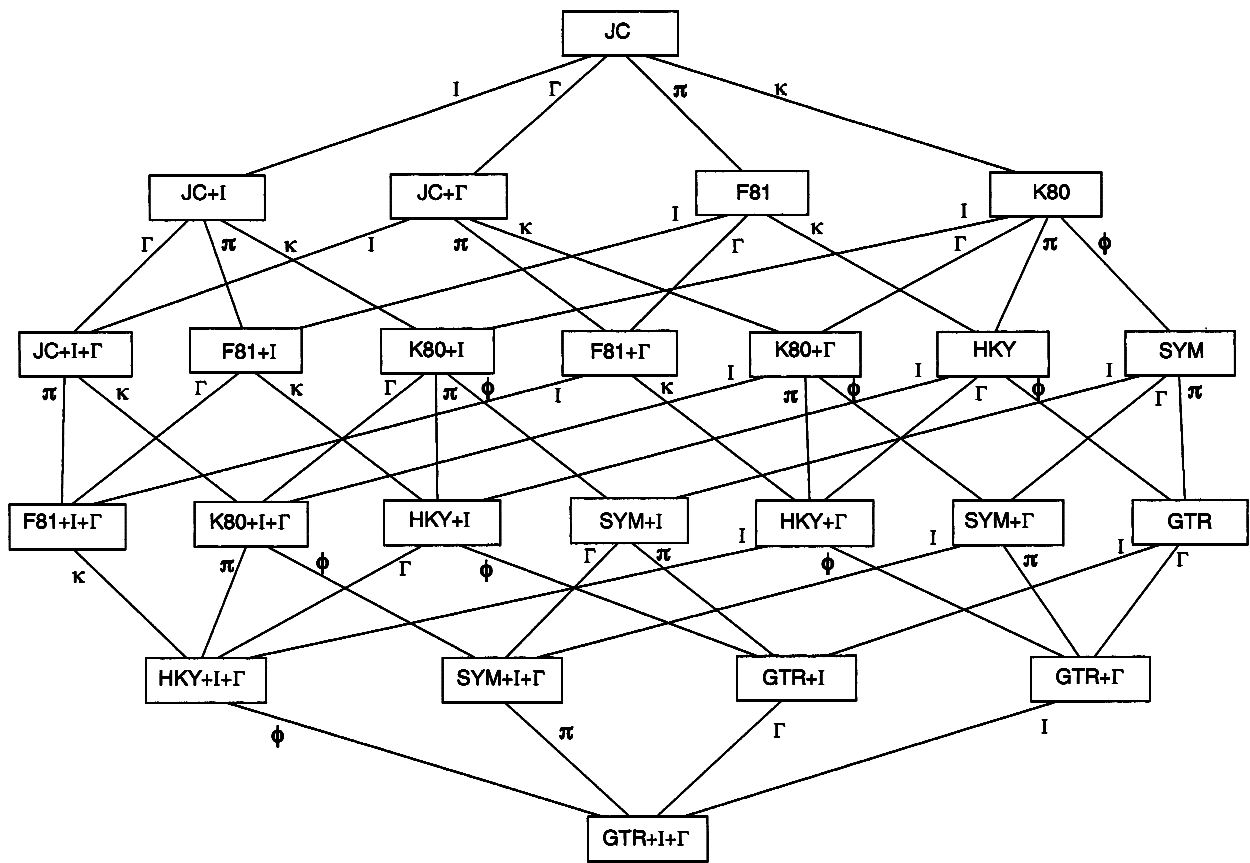


Fig. 3. Dynamic likelihood ratio tests. Starting with the simplest (JC) or the most complex model (GTR + I + Γ), LRTs are performed among the current model and several possible alternative models. π: base

frequencies. κ: transition/transversion bias. φ: substitution rates among nucleotides. Γ: rate heterogeneity among sites. I: proportion of invariable sites.

5. Repeat steps 2 and 3 until every hypothesis can be rejected. The current model is the selected model.

The alternative paths the algorithm can generate can be represented graphically (Fig. 3). Regarding multiple significance, it is not clear how to consistently apply the Bonferroni correction in this case. The number of tests performed may vary in each case. Also, several tests are performed but only some of them are actually considered. I 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 (AIC, Akaike (1974)). The AIC penalizes for the increasing number of parameters in the model:

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

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. Smaller AIC values indicate a better fit of the model to the data. I will use the term AIC_1 for this standard definition, as the penalty of the AIC_1 was empirically “tuned” to obtain an AIC_2 ($AIC_{2i} = -2 \ln L_i + 5 N_i$). This “tuning” was carried out by running several simulations and finding which penalty would increase the identification of the true model.

Bayesian information criterion (BIC, Schwarz (1974)). The BIC measures the relative support data give to different models:

$$BIC_i = -2 \ln L_i + N_i \ln n$$

where n is the sample size (sequence length). The smaller the BIC, the better the fit of the model to the data.

Results and Discussion

Accuracy of Model Fitting

The accuracy of the different model selection strategies was defined as the number of times a method recovered the correct model out of the 100 replicates, i.e. the probability of recovering the true model (Fig. 4). When the true model was simple (JC), most methods performed extremely well, with accuracies of 95–100%. However, the AIC_1 only recovered the true model about 50% of the time. In this case all hypotheses tested are true and type I error was on average around 4%. When the true model was of medium complexity (HKY + Γ), the accuracy of the ηLRT_1 method was over 90% in the presence of any short branches, but vastly decreased when all branches were medium or long, when the model recovered was more complex than the true one (due to type II error, see Fig. 5). The rest of the methods performed much worse

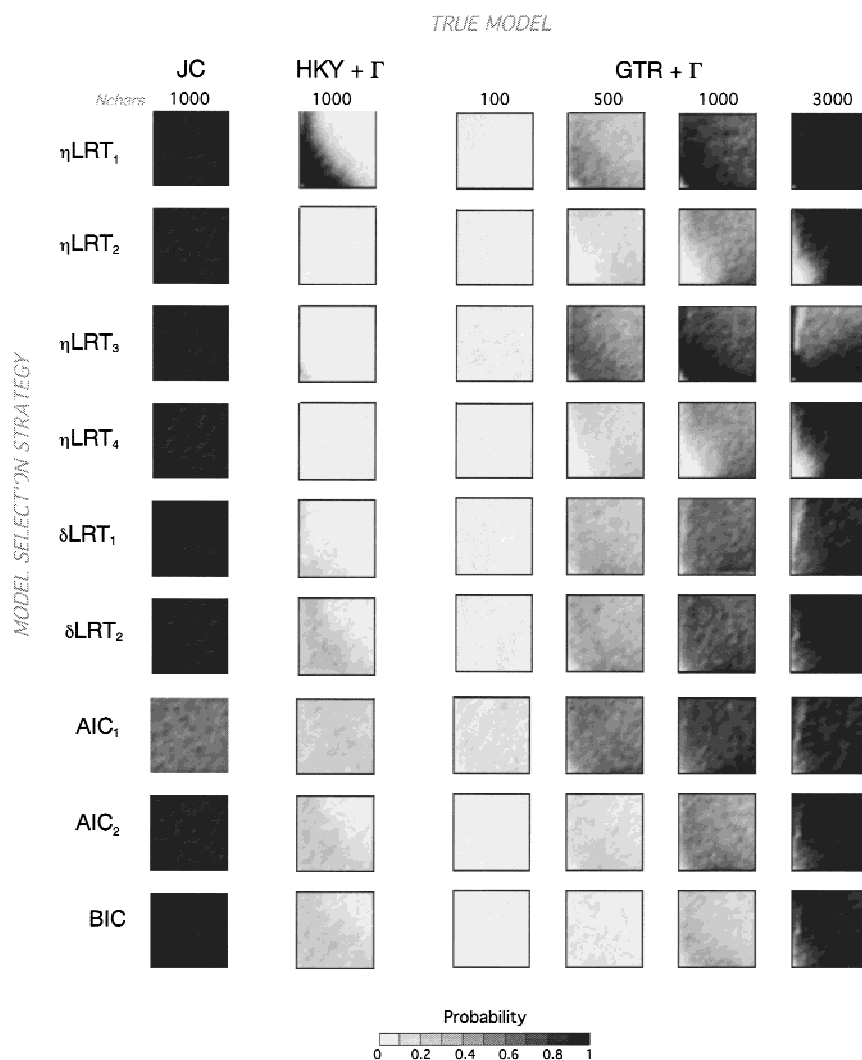


Fig. 4. Model selection accuracy. The probability of selecting the model that generated the data for nine different model selection strategies is presented. Colors indicate the number of times each strategy recovered the true model out of 100 replicates. The axes of each plot are the same as in Fig. 1. Each plot is a 16×16 array in which each point represents results from 100 simulated data sets. The P -values for the non-boundary LRTs were obtained by approximation to a standard χ^2 , while for the boundary LRTs a mixed χ^2 was used.

because an increase of type I, and especially, type II error (Fig. 5). When the true model was the complex GTR + Γ , accuracy increased with more characters, i.e., model selection methods are efficient. With 100 characters, all methods performed poorly (due mainly to type II error for the bottom-up approaches; data not shown), although AIC₁ did a little better than the rest. With 500 characters, the performance of η LRT₁, η LRT₃ (due to a big decrease in type II error) and AIC₁ increased considerably, especially for trees with short-medium branches. With 1000 characters, most methods recovered the true model 80–100% percent of the time over a large portion of the branch length space. The η LRT₁ and η LRT₃ strategies performed badly with short internal branches. With 3000 characters most methods showed performance values of 90–100%. The η LRT₃ showed an unexpected decrease over the upper part of the branch length region. The proportion of the parameter space more difficult was the lower left corner, where all branches are very short, and consequently, simpler models were inferred.

In general, the η LRT₁ performed better than the other strategies, due to its reduced probability of type I and II

error (Fig. 5). For as few as 4 taxa, 1000 characters were necessary for reasonable success when the true model was simple or complex. However, it has been shown that with 10 or more taxa, 500 characters are enough to obtain good performance (Posada and Crandall 2001a). Medium complexity models seem more difficult to recognize, and type II error becomes common. The AIC₁ is biased towards complex models, and that is the reason why it performed better than the other methods when GTR + Γ was the true model and only 100 or 500 characters were simulated, but also it is the reason it performed much worse than the other methods when the true model was the simple JC. Increasing its penalty (i.e., AIC₂) seems to correct for this bias. Failure to identify the GTR + Γ in the case of η LRT₂ and η LRT₄ is more often because of type I error, especially for trees with short internal branches (Fig. 5).

The η LRT₃ accuracy pattern was very unusual and deserves a tentative explanation. For the upper left region, this method increases its performance from 100 to 1000 characters, but suddenly drops with 3000 characters. If this pattern is an artifact, it might be due to a bad

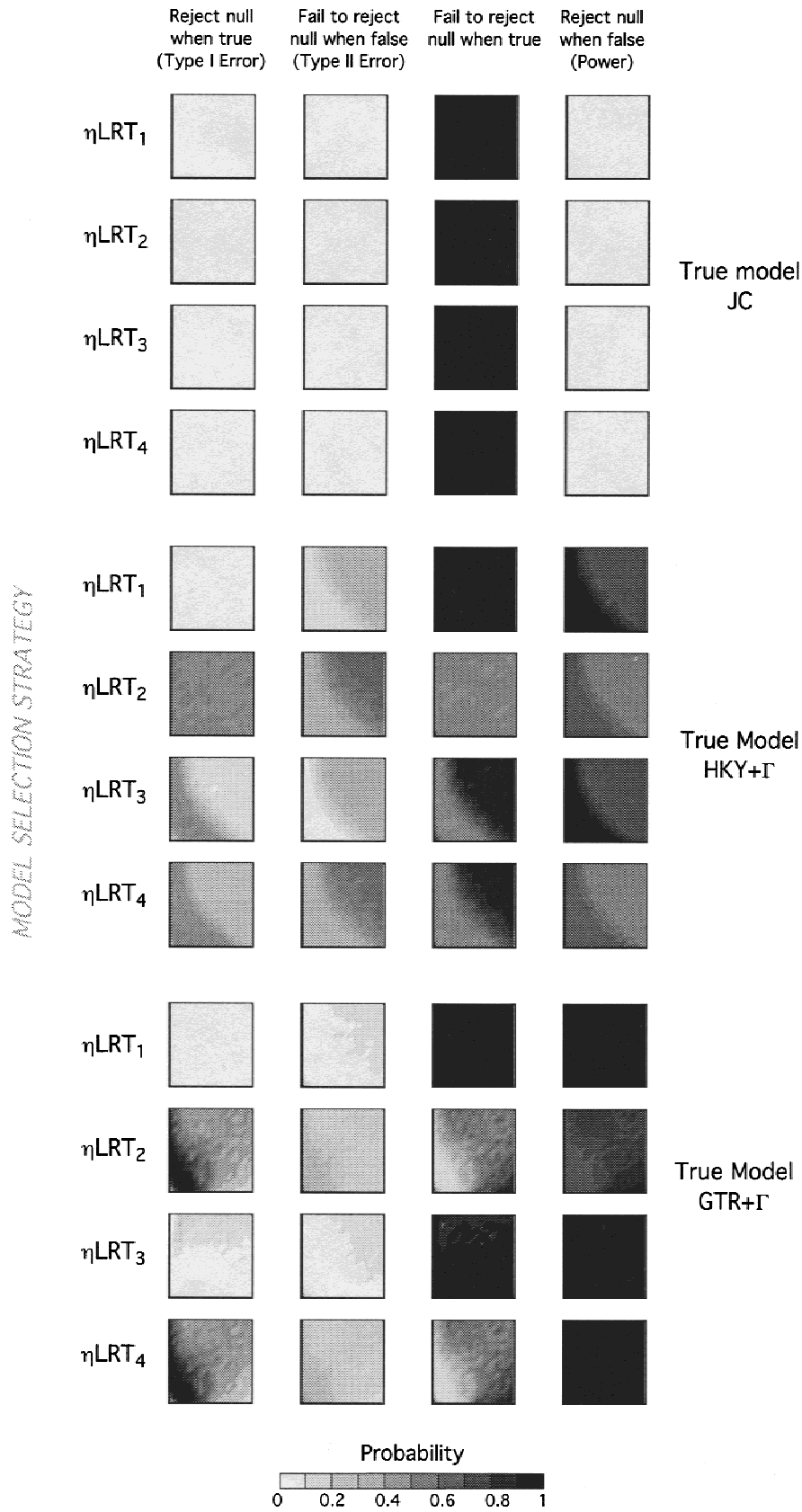


Fig. 5. Testing hypotheses. The conditional probability of rejecting or failing to reject the null hypothesis given that the null hypothesis is true or false is presented. The axes of each plot are the same as in Fig. 1. Each plot is a 16×16 array in which each point represents results from 100 simulated data sets with 1000 characters.

optimization of the likelihood scores for that region. It is known that, for small number of taxa, estimating the pattern of among-site rate variation is extremely difficult under mixed $I + \Gamma$ models (Gu et al. 1995; Sullivan et al. 1999). However, other methods use the same likelihood estimates without showing a similar behavior. It seems that there might be an alternative explanation. The model selected in this area is $GTR + I + \Gamma$, which, by looking at this hierarchy, indicates that the model $JC + \Gamma$ is being rejected incorrectly against $JC + I + \Gamma$ (type I error). While the $JC + \Gamma$ likelihood is always equal or worse than the $JC + I + \Gamma$ for all number of characters, and this difference increases with the number of characters, it only becomes significantly worse (and the parameter is incorrectly included in the model) with 3000 characters. Why this bias occurs precisely in this region of the branch length space is not obvious.

LRTs and Starting Models

Model selection procedures such as the LRTs necessarily start with one model, to which other models are compared. 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). In the context of nucleotide sequences, both approaches have been commonly used (e.g. Kelsey et al. 1999; Sullivan and Swofford 1997), although there is no evidence of that either strategy is superior (but see Posada and Crandall 2001a). For example, if we start with the simple JC model, we can add the parameter α (rate variation among sites) and check whether the likelihood improves significantly (i.e. perform the LRT JC vs. $JC + \Gamma$). If this is the case, rate variation is added to the model, while, if the likelihood does not improve significantly, rate variation is not added. In either case, the potential addition of other parameters is tested in a similar way. On the other hand, we can start with the complex $GTR + I + \Gamma$ and remove the α parameter from it, to test whether imposing the restriction of no rate variation decreases significantly its likelihood (i.e. perform the LRT $GTR + I + \Gamma$ vs. the $GTR + I$). If the likelihood does not decrease significantly, rate variation is removed from the model. If it does decrease, rate variation is kept in the model. Either way, the removal of additional parameters is tested in a similar fashion.

In this simulation, and for the hierarchical LRTs, the bottom-up strategies (ηLRT_1 and ηLRT_3) seem to perform better than the top-down ones (ηLRT_2 and ηLRT_4), although ηLRT_3 decreases its performance with 3000 characters. In general, bottom-up strategies have more power and smaller type I error, especially in the case of the ηLRT_1 . Top-down LRTs are more often subject to

type II error when the true model is $HKY + \Gamma$, but they are more prone to type I error when the true model was $GTR + \Gamma$ (Fig. 5). In the case of the dynamical LRTs, however, the top-down approach seems to work slightly better than the bottom-up approach.

LRTs and the Order of Parameter Addition/Removal

Given that a bottom-up or a top-down approach is taken, different parameters of the model may be added or removed, respectively, in a specific order. For example, we can test first for the addition of κ and later for the addition of π , or vice versa. 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 , where the parameter π is also present in both models.

If the presence of additional parameters does not affect the LRTs, we expect the order in which parameters are added or removed not to change the final model selected. Whelan and Goldman (1999), and Goldman and Whelan (2000) found that this is the case when the LRT is performed assuming the true model as the null hypothesis. However, this is not the situation here, as the null hypothesis will be the true model in only a few of the LRTs performed. Nor is it the case with real data, as the true model is unknown. On the other hand, Zhang (1999) showed that LRTs of the transition/transversion bias or rate variation are affected by the presence or absence of other parameters. For example, the failure to take in account unequal base frequencies led, in Zhang's simulations, to the rejection of the null hypothesis of no transition bias much more often than expected. Using an empirically generated phylogeny, Cunningham et al. (1998) observed that the choice of the best-fit models was affected by the order of addition of parameters, but their conclusion would be the opposite if they had corrected for type I error in their Table 1. In these simulations, the patterns of accuracy of ηLRT_1 and ηLRT_3 (except for 3000 characters), and of ηLRT_2 and ηLRT_4 were very similar, which indicates that the order in which parameters were added or removed to or from model had a weak effect. Indeed, other hierarchies of LRTs could exist where this might not be true. Only in the case of $HKY + \Gamma$ being the true model, testing first for equal base frequencies (ηLRT_1) reduced the probability of type I error.

LRT Distribution and Mixed χ^2

The wrong use of a standard χ^2 distribution instead of the appropriate mixed χ^2 distribution in the case of boundary

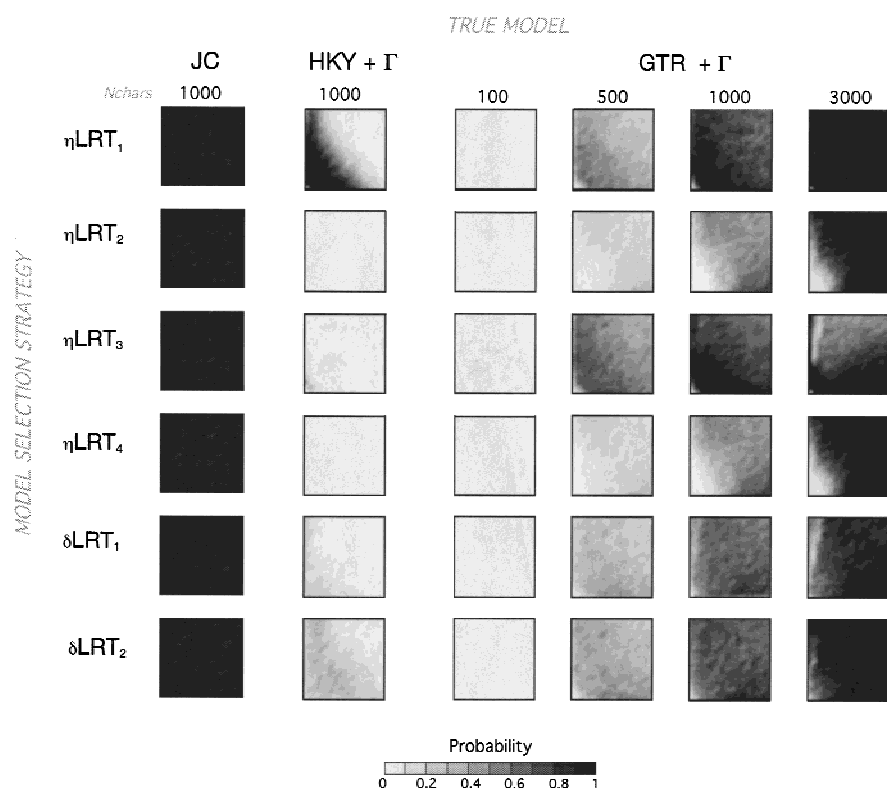


Fig. 6. Model selection accuracy using a standard χ^2 for all tests. The probability of selecting the model that generated the data for the LRTs selection strategies is presented. Colors indicate the number of times each strategy recovered the true model out of 100 replicates. The axes of each plot are the same as in Fig. 1. Each plot is a 16×16 array in which each point represents results from 100 simulated data sets. The P -values for all LRTs were obtained by approximation to a standard χ^2 distribution.

parameters, did not have a significant influence on the performance of the different model-fitting strategies. In fact, the accuracy patterns in the branch space were almost indistinguishable (compare Figs. 4 and 6). Although the standard χ^2 distribution may be significantly different from the true LRT distribution in the boundary case, the P -values obtained in LRTs of evolutionary hypotheses are often so small that this bias does not compromise previous analyses that used the incorrect χ^2 approximation. Indeed, the appropriate χ^2 distribution should be used in each case.

Relevance of Models

The relevance of models of nucleotide substitution to molecular evolution and phylogenetic studies is well documented. Through the selection of a model, the process of molecular evolution is characterized. Furthermore, the use of appropriate models is especially critical for parameter estimation.

The use of naively simple models, even when recovering the correct topology, can result in wrong estimation of parameters, especially when rate variation is ignored (Yang 1996a). Simple models tend to underestimate branch lengths (Adachi and Hasegawa 1995; Tamura 1992; Yang et al. 1994), sequence distances (Golding 1983), transition-transversion ratios (Wakeley 1994; Yang et al. 1994; Yang et al. 1995), or strength of rate variation among sites (Yang et al. 1995). Simple models

may be conservative in LRTs of the molecular clock hypothesis (Zhang 1999).

In general, more complex models will fit the data better than simpler ones. However, when using complex models a large number of parameters need to be estimated from the same amount of data, and more error is included in each estimate. Errors in parameter estimation may compromise phylogenetic accuracy, especially for small data sets. Over-parameterization may lead to a loss of discriminatory power.

Phylogenetic methods often perform worse when the model of evolution assumed is incorrect (Bruno and Halpern 1999; Felsenstein 1978; Huelsenbeck 1995; Huelsenbeck and Hillis 1993). When substitution rates vary among lineages, the use of an appropriate model is of utmost importance for obtaining a correct tree topology (Philippe and Germot 2000; Takezaki and Gojobori 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 always straightforward (Fukami-Kobayashi and Tateno 1991; Gaut and Lewis 1995; Russo et al. 1996; Takahashi and Nei 2000; Yang et al. 1995). Cases where the use of wrong models increases phylogenetic performance are the exception (e.g. Posada and Crandall 2001b; Yang 1997), and they might rather represent a bias towards the true tree associated with violated assumptions (Bruno and Halpern 1999). Finally, simple models tend to suggest that a tree is significantly supported when it cannot be (Yang et al. 1994), and can

cause rapidly evolving taxa to be confidently but incorrectly grouped (Bruno and Halpern 1999).

Conclusions

The choice of appropriate models is thought to be especially important when there is large branch length variation. Here it is shown that there are model selection procedures that perform well over the branch length space given enough characters. The order in which parameters were added or removed to a model did not have an effect, but the starting model influenced model selection. A specific hierarchy of LRTs performed slightly better than other hierarchies, or than the AIC or BIC strategies. The χ^2 distribution used for the LRTs did not seem to make a difference, although the proper distribution should be used in each case.

Models are necessary and useful simplifications, and none of them are exactly correct when dealing with real data. Even the best-fit model is far from the true model underlying the evolution of the sequences under study. Although these simulation results pertain to a perfect fit between models and data, they offer us some useful insights. If model-fitting procedures are able to recognize some features of the process of nucleotide substitution in simulated data sets (equal or unequal base frequencies, rate variation, etc), it can be expected that the same methods will recognize these same features in real data sets, selecting more realistic, although still imperfect, models. And as more assumptions of a method are justified, the performance of that method will become better and better. Indeed, the identification of best-fit models is of utmost importance to recognize and understand the process of molecular evolution.

Although it is clear that the use of a correct model improves parameter estimation, the relationship of models to phylogeny estimation is not straight forward. Appropriate studies are needed in order to understand whether the use of best-fit models actually improves phylogenetic reconstruction.

Model fitting should be routine in sequence studies. A program facilitating this task, Modeltest (Posada and Crandall 1998), can be downloaded free at http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm.

Acknowledgments. Keith Crandall and Jack Sites commented on the manuscript. Thanks to Andrew Rambaut, Nick Goldman and Simon Whelan for helpful discussion. Ziheng Yang and two anonymous reviewers provided detailed and useful comments. This work was supported by a BYU Graduate Studies Award and NSF DEB 0073154.

References

Adachi J, Hasegawa M (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 WJ, Halpern AL (1999) Topological bias and inconsistency of maximum likelihood using wrong models. *Mol Biol Evol* 16:564–566
- Corneli PS, Ward RH (2000) Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation. *Mol Biol Evol* 17:224–234
- Cunningham CW, Zhu H, Hillis DM (1998) Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 52:978–987
- 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, Simon C, Sullivan J, Swofford DL (1997) Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in *Colombola*. *J Mol Evol* 44:145–158
- Fukami-Kobayashi K, Tateno Y (1991) Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J Mol Evol* 32:79–91
- Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol Biol Evol* 12:152–162
- Golding BG (1983) Estimates of DNA and protein sequence divergence: a examination of some assumptions. *Mol Biol Evol* 1:125–142
- Goldman N (1993a) Simple diagnostic statistical test of models of DNA substitution. *J Mol Evol* 37:650–661
- Goldman N (1993b) Statistical tests of models of DNA substitution. *J Mol Evol* 36:182–198
- Goldman N, Whelan S (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol Biol Evol* 17:975–978
- Gu X, Fu Y, Li W (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* 12:546–557
- 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, Kishino K, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174
- Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. *Syst Biol* 44:17–48
- Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst* 28:437–466
- Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. *Syst Biol* 42:247–264
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HM (ed) *Mammalian Protein Metabolism*. Academic Press, New York, pp 21–132
- Kelsey CR, Crandall KA, Voevodin AF (1999) Different models, different trees: the geographic origin of PTLV-I. *Mol Phylogenet Evol* 13:336–347
- Kendall M, Stuart A (1979) *The advanced theory of statistics*. Charles Griffin, London
- Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120
- Leitner T, Kumar S, Albert J (1997) Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J Virol* 71:4761–4770

- Liò P, Goldman N (1998) Models of molecular evolution and phylogeny. *Genome Res* 8:1233–44
- Morozov P, Sitnikova T, Churchill G, Ayala FJ, Rzhetsky A (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. In: Crandall KA (ed) *The evolution of HIV*. Johns Hopkins University Press, Baltimore, pp 122–152
- Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol* 17:798–803
- Penny D, Lockhart PJ, Steel MA, Hendy MD (1994) The role of models in reconstructing evolutionary trees. In: Scotland RW, Siebert DJ, Williams DM (eds) *Models in phylogenetic reconstruction*. Clarendon Press, Oxford, pp 211–230
- Philippe H, Germot A (2000) Phylogeny of eukaryotes based in ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol Biol Evol* 17:830–834
- Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–818
- Posada D, Crandall KA (2001a) Selecting the best-fit model of nucleotide substitution. *Syst Biol* (in press)
- Posada D, Crandall KA (2001b) Simple (wrong) models for complex trees: empirical bias. *Mol Biol Evol* 18:271–275
- Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosciences* 13:235–238
- Russo CAM, Takezaki N, Nei M (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol Biol Evol* 13:525–536
- Saitou N, Nei M (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
- Self SG, Liang K-L (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 82:605–610
- Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol* 17:839–850
- Sullivan J, Markert JA, Kilpatrick CW (1997) Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst Biol* 46:426–440
- Sullivan J, Swofford D, Naylor G (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol Biol Evol* 16:1347–1356
- Sullivan J, Swofford DL (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J Mamm Evol* 4:77–86
- Swofford DL (1998) PAUP* Phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland
- Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol Biol Evol* 17:1251–1258
- Takezaki N, Gojobori T (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. In: Miura RM (ed) *Some mathematical questions in biology—DNA sequence analysis*. American Mathematical Society, Providence, pp 57–86
- Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. *Mol Biol Evol* 11:436–442
- Wang X-Q, Tank DC, Sang T (2000) Phylogeny and divergence times in *Pinaceae*: Evidence from three genomes. *Mol Biol Evol* 17:773–781
- Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol* 16:1292–1299
- 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) Estimating the pattern of nucleotide substitution. *J Mol Evol* 39:105–111
- Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314
- Yang Z (1996a) Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol Evol* 11:367–372
- Yang Z (1996b) Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol* 42:587–596
- Yang Z (1997) How often do wrong models produce better phylogenies? *Mol Biol Evol* 14:105–108
- Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* 11:316–324
- Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst Biol* 44:384–399
- Zhang J (1999) Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol Biol Evol* 16:868–875
- Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol* 39:315–329