Bayesian Model Adequacy and Choice in Phylogenetics

Jonathan P. Bollback

Department of Biology, University of Rochester

Bayesian inference is becoming a common statistical approach to phylogenetic estimation because, among other reasons, it allows for rapid analysis of large data sets with complex evolutionary models. Conveniently, Bayesian phylogenetic methods use currently available stochastic models of sequence evolution. However, as with other model-based approaches, the results of Bayesian inference are conditional on the assumed model of evolution: inadequate models (models that poorly fit the data) may result in erroneous inferences. In this article, I present a Bayesian phylogenetic method that evaluates the adequacy of evolutionary models using posterior predictive distributions. By evaluating a model's posterior predictive performance, an adequate model can be selected for a Bayesian phylogenetic study. Although I present a single test statistic that assesses the overall (global) performance of a phylogenetic model, a variety of test statistics can be tailored to evaluate specific features (local performance) of evolutionary models to identify sources failure. The method presented here, unlike the likelihood-ratio test and parametric bootstrap, accounts for uncertainty in the phylogeny and model parameters.

Introduction

The results of any phylogenetic analysis are conditional on the chosen model. Models that fit the data poorly can lead to erroneous or consistently biased inferences of phylogeny (Felsenstein 1978; Huelsenbeck and Hillis 1993; Gaut and Lewis 1995; Sullivan and Swofford 1997; Bruno and Halpern 1999). For example, a model that assumes equal rates across sites (rate homogeneity) may result in inconsistent inferences even if all other parameters of the model are correct (Gaut and Lewis 1995). The tremendous increase in computational power over the last few years has resulted in the development of a bewildering assortment of models of sequence evolution for researchers to choose from (for a review see Swofford et al. 1996; Huelsenbeck and Bollback 2001). Despite the potentially severe effects of poor model fit, inadequate models were used in four out of five recent articles in a primary systematics journal (Posada and Crandall 2001).

The parameters of a phylogenetic model describe the underlying process of sequence evolution. The maximum likelihood and Bayesian methods of statistical inference both estimate these parameters (including the topology) using the likelihood function, a quantity hereafter referred to as $p(\mathbf{X}|\theta)$ (which should be read as the probability of the data, $\mathbf{X}$, conditioned on a specific combination of model parameters, θ; more formally, the likelihood is proportional to the probability of observing the data). In maximum likelihood, inferences are based on finding the topology relating the species, branch lengths, and parameter estimates of the phylogenetic model that maximize the probability of observing the data. Bayesian inferences, on the other hand, are based on the posterior probability of the topology, branch lengths, and parameters of the phylogenetic model conditioned on the data. Posterior probabilities can be calculated using Bayes's theorem.

Determining which model is best suited to the data can be divided into two distinct criteria—model adequacy (or assessment) and model choice (or selection). Model adequacy is an absolute measure of how well a model under scrutiny fits the data. Model choice, on the other hand, is a relative measure: the best fitting model from those available is chosen. Although a model may be the best choice, it may be, by absolute standards, inadequate. The likelihood-ratio test (LRT) and Bayes factors are model choice tests: they measure relative merits of competing models but reveal little about their overall adequacy. (Although formally, the LRT evaluates the adequacy of a model [Goldman 1993], in practice it is used as a model choice strategy.) Although model adequacy and choice are distinct but related criteria, they are often evaluated simultaneously by comparing nested models which differ by a single parameter (see Goldman 1993). Ideally, we would use only adequate models for a phylogenetic analysis, but in practice we often settle for the best available model. In fact, most models appear to be poor descriptions of sequence evolution (Goldman 1993).

How does one choose an adequate phylogenetic model? Traditional maximum likelihood approaches to model selection employ the LRT (for hierarchically nested models) or the parametric bootstrap (for nonnested models) (Goldman 1993). Both methods depend on a particular topology, often generated by a relatively fast method such as parsimony or neighbor-joining. (See Posada and Crandall [2001] for an analysis of the effects of topology choice on model selection using the LRT, the Akaike information criterion [AIC; Akaike 1974], and the Bayesian information criterion [BIC; Schwarz 1974].) The LRT evaluates the merits of one model against another by finding the ratio of their maximum likelihoods. For nested models, the LRT statistic is asymptotically χ^2-distributed with q degrees of freedom (Wilks 1938), permitting comparison with standard χ^2 tables to determine significance. Unfortunately, significance cannot be evaluated in this way when models are not nested, or the null fixes parameters of the alternative

Key words: phylogenetics, Bayesian inference, model determination, model selection, model adequacy, posterior predictive densities, posterior predictive simulations.

Address for correspondence and reprints: Department of Biology, University of Rochester, Rochester, New York 14627. E-mail: bollback@brahms.biology.rochester.edu.

Mol. Biol. Evol. 19(7):1171–1180. 2002

model at the boundary of the parameter space, because the regularity conditions of the χ^2 are not satisfied.

The parametric bootstrap, alternatively, is not constrained by regularity conditions allowing comparison of nonnested models but is time-intensive and may require researchers to write computer simulations to approximate the null distribution. Unfortunately, this computationally expensive approach, the AIC, and the BIC remain the only current methods (apart from simple inspection of the log likelihood scores) to compare nonnested likelihood models.

The results of the LRT and the parametric bootstrap are conditional on the topology and model parameters chosen to conduct the test. The assumed topology may be chosen using a fast method, such as parsimony, known to be inconsistent under certain conditions (Felsenstein 1978). The branch lengths and model parameters (such as transition/transversion bias) are generally maximum likelihood–point estimates conditional on the assumed topology. Ideally, a statistical method should minimize the number of assumptions made.

Bayesian methods offer an efficient means of reducing this reliance on assumptions. These methods can accommodate uncertainty in topology, branch lengths, and model parameters. For example, Suchard, Weiss, and Sinsheimer (2001) recently developed a Bayesian method of model selection that uses reversible jump Markov chain Monte Carlo (MCMC) and employs Bayes factors for comparing models. This approach is a Bayesian analog of the LRT: the Bayes factor indicates relative superiority of competing models by evaluating the ratio of their marginal likelihoods. In this approach, prior probability distributions of the models must be proper but allowably vague. If the information contained in the data about model adequacy is small, then the priors will determine the outcome of the test. In this situation, most of the posterior will be placed on the more complicated model (Carlin and Chib 1995). Although the method of Suchard, Weiss, and Sinsheimer (2001) allows comparison of models without strict dependence on a particular set of assumptions, like traditional likelihood approaches, it does not explicitly evaluate the absolute merits of a model. The chosen model may well be severely inadequate.

Here, I present a Bayesian method using posterior predictive distributions to explicitly evaluate the overall adequacy of DNA models of sequence evolution. The approach I use, posterior predictive check by simulation (Rubin 1984; Gelman, Dey, and Chang 1992; Gelman et al. 1995; Gamerman 1997), is a Bayesian analog of classical frequentist methods such as the parametric bootstrap or randomization tests (Rubin 1984). A similar approach has been used recently to test molecular evolution hypotheses (Huelsenbeck et al. 2001; Nielsen and Huelsenbeck 2001; Nielsen 2002). The rationale motivating this approach is that an adequate model should perform well in predicting future observations. In the absence of future observations, predicted observations are simulated from the posterior distribution, under the model in question. These predicted data are then compared with the original data using a test statistic that summarizes the differences between them. Careful evaluation of the model parameters permits enhancement (addition of parameters) or simplification (elimination of irrelevant parameters) of the model to improve its overall fit to the data. Here, I use the multinomial test statistic to evaluate overall adequacy of phylogenetic models.

Materials and Methods

Models of Sequence Evolution

Models of sequence evolution used in phylogenetics model nucleotide substitutions as a stochastic process, most of which are time-homogenous, time-reversible Markov processes. Reversibility of a model is satisfied when the rate of forward and reverse changes are equal, such that $\pi_i q_{ij} = \pi_j q_{ji}$. However, general, nonreversible substitution models have also been developed and explored in a variety of phylogenetic contexts (Yang 1994; Huelsenbeck, Bollback, and Levine 2002). For the sake of brevity, this study restricts itself to reversible models, but the method is easily extended to nonreversible, time-heterogeneous, or other classes of models. Four models will be used in this study (1) the general-time-reversible model (GTR; Tavaré 1986), (2) Hasegawa-Kishino-Yano model (HKY85; Hasegawa, Kishino, and Yano 1985), (3) Kimura's two-parameter model (K2P; Kimura 1980), and (4) Jukes-Cantor model (JC69; Jukes and Cantor 1969). These represent the most commonly implemented models in the phylogenetic literature. The first three models are special cases of the GTR.

The GTR is the most general model of DNA sequence evolution allowing for different rates for each substitution class and accommodating unequal base frequencies. The instantaneous rate matrix, **Q**, for this model is:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} — & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & — & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & — & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & — \end{pmatrix}. \qquad (1)$$

The diagonals of the matrix are set such that the rows each sum to 0. When the rates in the aforementioned matrix are constrained such that $b = e = \kappa$, and $a = c = d = f = 1$, the GTR model collapses into the HKY85 model with the following rate matrix:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} — & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & — & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & — & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & — \end{pmatrix} \qquad (2)$$

where κ is a rate parameter describing a transition-transversion bias. If the HKY85 model is constrained such that the base frequencies are equal ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$), this model collapses into the K2P model with the following rate matrix:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} — & 1 & \kappa & 1 \\ 1 & — & 1 & \kappa \\ \kappa & 1 & — & 1 \\ 1 & \kappa & 1 & — \end{pmatrix}. \tag{3}$$

Finally, if the K2P model is constrained such that $\kappa = 1$, it collapses into the JC69 model with equal rates between all substitution classes.

Using instantaneous rates, substitution probabilities for a change from nucleotide i to j over a branch of length v can be calculated as $\mathbf{P} = \{P_{ij}\} = e^{\mathbf{Q}v}$. In the case of the JC69, K2P, and HKY85 models, closed-form analytical solutions for the substitution probabilities are available (Swofford et al. 1996). For the GTR model, closed-form solutions do not exist, and standard numerical linear algebra approaches are employed to exponentiate the term $\mathbf{Q}v$ (Swofford et al. 1996). With a matrix of substitution probabilities available, calculation of the likelihood is straightforward using Felsenstein's (1981) pruning algorithm.

Posterior Predictive Simulations

In evaluating a model's adequacy, we would like to know how well it describes the underlying process that generated the DNA sequence data in hand. Therefore, an ideal model should perform well in predicting future observations of the data. In practice, future observations are unavailable to researchers at the time of data analysis. However, surrogate future observations under the model being tested can be simulated by sampling from the joint posterior density of trees and model parameters (hence, posterior predictive simulations; Rubin 1984). Because of the complexity of the phylogeny problem—the large number of possible combinations of topology, branch lengths, and model parameters—the posterior density cannot be evaluated analytically. Luckily, we can use numerical methods to obtain an approximation of this density ($\hat{p}[\theta|\mathbf{X}]$) using the MCMC technique (Li 1996; Mau 1996; Mau and Newton 1997; Yang and Rannala 1997; Larget and Simon 1999; Mau, Newton, and Larget 1999; Newton, Mau, and Larget 1999; Huelsenbeck and Ronquist 2001).

Model assessment using this approach requires approximating the following predictive density:

$$p(\mathbf{X}\,|\,\mathbf{X}_{\text{obs}}) = \sum_{k=1}^{B(s)} \int_{v_k} \int_{\theta} p(\mathbf{X}\,|\,\theta, v_k, \tau_k, \mathbf{X}_{\text{obs}}) p(\theta, v_k, \tau_k\,|\,\mathbf{X}_{\text{obs}})\, dv_k\, d\theta. \tag{4}$$

Trees are labeled $\tau_1, \tau_2, \ldots, \tau_{B(s)}$, where $B(s) = (2s - 5)!/2^{s-3}(s - 3)!$ is the number of unrooted trees for s species. For all unrooted topologies, $B(s)$, we integrate over branch lengths (v_k) and parameters of the model (θ). Evaluation of this density requires knowledge of the joint posterior density, but once an approximation of the joint posterior density of model parameters and topologies, $\hat{p}(\theta, v, \tau|\mathbf{X})$, has been obtained, the posterior predictive density (eq. 4) can be approximated numerically by Monte Carlo simulation in the following way (1) Make a random draw from the joint posterior distribution of trees and model parameters, under the model being tested. (In practice, this can be accomplished by sampling the posterior output of a program that approximates posterior distributions, such as MrBayes [Huelsenbeck and Ronquist 2001]). (2) Using these random draws (which include values for the parameters of the substitution process, topology, and branch lengths) and the model being tested, simulate a data set, $\mathbf{X}_1$, of the same size as the original data set. (3) Repeat steps 1 and 2 N times to create a collection of data sets, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. (4) These simulated data sets are a numerical Monte Carlo approximation of the posterior predictive density (eq. 4):

$$\hat{p}(\mathbf{X}\,|\,\mathbf{X}_{\text{obs}}) = \frac{1}{N} \sum_{k=1}^{N} \sum_{j=1}^{N} p(\mathbf{X}\,|\,\theta_j, v_k, \tau_k, \mathbf{X}_{\text{obs}}). \tag{5}$$

Test Statistics

We now have an approximation of the posterior predictive density of the data, simulated under the phylogenetic model being scrutinized. But we are still left with the following problem: how can we use this posterior predictive distribution to assess the phylogenetic model's adequacy? This requires a descriptive test statistic (or discrepancy variable; Gelfand and Meng 1996) that quantifies discrepancies between the observed data and the posterior predictive distribution. The test statistic is referred to as a realized value when summarizing the observed data. An appropriate test statistic can be defined to measure any aspect of the predictive performance of a model (Gelman et al. 1995). I use the general notation $T(\cdot)$, where $\cdot$ refers to the variable being tested. To use this statistic, calculate $T(\cdot)$ (an example of the proposed statistic will be shown later), for the posterior predictive data sets to arrive at an approximation of the predictive distribution of this test quantity. This distribution can then be compared with the realized test statistic, which is calculated from the original data.

To asses how well a phylogenetic model is able to predict future nucleotide observations (overall adequacy), a test statistic that quantifies the frequency of site patterns is appropriate. Here I use the multinomial test statistic to summarize the difference between the observed and posterior predictive frequencies of site patterns (Goldman 1993). A minor limitation of the multinomial is its assumption of independence among sites, restricting its application to phylogenetic models that assume independence. Deviations in the posterior predictive frequency of site patterns from the observed occur because the phylogenetic model is an imperfect description of the evolutionary process. If the evolutionary process that generated the data exhibits a GC bias, for instance, then site patterns containing a predominance of these bases will be overrepresented. An adequate model should be able to predict this deviation, given the information contained in the original sequence data.

The multinomial test statistic of the data, $T(\mathbf{X})$, is calculated in the following way. Let $\xi_{(i)}$ be the ith unique observed site pattern and $N_{\xi(i)}$ the number of instances this pattern is observed. For a total number of N sites, $S = 4^k$ possible site patterns, and n unique site patterns observed, the multinomial test statistic ($T[\mathbf{X}]$) can be calculated as follows:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\xi(i)}}{N}\right)^{N_{\xi(i)}}\right). \tag{6}$$

This is the natural log density of the maximum likelihood estimator of the multinomial. Alternatively, for ease of computation, equation 6 can be rewritten as:

$$T(\mathbf{X}) = \left(\sum_{\xi\in S}^{n} N_{\xi}\ln(N_{\xi})\right) - N\ln(N). \tag{7}$$

To illustrate the multinomial test statistic, let us find the realized $T(\mathbf{X})$ for $k = 4$ sequences with $N = 10$ sites from the following hypothetical aligned matrix of DNA sequences:

$$\mathbf{X} = \begin{pmatrix} \text{AAATCCAGGG} \\ \text{AAACCCAACA} \\ \text{AATCGGTTCA} \\ \text{AATCGGTATT} \end{pmatrix}.$$

There are seven unique site patterns in the matrix. Site patterns $x'_{1,2} = \{AAAA\}$, $x'_{3,7} = \{AATT\}$, and $x'_{5,6} = \{CCGG\}$ are observed twice; the four remaining site patterns are observed only once each. The realized test statistic for this data, using equation 7 is then:

$$T(\mathbf{X}) = 3 \cdot 2\ln(2) + 4 \cdot 1\ln(1) - 10\ln(10) = -18.867.$$

Numerous test statistics can be formulated, but to be useful these test statistics should represent a relevant summary of the model parameters and data.

Predictive *P* Values

Classical frequency statistics rely on tail-area probabilities to assign statistical significance; values that lie in the extremes of the null distribution of the test quantity are considered significant. Under classical statistics, the distributions are conditioned on point estimates for model parameters. Predictive densities, on the other hand, are not. Because values are sampled from the posterior distribution of model parameters and trees, they are sampled in proportion to their marginal probabilities. This sampling scheme allows them to be treated as nuisance parameters—values not of direct interest—and to be integrated out. The predictive distribution of the test statistic allows us to evaluate the posterior predictive probability of the model. The posterior predictive P value for the test statistic is:

$$P_T = \frac{1}{N}\sum_{i=1}^{N} I(T(\mathbf{X}_i) \geq T(\mathbf{X})), \tag{8}$$

where I is an indicator function that takes on the value 1 when the equality is satisfied and 0 otherwise, $T(\mathbf{X}_i)$ the multinomial test statistic for the ith simulated data set, and $T(\mathbf{X})$ the realized test statistic. Probabilities less than the critical threshold, say $\alpha = 0.05$, suggest that the model under examination is inadequate and should be rejected or refined. Predictive P values are interpreted as the probability that the model would produce with as extreme a test value as that observed for the data (Gelman et al. 1995). For an adequate model, the predictive distribution of $T(\mathbf{X})$ should be centered around the realized test statistic (i.e., $P_T = 0.5$). This approach evaluates the practical fit of the model to a data set; inclusion of additional taxa or new sequences requires a revaluation of the model and its fit.

Simulations

To determine the utility and power of this approach, I simulated 300 data sets under a variety of models and parameter values (see table 1 for a description of the specifics for each analysis). For all data sets the true (model data was simulated under) and the JC69 models are examined. Briefly, I performed three sets of simulations to examine (1) the overall model adequacy, (2) the effects of sequence divergence, and (3) the model sensitivity. I discuss each of these in turn subsequently.

To test overall model adequacy, I simulated data sets of 500, 2,000, and 4,000 sites under the GTR model. The parameters of the model, for each data set, were assigned in the following way (1) instantaneous rates were randomly chosen from the uniform interval, U(0.0, 6.0], (2) values for the base frequencies were drawn from a Dirichlet distribution with parameters (α_A, α_C, α_G, α_T) randomly chosen from the interval U[1.0, 4.0], and (3) the overall substitution rate (m) was fixed at 0.5. Trees were simulated under the birth-death process as described by Rannala and Yang (1996). Speciation (λ), extinction (μ), and taxon sampling (ρ) rates were fixed at 2.0, 1.0, and 0.75, respectively.

To test the effects of sequence divergence, I simulated data sets of 2,000 sites under the GTR model. Parameters of the model were chosen as in the test of overall adequacy. For all data sets the tree in figure 1 was used. The overall substitution rate was varied from low ($m = 0.1$) to high ($m = 0.75$) divergence.

Finally, for the model sensitivity analyses, data sets of 1,000 and 5,000 sites were simulated under the K2P model. Violation of the JC69 model's assumptions varied from none ($\kappa = 1$) to extreme ($\kappa = 12$). The tree in figure 1, with an intermediate substitution rate ($m = 0.5$), was used to simulate data.

Power Analysis

Under the posterior predictive simulation approach the null hypothesis is that the model is an adequate fit to the data. A model is rejected if the realized test statistic is less than the critical value ($\alpha = 0.05$). Otherwise the model was accepted. The fraction of times the null model is accepted falsely is an estimate of Type II error rate, β—the complement ($1 - \beta$) is the power of a test. The power of the multinomial test statistic to reject a false model is determined by the analysis of all the data sets described previously using the JC69 model.

Table 1
Simulation Conditions

Test	True Model	True Model Parameters	True Tree	Number Taxa	Number Characters	Substitution Rate (m)	Models Tested	Replicates
Overall adequacy	GTR	a, b	c	10	500	0.50	GTR	20
	GTR	a, b	c	10	2,000	0.50	GTR	20
	GTR	a, b	c	10	4,000	0.50	GTR	20
	GTR	a, b	c	10	500	0.50	JC69	20
	GTR	a, b	c	10	2,000	0.50	JC69	20
	GTR	a, b	c	10	4,000	0.50	JC69	20
Sequence divergence . .	GTR	a, b	d	10	2,000	0.10	GTR	20
	GTR	a, b	d	10	2,000	0.25	GTR	20
	GTR	a, b	d	10	2,000	0.50	GTR	20
	GTR	a, b	d	10	2,000	0.75	GTR	20
	GTR	a, b	d	10	2,000	0.10	JC69	20
	GTR	a, b	d	10	2,000	0.25	JC69	20
	GTR	a, b	d	10	2,000	0.50	JC69	20
	GTR	a, b	d	10	2,000	0.75	JC69	20
Model sensitivity	K2P	$\kappa = 1.0$	d	10	1,000	0.50	K2P	20
	K2P	$\kappa = 3.0$	d	10	1,000	0.50	K2P	20
	K2P	$\kappa = 6.0$	d	10	1,000	0.50	K2P	20
	K2P	$\kappa = 12.0$	d	10	1,000	0.50	K2P	20
	K2P	$\kappa = 1.0$	d	10	5,000	0.50	K2P	20
	K2P	$\kappa = 3.0$	d	10	5,000	0.50	K2P	20
	K2P	$\kappa = 6.0$	d	10	5,000	0.50	K2P	20
	K2P	$\kappa = 12.0$	d	10	5,000	0.50	K2P	20
	K2P	$\kappa = 1.0$	d	10	1,000	0.50	JC69	20
	K2P	$\kappa = 3.0$	d	10	1,000	0.50	JC69	20
	K2P	$\kappa = 6.0$	d	10	1,000	0.50	JC69	20
	K2P	$\kappa = 12.0$	d	10	1,000	0.50	JC69	20
	K2P	$\kappa = 1.0$	d	10	5,000	0.50	JC69	20
	K2P	$\kappa = 3.0$	d	10	5,000	0.50	JC69	20
	K2P	$\kappa = 6.0$	d	10	5,000	0.50	JC69	20
	K2P	$\kappa = 12.0$	d	10	5,000	0.50	JC69	20

[a] Rate parameters of the GTR model (a, b, c, d, e, f) are chosen for each replicate by drawing a uniform random number from the interval, U(0.0, 6.0].

[b] Base frequencies of the GTR model ($\pi_A\pi_C\pi_G\pi_T$) are drawn from a Dirichlet distribution using parameters drawn randomly from the interval, U[1,0, 4.0].

[c] For each replicate a birth-death tree was simulated as described in *Materials and Methods.*

[d] Topology was kept fixed for these simulations (fig. 1). Branch lengths were multiplied by the overall rate of substitution (m).

Analysis of the $\psi\eta$-Globin Pseudogene

To illustrate the method of model determination using posterior predictive distributions, a DNA sequence data set was analyzed under the JC69, HKY85, and GTR models. The data set is the primate $\psi\eta$-globin pseudogene (Koop et al. 1986; Goldman 1993) with the addition of one species—the pygmy chimpanzee. This data set consists of seven species—human beings (*Homo sapiens*), chimpanzee (*Pan troglodytes*), pygmy chimpanzee (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), rhesus monkey (*Macaca mulatta*), and owl monkey (*Aotus trivirgatus*). The original DNA data matrix was 2,205 sites. Indels (c = 183 sites) were excluded from the analyses, yielding a matrix of 2,022 sites.

Programs

MrBayes v2.0 was used to approximate the posterior distribution of a model's parameters and trees (Huelsenbeck and Ronquist 2001). The Metropolis-coupled MCMC algorithm was used with four chains (Huelsenbeck and Ronquist 2001). The Markov chains were run for 100,000 generations and sampled every 100th generation. The first 10,000 generations were discarded as burn-in to ensure sampling of the chain at stationarity. Convergence of the Markov chains was verified by plotting the log probability of the chain as a function of generation to verify that they had plateaued. A program that reads the posterior output of MrBayes, simulates predictive data sets, and evaluates the multinomial test statistic was written in the C language. The code is available upon request.

Results and Discussion

Overall Model Adequacy

The overall adequacy of an evolutionary model was explored by simulating nucleotide data sets of a variety of sequence lengths, on a birth-death tree, under the GTR model (see table 1). The birth-death process was used to explore the effects of different branch lengths and branching order. A comparison of the JC69 model with data sets simulated under the GTR, because of the

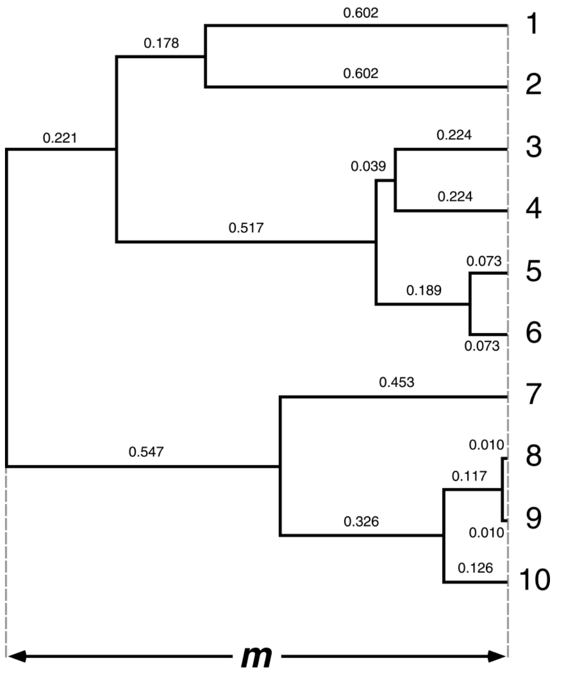

FIG. 1.—Tree used in simulations, testing sequence divergence, and sensitivity to model violations. The branch lengths were multiplied by the overall rate of substitution, m. Values for m can be found in table 1.

large difference in the number of parameters (eight), represents a conservative estimate of power.

An illustration of the predictive distribution of the multinomial test statistic for three data sets of 500, 2,000, and 4,000 sites is presented in figure 2. As expected the true (GTR) model centers the simulated distributions around the realized test statistic (fig. 2*A*, *C*, and *E*). The false (JC69) model performed poorly (P_T = 0.000; fig. 2*B*, *D*, and *F*). Increasing the number of sites in the data set increased the power of the test to reject the JC69 model: the predictive distributions under the JC69 model moved farther from the realized test statistic. This is because of the higher number of unique site patterns: increasing the number of sites increases the probability of observing rare patterns.

The effect of an increasing number of sites was measured in two ways: (1) using the mean posterior predictive *P* value ($\bar{P}_T$; table 2), and (2) using the power of the test (table 3). The first measure, mean *P* value, decreases below the critical value as the number of sites increases. For the true (GTR) model, the mean *P* value was close to the expected value of 0.5 for data sets of all sizes. For the false (JC69) model, the mean *P* value decreased as expected, as the number of sites increased, dropping below 0.05.

The second measure, the power of the test, increases as the number of sites increases. The true (GTR) model was accepted 100% of the time for data sets of all sizes. Interestingly, the false (JC69) model was often accepted for small data sets (table 3). The low power (or high Type II error rate) of the multinomial test statistic to reject a model, with small amounts of data, could be attributed to a number of causes. First, the small number simulations performed result in fairly large 95% confidence intervals (CI) for the Type II error rate (24%–68%). Second, the test statistic might not represent a complete summary of the underlying process of sequence evolution. Third, the approximation of the joint posterior distribution from small amounts of data may result in a large amount of uncertainty and increased Type II error rates.

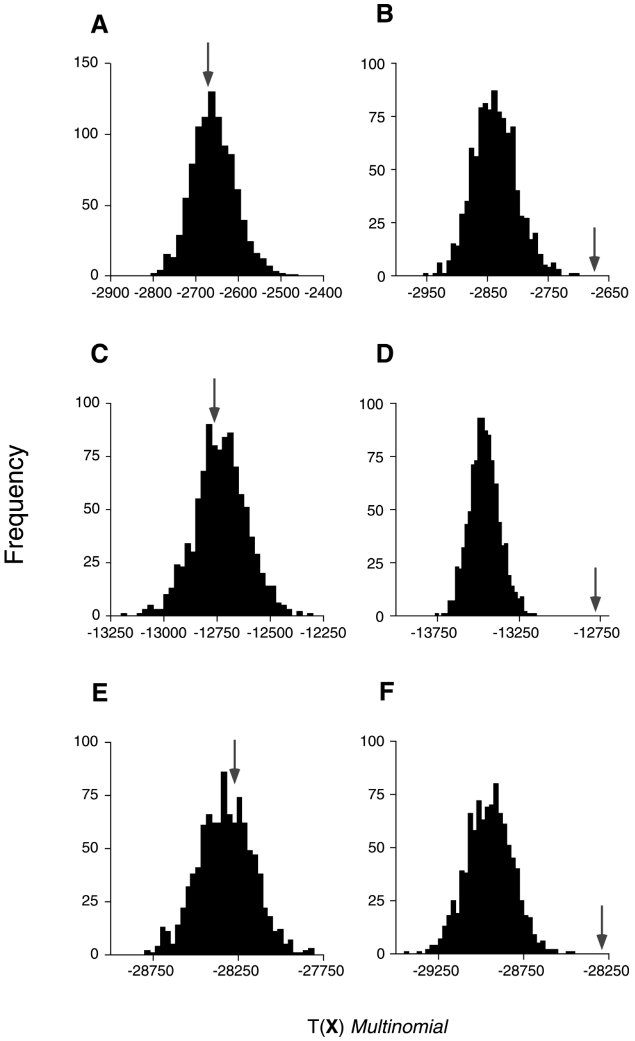

FIG. 2.—Illustration of the method comparing GTR versus JC69. Data sets of c = 500 (*A*, *B*), c = 2,000 (*C*, *D*), and c = 4,000 (*E*, *F*) sites were simulated under the GTR model. Predictive distributions were simulated under the GTR (*A*, *C*, *E*) and JC69 (*B*, *D*, *F*) models. Arrows indicate the values for the realized statistic from the original data. In all cases the GTR model, as expected, produced an adequate fit to the data, whereas the JC69 did not (P_T = 0.000).

Sequence Divergence

The effect of an increase in sequence divergence on power was explored by varying the overall rate of substitution across the tree shown in figure 1 (m = 0.10, 0.25, 0.50, and 0.75). Test data sets were simulated under the GTR model (table 1). The results of sequence divergence are shown in tables 2 and 3. As previously done, two measurements to evaluate the method are pre-

Table 2
Mean Posterior Predictive *P* Values for Simulations

Test	True Model	Number Characters	Substitution Rate (m)	Kappa (κ)	Model Tested	$\tilde{P}_T$[a]	Standard Deviation
Overall adequacy	GTR	500	0.50	NA	GTR	0.457	0.131
	GTR	2,000	0.50	NA	GTR	0.490	0.177
	GTR	4,000	0.50	NA	GTR	0.439	0.123
	GTR	500	0.50	NA	JC69	0.087	0.107
	GTR	2,000	0.50	NA	JC69	**0.007**	0.013
	GTR	4,000	0.50	NA	JC69	**0.000**	0.000
Sequence divergence	GTR	2,000	0.10	NA	GTR	0.411	0.082
	GTR	2,000	0.25	NA	GTR	0.466	0.095
	GTR	2,000	0.50	NA	GTR	0.457	0.142
	GTR	2,000	0.75	NA	GTR	0.486	0.137
	GTR	2,000	0.10	NA	JC69	0.094	0.090
	GTR	2,000	0.25	NA	JC69	**0.030**	0.041
	GTR	2,000	0.50	NA	JC69	**0.049**	0.091
	GTR	2,000	0.75	NA	JC69	**0.006**	0.013
Model sensitivity	K2P	1,000	0.50	1.0	K2P	0.433	0.185
	K2P	1,000	0.50	3.0	K2P	0.416	0.151
	K2P	1,000	0.50	6.0	K2P	0.426	0.138
	K2P	1,000	0.50	12.0	K2P	0.414	0.150
	K2P	5,000	0.50	1.0	K2P	0.470	0.120
	K2P	5,000	0.50	3.0	K2P	0.473	0.160
	K2P	5,000	0.50	6.0	K2P	0.472	0.077
	K2P	5,000	0.50	12.0	K2P	0.495	0.081
	K2P	1,000	0.50	1.0	JC69	0.426	0.192
	K2P	1,000	0.50	3.0	JC69	0.175	0.088
	K2P	1,000	0.50	6.0	JC69	**0.018**	0.018
	K2P	1,000	0.50	12.0	JC69	**0.000**	0.000
	K2P	5,000	0.50	1.0	JC69	0.471	0.128
	K2P	5,000	0.50	3.0	JC69	**0.019**	0.027
	K2P	5,000	0.50	6.0	JC69	**0.000**	0.000
	K2P	5,000	0.50	12.0	JC69	**0.000**	0.000

[a] Values in bold are significant at the α = 0.05 level.

sented: the mean predictive *P* value ($\bar{P}_T$) and the power of the test. Using the first, the GTR model performed well, approaching a mean predictive *P* value of 0.5 as *m* increased. An increase in the standard deviation of the predictive *P* value, from 0.082 to 0.137, was observed with an increase in divergence. This may be the result of a decrease in the diversity of site patterns as sites experience multiple hits and states begin to converge. The JC69 model, on the other hand, performed poorly at all values of *m*. At divergence levels of $m \geq 0.25$, the mean posterior predictive *P* value ($\bar{P}_T$; table 2) was below the critical level of α = 0.05.

Using the second measurement, the GTR model again performed well—it was accepted 100% of the time at all levels of divergence. The JC69 model performed poorly at low levels of sequence divergence (*m*

Table 3
Power of Test Statistic for Simulations

Test	True Model	Number Characters	Substitution Rate (m)	Kappa (κ)	Model Tested	Power (1 − β) (%)
Overall adequacy	GTR	500	0.50	NA	JC69	55
	GTR	2,000	0.50	NA	JC69	95
	GTR	4,000	0.50	NA	JC69	100
Sequence divergence. . . .	GTR	2,000	0.10	NA	JC69	50
	GTR	2,000	0.25	NA	JC69	80
	GTR	2,000	0.50	NA	JC69	70
	GTR	2,000	0.75	NA	JC69	100
Model sensitivity.	K2P	1,000	0.50	1.0	JC69	0
	K2P	1,000	0.50	3.0	JC69	5
	K2P	1,000	0.50	6.0	JC69	95
	K2P	1,000	0.50	12.0	JC69	100
	K2P	5,000	0.50	1.0	JC69	0
	K2P	5,000	0.50	3.0	JC69	90
	K2P	5,000	0.50	6.0	JC69	100
	K2P	5,000	0.50	12.0	JC69	100

= 0.10); the power of the test was relatively low (50) but rapidly increased to 100 at larger divergences (m = 0.75). For $m = 0.50$, the power of the test was considerably lower than in data sets with identical simulation conditions in the test of overall model adequacy (95%; see *Overall Model Adequacy*).

This reduction in power may be the result of a number of factors. The first, and most likely, explanation is sampling error; the small number of replicates leads to large confidence intervals (CI) around the Type II error rate (95% CI, 14%–53%). Second, the JC69 model may be robust to minor violations of its assumptions. For example, in replicates for which the JC69 model was accepted, assumptions were not severely violated. Analyses of model sensitivity support this explanation (see later). Third, the simulation tree for these analyses had a smaller sum of branch lengths than in the analysis of overall adequacy—branch lengths are in terms of the expected number of substitutions per site. When m = 0.5, figure 1 has a tree length of 2.266, whereas the mean tree length in the adequacy analysis was 2.601 (15 of 20 overall adequacy replicates had longer tree lengths, some as much as 36% longer). Therefore, the effects of divergence on power should be interpreted as a function of the total number of expected substitutions per site across the phylogeny—not simply the rate from the root to the tips of the tree (m). Finally, the statistic may be sensitive to the shape of the topology or variations in branch lengths across the tree.

Sensitivity to Model Violations

The sensitivity of the multinomial test statistic to reject inadequate models was explored by simulating data sets under the K2P model, varying κ from 1 to 12, followed by analysis with both the K2P (true) and JC69 models. When $\kappa = 1$, the K2P model collapses into the JC69 model. Under these conditions, the JC69 model is not violated and is expected to perform as well as the K2P model. As κ increases, reflecting an increase in the transition-transversion bias, the JC69 model becomes more severely violated and is expected to perform more poorly.

The effects of model violations were explored on data sets of two sizes: 1,000, and 5,000 sites (tables 2 and 3). Both the K2P and JC69 models performed well for data sets of 1,000 sites simulated with a κ value of 1. The mean posterior predictive P values for the K2P and JC69 models were 0.433 and 0.426, respectively. Both models were accepted in 100% of the replicates. For the JC69 model, as κ increased the mean P values declined, whereas the K2P model continued to perform well. The probability of accepting the K2P model was 100% for all replicates except one ($\kappa = 12$, 95%). The JC69 model performed well at $\kappa = 3$ (95% accepted), but as the model became increasingly violated, the power increased to 95% and 100%, at κ values of 6 and 12, respectively.

A fivefold increase in the number of sites moved the mean posterior predictive P values for the K2P model toward 0.5 (table 2), and all replicates analyzed under the K2P model were accepted 100% of the time. As the number of sites increased from 1,000 to 5,000, the discriminating power of the test statistic increased, as shown by the rapid decline in the mean predictive P values with increasing κ (table 2) and by the increased power to reject the JC69 model (table 3). For example, there was a nearly 10-fold drop in the mean predictive P value between 1,000 and 5,000 sites under the JC69 model with moderate violation—for $\kappa = 3$ the mean P value decreased from 0.175 to 0.019 (table 2). In addition, the variance across the replicate data sets decreased markedly. The JC69 model was accepted 100% of the time when κ was 1 but declined with an increase in κ compared with the 1,000 site data sets. This pattern is most dramatically demonstrated in a comparison of data sets simulated with $\kappa = 3$. For 1,000 sites there was a Type II error rate of 95% as compared with a Type II error rate of 10% with 5,000 sites under the JC69 model.

Analysis of the $\psi\eta$-Globin Pseudogene

The primate $\psi\eta$-globin pseudogene data set was analyzed under the GTR, HKY85, and JC69 models. Pseudogenes are nonfunctional copies in which mutations are not constrained by selection, and thus substitution biases should reflect mutational biases. Biases in the mutational spectrum will give rise to biases in the observed frequency of site patterns. The analysis of the mean base frequencies for the $\psi\eta$-globin pseudogene indicates an AT bias ($\pi_A = 0.296$, $\pi_C = 0.190$, $\pi_G = 0.238$, $\pi_T = 0.277$). Consequently, models that assume equal base frequencies (i.e., JC69) are not expected to perform as well as models that allow for unequal frequencies (i.e., HKY85 and GTR). The HKY85 and GTR models are adequate summaries of the true underlying process (GTR, $P_T = 0.199$; HKY85, $P_T = 0.303$), although the HKY85 represents a better fit to the data—the HKY85 model was better able to center the predictive distribution of the test statistic around the realized value (fig. 3). This difference may be because of a better model fit or stochastic error. The JC69 model represents a poor fit to the data (fig. 3, $P_T = 0.053$), even though it cannot be explicitly rejected at the 0.05 level.

Interestingly, Goldman (1993), using the parametric bootstrap, rejected the JC69 model for a similar data set that excluded the pygmy chimpanzee. The JC69 model performed less poorly with the method presented here. What can we attribute this apparent discrepancy to? One possible explanation is that with small numbers of taxa, and subsequently a smaller number of possible site patterns, there is low power—assuming that the JC69 model is inadequate, which, of course, may not be the case. Removal of the pygmy chimpanzee sequence and reanalysis of the model results in an increase in the predictive P value (JC69, $P_T = 0.123$). For the six-species data set, the JC69 model performs better at predicting the data than in the seven-species data set. This is not surprising because with fewer taxa there are fewer possible site patterns and the JC69 model, even with minor violations, should perform well. Another explanation is that accommodating uncertainty in the topol-

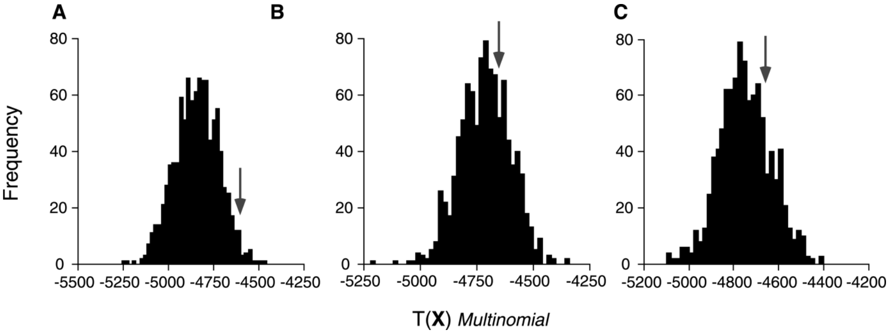

FIG. 3.—Analysis of the ψη-globin pseudogene under the JC69 (*A*), HKY85 (*B*) and GTR (*C*) models. The data set consisted of seven taxa and 2,022 nucleotides (see *Materials and Methods*). The MCMC analysis, performed using MrBayes v2.0, was used to approximate the joint posterior distribution of model parameters and topologies. The chain was run for 100,000 generations, sampling every 100th generation. The first 10,000 generations were discarded as burn-in. A total of 1,000 samples was randomly drawn from the joint posterior distribution of model parameters, and topologies and simulated data sets of 2,022 sites were generated under each of the models. The arrows above the distributions are the realized test statistic for the original data set (−4,651.32). Posterior predictive *P* values for JC69 (*A*), HKY85 (*B*), and GTR (*C*) are $P_T = 0.053$, $P_T = 0.303$, and $P_T = 0.199$, respectively.

ogy, using the present method, more accurately describes model variance. The JC69 model performed less poorly at predicting the observations than the parametric bootstrap when uncertainty was accommodated. The parametric bootstrap, by not accounting for uncertainty, may be more liberal in rejecting models. Analysis of the posterior distribution of trees for this data set suggests a high degree of uncertainty in the relationships between humans, gorillas, and chimpanzees—there was equal posterior support for the three possible subtrees of these species. This uncertainty was recognized in Goldman's (1993) analysis as a polytomy. Therefore, accounting for uncertainty in the topology, branch lengths, and model parameters appears to be important in determining model adequacy.

When we are confronted with two models that appear to perform equally well, how do we proceed in choosing between them? One approach would be to simply choose the less complex model, thus favoring a reduction in the number of free parameters to be estimated. Another alternative would be to use the method presented here, with a test statistic that summarizes local features of the models. In this way, identification of particular features of a model that do not contribute explanatory power can be identified and eliminated. Conversely, testing the addition of new parameters to a simpler model could lead to a better fit to the data using an expanded model. In these ways, we can identify the best model and arrive at a sound statistical choice.

Conclusions

The method I present here permits explicit evaluation of a phylogenetic model's adequacy using posterior predictive simulations. An adequate model should perform well in predicting future observations of the data; in the absence of such observations, simulations from the posterior distribution are used as surrogate observations. This approach differs, most importantly, from the traditional likelihood-based approaches by taking into account uncertainty in topology, branch lengths, and model parameters. Therefore, model choice has been freed from conditioning on these parameters and results in a more accurate estimate of model variance.

The multinomial test statistic is presented to evaluate the global (or overall) performance of a model through the posterior predictive distribution. The power of the multinomial test statistic was explored under a wide range of conditions. A number of factors have been shown here to increase power (1) increasing the number of sites, (2) increasing sequence divergence (expected number of substitutions per site), and (3) the degree of violation to a model's assumptions.

An appealing aspect of posterior predictive distributions, when used for model checking, is that a wide variety of test statistics can be formulated to check various aspects of phylogenetic models. For example, posterior predictive distributions can be used to detect variation in rates across data partitions, allowing models to be expanded to accommodate rate heterogeneity. The generality of the posterior predictive approach, and the development of new test statistics, will permit further exploration and development of more complex and realistic phylogenetic models.

Acknowledgments

This work was supported by grants from the NSF to John Huelsenbeck (MCB-0075404 and DEB-0075406), to whom I am grateful for his support throughout this project. Also, I would like to express my deep thanks to Andrea Betancourt, John Huelsenbeck, Kelly Dyer, Rasmus Nielsen, and Frederick Ronquist for taking the time to read early versions of the manuscript. Each and every one of them provided invaluable comments, that ultimately made the manuscript better. John Huelsenbeck, Bret Larget, Rasmus Nielsen, Ken Karol, and Andrea Betancourt patiently listened to me drone on about this project, and offered insightful comments that benefited this work, and for this they have my deepest gratitude. And finally, I would like to thank two anonymous reviewers who gave critical attention to the manuscript and provided valuable comments.

LITERATURE CITED

AKAIKE, H. 1974. A new look at 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.

CARLIN, B. P., and CHIB, S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. B **57**: 473–484.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

GAMERMAN, D. 1997. Markov Chain Monte Carlo: stochastic simulation for Bayesian Inference. Chapman and Hall, New York.

GAUT, B., and P. LEWIS. 1995. Success of maximum likelihood in the four taxon case. Mol. Biol. Evol. **12**:152–162.

GELFAND, A. E., and X.-L. MENG. 1996. Model checking and model improvement. Pp. 189–198 *in* W. R. GILKS, S. RICHARDSON, and D. J. SPIEGELHALTER, eds. Markov chain Monte Carlo in practice. Chapman and Hall, New York.

GELMAN, A., J. B. CARLIN, H. S. STERN, and D. B. RUBIN. 1995. Bayesian data analysis. Chapman and Hall, New York.

GELMAN, A. E., D. K. DEY, and H. CHANG. 1992. Model determination using predictive distributions with implementation via sampling-based methods. Pp. 147–167 *in* J. M. BERNARDO, J. O. BERGER, A. P. DAWID, and A. F. M. SMITH, eds. Bayesian statistics 4. Oxford University Press, New York.

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

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

HUELSENBECK, J. P., and J. P. BOLLBACK. 2001. Application of the likelihood function in phylogenetic analysis. Chap. 15, pp. 415–439 *in* D. J. BALDING, M. BISHOP, and C. CANNINGS, eds. Handbook of statistical genetics. John Wiley and Sons Inc., New York.

HUELSENBECK, J. P., J. P. BOLLBACK, and A. LEVINE. 2002. Inferring the root of a phylogenetic tree. Syst. Biol. **51**:32–43.

HUELSENBECK, J. P., and D. M. HILLIS. 1993. Success of phylogenetic methods in the four taxon case. Syst. Biol. **42**: 247–264.

HUELSENBECK, J. P., and F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformat. Appl. Note **17**:754–755.

HUELSENBECK, J. P., F. RONQUIST, R. NIELSEN, and J. P. BOLLBACK. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310–2314.

JUKES, T., and C. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

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.

KOOP, B. F., M. GOODMAN, P. XU, K. CHAN, and J. L. SLIGHTOM, 1986. Primate eta-globin DNA sequences and man's place among the great apes. Nature **319**:234–238.

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

LI, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Doctoral dissertation, Ohio State University, Columbus.

MAU, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Doctoral dissertation, University of Wisconsin, Madison.

MAU, B., and M. NEWTON. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. **6**:122–131.

MAU, B., M. NEWTON, and B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

NEWTON, M., B. MAU, and B. LARGET. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. *In* F. SEILLER-MOSEIWITCH, T. P. SPEED, and M. WATERMAN, eds. Statistics in molecular biology. Monograph series of the Institute of Mathematical Studies.

NIELSEN, R. 2002. Mapping mutations on phylogenies. Syst. Biol. (in press).

NIELSEN, R., and J. P. HUELSENBECK. 2001. Detecting positively selected amino acids sites using posterior predictive *p*-values. Pp. 576–588 *in* R. B. ALTMAN, A. K. DUNKER, L. HUNTER, K. LAUDERDALE, and T. E. KLEIN, eds. Pacific symposium on biocomputing. World Scientific, New Jersey.

POSADA D., and K. A. CRANDALL. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. **50**:580–601.

RANNALA, B., and Z. YANG. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. **43**:304–311.

RUBIN, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. **12**:1151–1172.

SCHWARZ, G. 1974. Estimating the dimension of a model. Ann. Stat. **6**:461–464.

SUCHARD, M. A., R. E. WEISS, and J. S. SINSHEIMER. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. **18**:101–1013.

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., G. OLSEN, P. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–511 *in* D. HILLIS, C. MORITZ, and B. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pp. 57–86 *in* Lectures in mathematics in the life sciences. Vol. 17[Please provide the publisher name and location].

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

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

WILKS, S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**: 554–560.

BRANDON GAUT, reviewing editor

Accepted March 25, 2002