

# Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo

R.H. Substitution model selection

Key words: Bayesian phylogenetic inference, Markov chain Monte Carlo, maximum likelihood, reversible jump Markov chain Monte Carlo, substitution models

JOHN P. HUELSENBECK<sup>1</sup>, BRET LARGET<sup>2</sup>, AND MICHAEL E. ALFARO<sup>1</sup>

<sup>1</sup>*Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0116, U.S.A.*

<sup>2</sup>*Departments of Statistics and of Botany, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706, U.S.A.*

John P. Huelsenbeck  
Section of Ecology, Behavior and Evolution  
Division of Biological Sciences  
University of California, San Diego  
La Jolla, CA 92093-0116  
U.S.A.  
Phone: (858) 822-1832  
FAX: (858) 534-7108  
E-mail: johnh@biomail.ucsd.edu

## Abstract

A common problem in molecular phylogenetics is choosing a model of DNA substitution that does a good job of explaining the DNA sequence alignment without introducing superfluous parameters. A number of methods have been used to choose among a small set of candidate substitution models, such as the likelihood ratio test, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Bayes factors. Current implementations of any of these criteria suffer from the limitation that only a small set of models are examined, or that the test does not allow easy comparison of non-nested models. In this paper, we expand the pool of candidate substitution models to include all possible time-reversible models. This set includes seven models that have already been described. We show how Bayes factors can be calculated for these models using reversible jump Markov chain Monte Carlo, and apply the method to 16 DNA sequence alignments. For each data set, we compare the model with the best Bayes factor to the best models chosen using AIC and BIC. We find that the best model under any of these criteria is not necessarily the most complicated one; models with an intermediate number of substitution types typically do best. Moreover, almost all of the models that are chosen as best do not constrain a transition rate to be the same as a transversion rate, suggesting that it is the transition/transversion rate bias that plays the largest role in determining which models are selected. Importantly, the reversible jump Markov chain Monte Carlo algorithm described here allows estimation of phylogeny (and other phylogenetic model parameters) to be performed while accounting for uncertainty in the model of DNA substitution.

## Introduction

Currently, a universal assumption of model-based methods of phylogenetic inference is that character change occurs according to a continuous-time Markov chain. At the heart of any continuous-time Markov chain is a matrix of rates, specifying the rate of change from one character state to another. For many phylogenetic analyses using DNA sequence data, it is assumed that there are four states (the nucleotides A, C, G, T/U) with a  $4 \times 4$  matrix of rates among the 12 possible nucleotide substitutions. A few standard models of DNA substitution have been proposed. These include those first described by Jukes and Cantor (1969), Kimura (1980, 1981), Felsenstein (1981, 1984), Hasegawa et al. (1984, 1985), Tamura and Nei (1993), and Tavaré (1986). Importantly, the parameterizations of the substitution process suggested by these authors can be directly applied to doublet models (used to describe the substitution process in stem-paired nucleotides; Schöniger and von Haeseler, 1994) and codon models (used to describe substitutions in the three nucleotides of a codon; Goldman and Yang, 1994, Muse and Gaut, 1994). Moreover, these models are used in a wide variety of analysis methods, including distance, maximum likelihood, and Bayesian methods.

One of the challenges facing the biologist is to choose a model of DNA substitution that best describes the data in hand. Generally speaking, the aim is to pick a model that adequately explains the data (in this case an alignment of DNA sequences) without introducing superfluous parameters; one wants to strike a balance between bias avoidance (achieved by using models with many parameters) and reduced variance in the parameter estimates (achieved by using models with few parameters). Fortunately, there are several methods available that allow the biologist to choose among substitution models in a statistical framework. When two nested models are compared, one can use likelihood ratio tests (Goldman, 1993). Models are nested when one model is a special case of a more general model. For example, the model proposed by Jukes and Cantor (1969) is a special case of the model proposed by Kimura (1980) which introduces an additional parameter describing the transition/transversion rate bias. When the transition/transversion bias is set to one, then Kimura's model is equivalent to the Jukes-Cantor model. In such nested model comparisons, minus twice the log likelihood ratio is asymptotically distributed as a  $\chi^2$  distribution, where the degrees of freedom of the  $\chi^2$  is the difference in the number of free parameters between the general and restricted models. Phylogeneticists commonly employ a series of nested likelihood ratio tests (LRTs) to select a substitution model, either by calculating likelihood ratios by hand for a pool of candidate models or by using Modeltest (Posada and Crandall, 1998), which automates this procedure. This is a reasonable approach because many of the most familiar substitution models can be arranged linearly in a nested hierarchy.

Unfortunately, not all of the interesting substitution model comparisons are between nested models. In these cases, it may make sense to use a method that does not assume nesting of models, such as Akaike's Information Criterion (AIC; Akaike, 1973), which is based on information theory (specifically, the Kullback-Liebler distance between models) and introduces a penalty term for models that have more parameters. The idea is to calculate the AIC for every candidate model, and choose the model with the smallest AIC for data analysis. (One might also consider other models that achieve a low AIC score.) Methods related to the AIC include the small sample bias adjusted AIC proposed by Hurvich and Tsai (1989), a quasi-likelihood modification (QAIC) suggested by Lebreton et al. (1992), and the Bayesian information criterion (BIC; Schwarz, 1978). AIC has been commonly used for choosing among a limited number of candidate substitution models, and is implemented in the program Modeltest (Posada and Crandall, 1998).

The criteria described above all work well when maximum likelihood is the criterion. In a Bayesian analysis, model choice is often guided by the Bayes factor (Kass and Raftery, 1995; Raftery, 1996), which is the ratio of the marginal likelihoods of two models; instead of maximizing the likelihood with respect to the parameters, the parameters are integrated over a prior probability distribution. Relatively little work has been done on Bayesian model selection in phylogenetics, with the notable exception of that by Suchard et al. (2001). Suchard et al. (2001) showed how Bayes factors can be calculated for nested models using the Savage-Dickey ratio (Verdinelli and Wasserman, 1995).

In this paper, we extend the pool of candidate substitution models to all possible time-reversible models. We develop a new Markov chain Monte Carlo algorithm that jumps between models, allowing accurate

calculation of the Bayes factor for any of the models. We also evaluate the possible DNA substitution models using AIC and BIC. We show that although the named models often do a good job of explaining much of the pattern of substitution for real data sets that (1) many other (unnamed) models also do quite well, (2) the most complicated time-reversible model is not universally chosen as best, and (3) phylogeny estimation can be performed while averaging over the possible models.

## Methods

How many nucleotide substitution models are there?

In this paper, we restrict our attention to time-reversible models of DNA substitution. Although we focus entirely on substitution models with four states (A, C, G, and T/U), all of the results of this study apply equally well to doublet and codon models. The most general time-reversible model, called the GTR model, has rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & r_{AC} \pi_C & r_{AG} \pi_G & r_{AT} \pi_T \\ r_{AC} \pi_A & - & r_{CG} \pi_G & r_{CT} \pi_T \\ r_{AG} \pi_A & r_{CG} \pi_C & - & r_{GT} \pi_T \\ r_{AT} \pi_A & r_{CT} \pi_C & r_{GT} \pi_G & - \end{pmatrix} \mu$$

and was first described by Tavaré (1986). The entry in the  $i$ th row and  $j$ th column of the matrix specifies the rate of change from nucleotide  $i$  to nucleotide  $j$ . The diagonal entries of the matrix, here shown with a dash, are specified such that each row sums to zero. The frequency of the  $i$ th nucleotide is denoted  $\pi_i$ . The commonly used DNA substitution models all have the unusual feature that the stationary frequency of the process is built directly into the rate matrix. The sum of the four base frequencies must, of course, equal one. We introduce the additional constraint that the sum of the six rate parameters is equal to six ( $r_{AC} + r_{AG} + r_{AT} + r_{CG} + r_{CT} + r_{GT} = 6$ ) so that their mean is one. We introduce this constraint for the same reason that most parameterizations of the GTR model fix the rate of  $G \leftrightarrow T$  to one, measuring the rates of the other five substitutions relative to the  $G \leftrightarrow T$  rate; because the absolute times on a phylogenetic tree are typically unknown, it is convenient to measure the branch lengths of the tree in terms of amount of nucleotide substitution (specifically, the branch lengths on a tree are in terms of expected number of substitutions per site). This means that the substitution rate matrix must be rescaled such that the mean rate is one. This is achieved by setting  $\mu = -1 / \sum_{i \in (A,C,G,T)} \pi_i q_{ii}$ . It also means that we cannot estimate the absolute rates of substitution, but only their relative rates. It is just as easy to measure the relative rates of the six substitution types on a scale where they all sum to six as it is to measure the substitution rates relative to the  $G \leftrightarrow T$  rate.

The models typically used in phylogenetic analysis are all special cases of the GTR model, involving constraints on the nucleotide frequencies or on the rate parameters. For example, the Felsenstein (1981) model constrains the rates of substitution to be equal ( $r_{AC} = r_{AG} = r_{AT} = r_{CG} = r_{CT} = r_{GT}$ ) and the Jukes-Cantor (1969) model has the additional constraint that the nucleotide frequencies are equal. For the remainder of this paper, we will assume that the base frequencies are a common parameter across all possible models, and that they are free to vary. We concentrate our attention on constraints of the substitution rate parameters. We also introduce a notation that allows any possible time-reversible model to be described. We assign index values to each of the six substitution rates, in the order  $AC, AG, AT, CG, CT, GT$ . If a model has the constraint that  $r_i = r_j$ , then the index value for those two rates is the same. Moreover, the index number for the first rate is always 1, and indices are labelled sequentially. So, for example, “111111” denotes the Jukes-Cantor (1969) or Felsenstein (1981) model and “121121” denotes the Kimura (1980), Hasegawa et al. (1984, 1985), or Felsenstein (1984) model. There are a total of 203 such models, with the simplest being “111111” and the most complex being the GTR model, “123456”. All of the possible models are shown in Table 1. Of the 203 models, one of them has one substitution type (111111), 31 of them have two substitution types (e.g., 121121), 90 of them have three substitution types (e.g., 123121), 65 of them have four substitution types (e.g., 112341), 15 of them have five substitution types (e.g., 123415), and one of them

has six substitution types (123456). Only a handful of the 203 models have been named<sup>1</sup>. These include 111111 (Jukes and Cantor, 1969; Felsenstein, 1981), 121121 (Kimura, 1980; Felsenstein, 1984; Hasegawa et al., 1984, 1985), 121131 (Tamura and Nei, 1993), 123321 (Kimura, 1981), 123341 (Posada, 2003), 123425 (Posada, 2003), and 123456 (Tavaré, 1986).

How is the number of possible substitution models determined? The answer can be found in the combinatorics literature. Specifically, the number of ways a set with  $n$  objects can be partitioned into disjoint and non-empty subsets is described by the Bell numbers (Bell, 1934). The Bell numbers,  $B_n$ , are calculated as the sum

$$B_n = \sum_{k=0}^n S(n, k)$$

where  $S(n, k)$  is a Stirling number of the second kind. The Stirling numbers of the second kind,  $S(n, k)$ , give the number of ways to partition a set of  $n$  objects into  $k$  non-empty sets. The Stirling number of the second kind is

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n$$

Special cases of the Stirling number of the second kind include  $S(n, 0) = 0$ ,  $S(n, 1) = 1$ ,  $S(n, n) = 1$ ,  $S(n, 2) = 2^{n-1} - 1$ , and  $S(n, n-1) = \binom{n}{2}$ . The Bell numbers give the sequence 1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, 4213597, ... for  $n = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \dots$ . For the problem investigated here, the Bell number for  $n = 6$  objects gives the number of possible substitution models ( $B_6 = 203$ ) and the Stirling numbers of the second kind give the number of models with one [ $S(6, 1) = 1$ ], two [ $S(6, 2) = 31$ ], three [ $S(6, 3) = 90$ ], four [ $S(6, 4) = 65$ ], five [ $S(6, 5) = 15$ ], and six [ $S(6, 6) = 1$ ] substitution types.

To improve understanding of the remainder of this paper, we introduce the following additional notation describing the substitution models. Each substitution model has  $K$  groups of substitution types. The membership of the  $i$ th group is contained in the vector  $\mathbf{k}_i$ . For example, model  $M_{101}$  (112231) has a total of  $K = 3$  substitution groups:  $\mathbf{k}_1 = (AC, AG, GT)$ ,  $\mathbf{k}_2 = (AT, CG)$ , and  $\mathbf{k}_3 = (CT)$ . The number of substitution types in the  $i$ th substitution group is denoted  $n_i$  and the rate for the  $i$ th substitution group is denoted  $r_i$ . For model  $M_{101}$ ,  $n_1 = 3$ ,  $n_2 = 2$  and  $n_3 = 1$ . Finally,  $N(M)$  is the number of substitution groups in model  $M$  with  $n_i > 1$ . For model number 101,  $N(M_{101}) = 2$ ; the GTR model (123456) has  $N(M_{203}) = 0$ .

### Choosing among the possible substitution models

A number of criteria can be used to choose among a set of  $d$  candidate models, where the candidate models are labelled  $M_1, M_2, \dots, M_d$ . Here we review a few of these methods. However, we mainly concentrate on the Bayesian framework for choosing models, showing how Bayes factors can be calculated for substitution models of interest. We will also point out how AIC and BIC can be used to choose among all possible substitution models when using maximum likelihood.

*Likelihood ratio testing*—The likelihood ratio

$$\Lambda = \frac{\max[\ell(M_1)]}{\max[\ell(M_2)]}$$

measures the relative tenability of two competing models. (The likelihood,  $\ell(M_i)$ , is maximized over all the parameters.) Model  $M_1$  is favored when the likelihood ratio is greater than one, whereas the opposite is true for likelihood ratios less than one. When the models are nested, with  $M_1$  being a special case of  $M_2$ ,  $\Lambda \leq 1$  and  $-2 \log_e \Lambda$  is asymptotically distributed as a  $\chi^2$  distribution with the degrees of freedom being the difference in the number of free parameters between  $M_2$  and  $M_1$ .

---

<sup>1</sup>In a talk at the 1995 Evolution meetings held in Montreal, Canada, D. L. Swofford pointed out the large number of possible substitution models and named one after his, as it turns out fictional, Aunt Emily. Unfortunately, everyone involved, including D. L. Swofford, has forgotten which of the 203 models was named after ‘Aunt Emily’, so this model is not included as one of the named ones in this paper.

Likelihood ratio tests have proven very useful in molecular phylogenetics (see review by Huelsenbeck and Rannala, 1997) where they have been used to examine questions ranging from the molecular clock hypothesis to cospeciation in hosts and parasites. However, to use the  $\chi^2$  approximation of the null distribution, the requirement that the models are nested restricts the hypotheses that can be compared. With non-nested models, one can use the Cox test (Cox, 1962), but in these cases it is not clear which hypothesis the null distribution should be generated under. In this paper, where many of the models of interest are not nested (e.g., 111222 vs. 112211), likelihood ratio tests are of only limited value.

*Information criteria*—Another approach to model choice is to attempt to find a model that in some sense minimizes the distance to the ‘true’ model. A number of methods attempt to estimate the Kullback-Liebler information, which is a measure of the distance between models. The Akaike Information Criterion (AIC; Akaike, 1973) is based on the maximum likelihood score and the number of parameters for a model, and represents an attempt to estimate the Kullback-Liebler information (distance) between the true model and the fitted model. The AIC for model  $i$  is calculated as

$$AIC_i = 2 \times \max[\log_e \ell(M_i)] - 2p$$

where  $\max[\log_e \ell(M_i)]$  is the maximum likelihood and  $p$  is the number of free parameters in the model. A Bayesian variant of the AIC is the Bayesian Information Criterion (BIC; Schwarz, 1978) which assumes equal priors on each model and vague priors on the parameters. The BIC for model  $i$  is calculated as

$$BIC_i = 2 \times \max[\log_e \ell(M_i)] - p \log_e(n)$$

where  $n$  is the number of observations. For both AIC and BIC, the first term can be interpreted as a penalty for using too simple a model; the maximum likelihood is lower for models that fail to include important parameters. The second term, on the other hand, is a penalty against overfitting. The second term in AIC and BIC favors more parsimonious models.

Model selection using AIC or BIC works by calculating the AIC/BIC scores for each candidate model and selecting the model with the lowest AIC/BIC as best. AIC and BIC have been used previously to choose among a handful of competing phylogenetic models (e.g., Posada and Crandall, 1998, Yang et al., 2000).

*Bayes factors*—The Bayes factor for a comparison of model  $M_1$  to model  $M_2$  is

$$BF_{12} = \frac{f(X|M_1)}{f(X|M_2)}$$

where the observations are denoted  $X$  (in this case, an alignment of DNA sequences) and  $f(X|M_i)$  is the marginal likelihood of model  $i$ . The Bayes factor has the same intuitive meaning as the likelihood ratio, described above; values for the Bayes factor greater than one support  $M_1$  while a Bayes factor less than one supports  $M_2$ . The marginal likelihoods are calculated by integrating (marginalizing) over the other parameters of the model. Say that model  $M_i$  has parameters  $\theta_i$ . The marginal likelihood is calculated by integrating the likelihood over a prior probability distribution,  $f(\theta_i|M_i)$ :  $f(X|M_i) = \int_{\theta_i} f(X|\theta_i, M_i)f(\theta_i|M_i)d\theta_i$ . Note that in contrast to the likelihood ratio test statistic in which the ratio of the maximum likelihood scores is used, for the Bayes factor the ratio of the average likelihoods is used instead. Importantly, the Bayes factor accounts for uncertainty in the parameters of the model,  $\theta_i$ . Also note that the Bayes factor does not rely on nesting of the models. In fact, the parameters of the two compared models need not be the same. That is, the parameters of model  $M_1$  ( $\theta_1$ ) are not necessarily the same parameters as those in model  $M_2$  ( $\theta_2$ ).

The Bayes factor can also be calculated as the ratio of the posterior odds to the prior odds of the models

$$BF_{12} = \frac{f(X|M_1)}{f(X|M_2)} = \frac{f(M_1|X)/f(M_2|X)}{f(M_1)/f(M_2)}$$

where  $f(M_i|X)$  is the posterior probability and  $f(M_i)$  is the prior probability of model  $i$ . Often the Bayes factor can be approximated using the output of a Markov chain Monte Carlo (MCMC) analysis. This is

the approach that will be taken in this paper, with the details of the MCMC algorithm described in the next section. Suchard et al. (2001) calculated the Bayes factor for the special case where the compared models are nested. Specifically, they used the Savage-Dickey ratio to approximate the Bayes factor, where the Savage-Dickey ratio is

$$BF_{21} = \frac{1}{BF_{12}} = \frac{f(\theta = \theta_0 | M_2, X)}{f(\theta = \theta_0 | M_2)}$$

Model  $M_1$  is a special case of  $M_2$ . Specifically,  $M_2$  is the same as  $M_1$  when the parameter  $\theta = \theta_0$ . For example, the Jukes-Cantor (1969) and Kimura (1980) models are the same when the parameter  $\kappa$  of the Kimura model is set to one ( $\kappa = 1$ ). The Bayes factor could be calculated using the Savage-Dickey ratio for a comparison of the Jukes-Cantor ( $M_1$ ) to the Kimura ( $M_2$ ) models as

$$BF_{12} = \frac{f(\kappa = 1 | M_2, X)}{f(\kappa = 1 | M_2)}$$

This involves performing a Bayesian analysis under the more general, Kimura (1980), model and evaluating the posterior and prior probability densities of the parameter  $\kappa$  at one.

#### Reversible jump MCMC for model selection

We do not pursue calculation of Bayes factors using the Savage-Dickey ratio for the same reason that we do not examine likelihood ratio tests; many of the comparisons of interest involve non-nested models. Instead, we calculate Bayes factors for all possible substitution models by constructing a Markov chain that visits substitution models in proportion to their posterior probability. The joint posterior probability of all the model parameters is calculated using Bayes' theorem as

$$f(\tau, \mathbf{v}, \alpha, \pi, \theta_i, M_i | X) = \frac{f(X | \tau, \mathbf{v}, \alpha, \pi, \theta_i, M_i) f(\tau, \mathbf{v}, \alpha, \pi, \theta_i, M_i)}{f(X)}$$

where  $\tau$  is the unrooted tree topology,  $\mathbf{v}$  is the vector of branch lengths for the tree,  $\alpha$  is the gamma shape parameter describing the amount of among-site rate variation,  $\pi$  is a vector of nucleotide frequencies,  $\theta_i$  is the vector of substitution rates, and  $M_i$  is the substitution model. The posterior probability,  $f(\cdot | X)$ , is equal to the likelihood,  $f(X | \cdot)$ , times the prior probability distribution,  $f(\cdot)$ , divided by a normalizing constant,  $f(X)$ . In this section, we describe the modelling assumptions we make in calculating posterior probabilities of models.

*Likelihood*—We calculate the likelihood (probability of observing the DNA sequence alignment conditional on a specific model) under the GTR+ $\Gamma$  model of DNA substitution, or submodels of the GTR+ $\Gamma$ . We approximate the continuous gamma distribution for among site rate variation using the discrete approximation of Yang (1994). We use four categories to approximate the gamma distribution. The pruning algorithm of Felsenstein (1981) was used to calculate the likelihood for a tree.

*Prior*—We assume that all 203 substitution models have equal prior probability. Moreover, we assume a uniform prior on unrooted tree topologies, an exponential(10) prior on branch lengths, a flat Dirichlet distribution on nucleotide frequencies, and a uniform(0, 50) prior on the gamma shape parameter for among-site rate variation. We assume that the  $K$  rates of the substitution model are a linear transformation of a flat Dirichlet prior. In particular, if  $Y_1, \dots, Y_K$  have a flat Dirichlet distribution we let  $r_i = 6Y_i/n_i$  for  $i = 1, \dots, K$ . The prior density of the rates is

$$\frac{\Gamma(K) \prod_{i=1}^K n_i}{6^K}, \quad 0 \leq r_i \leq \frac{6}{n_i}, \quad \sum_{i=1}^K n_i r_i = 6, \quad \text{for } i = 1, \dots, K.$$

(The standard flat Dirichlet distribution has a normalizing constant of  $\Gamma(K)$ . Our linear transformation of the flat Dirichlet requires a different normalizing constant.)

*Markov chain Monte Carlo*—The posterior probability distribution of the phylogenetic model parameters cannot be calculated analytically. We use Markov chain Monte Carlo (MCMC; Metropolis et al., 1953;

Hastings, 1970; Green, 1995) to approximate the posterior probability distribution of the phylogenetic model parameters. MCMC for phylogeny inference has been described elsewhere (see Larget and Simon, 1999; Huelsenbeck et al., 2002). The general idea is to construct a Markov chain that has as its state space the parameters of the phylogenetic model and a stationary distribution that is the posterior probability distribution of the parameters. The sequence of parameter states visited by the Markov chain form a valid, albeit dependent, sample from the posterior probability distribution of interest (Tierney, 1994). In this paper, we are interested in the posterior probabilities of alternative substitution models. We construct a Markov chain that, in addition to exploring the other model parameters such as the phylogenetic tree, visits alternative substitution models. The posterior probability of the  $i$ th model is approximated as the proportion of the time the Markov chain visited model  $i$ .

In each iteration of the MCMC algorithm, we pick a parameter at random to update. We update the tree and branch lengths simultaneously using the LOCAL mechanism described in Larget and Simon (1999). We update the nucleotide frequencies and substitution model parameters using a Dirichlet proposal mechanism and we update the gamma shape parameter using a sliding window mechanism. Each of these proposal mechanisms has been described elsewhere, and we will not go into any more detail about them here. However, two of the proposal mechanisms are unique to the problem of exploring substitution models. Specifically, we have proposal mechanisms that explore the different possible substitution models by merging or splitting groups of substitution types. The main complication is that moves from one substitution model to another involve changes in the dimension of the parameter space and the normal MCMC theory does not apply. Instead, we construct a Markov chain that jumps dimensions using the theory described in Green (1995).

The current substitution model will be denoted  $M$ . We attempt a merge of two substitution groups with probability  $p_M(M)$  and attempt a split of a substitution group into two groups with probability  $p_S(M)$ . When the current model is the Jukes-Cantor (111111), then the probability of attempting a merge is zero and the probability of attempting a split is set to one. When the current model is GTR (123456) the probability of attempting a split is set to zero, and the probability of a merge is set to one.

A merge move works as follows. Two of  $K$  substitution groups of the current model,  $M$ , are chosen at random. These groups are denoted  $\mathbf{k}_i$  and  $\mathbf{k}_j$ . The new group that results from the merging of these two groups is denoted  $\mathbf{k}'$ . For example, consider a merge of the model 111222. There are only two substitution groups that can be merged:  $\mathbf{k}_1 = (AC, AG, AT)$  and  $\mathbf{k}_2 = (CG, CT, GT)$ . The group that results from the merging of these two groups is  $\mathbf{k}' = (AC, AG, AT, CG, CT, GT)$ . Of course, this is the model 111111. The rates of substitution must also be chosen for the new, merged, group. For the merge move, the new rate is deterministically chosen. The rates of substitution for the two merged groups are denoted  $r_i$  and  $r_j$ . The new rate,  $r'$  is

$$r' = \frac{n_i r_i + n_j r_j}{n_i + n_j}$$

The new group size is  $n_i + n_j$ . Remember that we constrain the sum of the rates to be 6 ( $\sum_i n_i r_i = 6$ ), so that  $\bar{r} = 1$ . Notice that  $(n_i + n_j)r' = n_i r_i + n_j r_j$ : the sum of all rates remains unchanged.

A split move works by picking one of the  $N(M)$  groups with at least two substitution types at random. The group to be split will be designated  $\mathbf{k}$ . The substitution types in that group are then divided randomly, with the constraint that there is at least one item in each group. This results in two groups,  $\mathbf{k}'_i$  and  $\mathbf{k}'_j$ . For example, if the current model,  $M$ , is 112334, then only two groups are available that could be potentially split: the groups with substitutions  $AC$  and  $AG$  or with substitution types  $CG$  and  $CT$ . One of these two groups would be chosen at random (with each having an equal probability of being chosen), and the substitution types in the group randomly divided into two new groups. In this case, if  $\mathbf{k} = (AC, AG)$ , then the resulting groups would be  $\mathbf{k}'_i = (AC)$  and  $\mathbf{k}'_j = (AG)$ . To find the new rates for the two new groups,  $r'_i$  and  $r'_j$ , we first generate a uniform random variable on the interval  $(-n_i r, n_j r)$ , denoted  $u$ , where  $r$  is the rate of the group to be split. The new rates are then

$$r'_i = r + \frac{u}{n_i}$$

$$r'_j = r - \frac{u}{n_j}$$

Notice as well that  $n_i r'_i + n_j r'_j = (n_i + n_j)r$ , so that the sum of all the rates remains unchanged.

The probability of accepting a merge or split move,  $R$ , is

$$R = \min(1, \text{Likelihood Ratio} \times \text{Prior Ratio} \times \text{Proposal Ratio} \times \text{Jacobian})$$

For a merge move, the prior ratio is

$$\text{Prior Ratio} = \frac{(1/203) \times \Gamma(K-1) \times \left( (n_i + n_j) \prod_{a \neq i,j} n_a \right) / 6^{K-1}}{(1/203) \times \Gamma(K) \times \left( n_i n_j \prod_{a \neq i,j} n_a \right) / 6^K}$$

the proposal ratio is

$$\text{Proposal Ratio} = \frac{p_S(M') \times (1/N(M')) \times (1/(2^{n_i+n_j-1} - 1)) \times (1/(r(n_i + n_j)))}{p_M(M) \times 1/\binom{K}{2}}$$

and the Jacobian is

$$\text{Jacobian} = \frac{n_i n_j}{n_i + n_j}$$

The acceptance probability for a merge move is then

$$R = \min \left( 1, \text{Likelihood Ratio} \times \frac{p_S(M')}{p_M(M)} \times \frac{6K}{N(M')(2^{n_i+n_j} - 2)r(n_i + n_j)} \right)$$

For a split move from current model  $M$  with  $K-1$  rates to a model  $M'$  with  $K$  rates the acceptance probability is

$$R = \min \left( 1, \text{Likelihood Ratio} \times \frac{p_M(M')}{p_S(M)} \times \frac{N(M)(2^{n_i+n_j} - 2)r(n_i + n_j)}{6K} \right)$$

where  $r$  is the rate of the group to be split and  $n_i$  and  $n_j$  are the sizes of the new groups.

Data

We examined 16 alignments of DNA sequences. Table 2 summarizes the characteristics of the data sets examined. The posterior probabilities of substitution models (and other phylogenetic model parameters) were approximated using the MCMC algorithm described above as implemented in a program written by J.P.H. PAUP\* (Swofford, 2002) was used to calculate the maximum likelihood scores of the 203 models for each data set. All of the data sets were analyzed using the same model: the GTR+ $\Gamma$  model of DNA substitution, and submodels thereof. The MCMC algorithm was run for 10,000,000 cycles for each data set, and sampled every 1,000th cycle. Each data set was analyzed three times. For each chain, we excluded samples taken during the first million cycles as the “burn in”, basing inferences on the last 9,000 samples taken in each analysis. All probabilities were based upon the pooled post-burn in samples from the three chains.

## Results and Discussion

For most of the data sets analyzed in this study, the posterior probability is spread out among a handful of models. Table 3 shows the best models selected for each analysis. The posterior probability was concentrated on only a few substitution models for the mammal and Parrotfish data set, where two and three models, respectively, accounted for 95% of the posterior probability. For some of the smaller alignments, however, there was considerable model uncertainty. For example, for the HIV-*env* alignment 22 models accounted for 95% of the posterior probability. Another interesting finding is that the most complicated model (GTR:

123456) was not found to be best in any of the analyses; in fact, it was chosen with a high probability only in the mammalian mitochondrial data set, where it had a posterior probability of 0.36.

The Bayes factors for the best models found for each alignment ranged from 52 to 481. Jeffreys (1961) provided a table to help interpret Bayes factors, that was modified by Raftery (1996). Generally speaking, Bayes factors in the range 12 to 150 are considered “strong” evidence in favor of a model while Bayes factors greater than 150 are considered “very strong” evidence in favor of a model. Another way of interpreting the Bayes factor is as a measure of how one changes belief in a hypothesis in the light of new evidence. Lavine and Schervish (1999:120) argue that the Bayes factor should be interpreted as measuring “the *change* in the odds in favor of the hypothesis when going from the prior to the posterior.” For either interpretation, the large Bayes factors indicates that the alignments contained considerable information about the substitution model.

Generally speaking, substitution models with two to five substitution types did best. Table 4 shows the Bayes factors for the number of substitution types ( $K$ ) and Figure 1 shows the posterior probability of the number of substitution types. The Jukes-Cantor (1969) and Felsenstein (1981) model (111111) did poorly for all of the data sets examined in this study. The seven named models did well for 10 of the data sets, having a Bayes factor greater than one, meaning there was more posterior probability associated with the named models than there was prior probability. However, the Bayes factor for the named models was never greater than 115, and was less than one for four of the data sets.

The results for model choice using AIC and BIC are largely concordant with the results from Bayes factors. Table 3 also shows the best models chosen using the AIC and BIC. For four of the alignments, AIC, BIC, and posterior probabilities were in agreement, choosing precisely the same model as best. For many of the other cases, AIC or BIC chose models that were still in the credible set of models, calculated using the reversible jump MCMC algorithm. In only one instance did AIC/BIC not choose a model that was at least in the credible set of models; for the butterfly alignment of *wingless* sequences, BIC chose model 193, which was not in the credible set of models. (Model 125, chosen by AIC for the butterfly data set was in the 95% credible set of models.)

In this study, we ran the Markov chains assuming a uniform prior probability on all models, and Bayes factors were calculated as the ratio of the posterior odds to the prior odds in favor of the model. The posterior probabilities of the best models happened to be of intermediate value for all 16 data sets examined, so that it was straight-forward to calculate the Bayes factor for the best model. However, it is possible that the posterior probability for the best model would be approximated to be 1.0 under a uniform prior on models, or alternatively, the biologist might be interested in the Bayes factor for a model that does a very poor job of explaining the data, having a posterior probability that is approximated using MCMC to be 0.0. However, in both cases, the true posterior probability is neither 1.0 nor 0.0, but rather a number very close to 1.0 or 0.0. How does one accurately approximate the Bayes factors for very good or very poor substitution models? One solution involves seeding the prior, giving more weight to a poor model for which one wants to approximate a Bayes factor, or giving less prior weight to a good model. One then runs the MCMC algorithm using the non-uniform prior, calculating the posterior probabilities for the model of interest as the fraction of the time that the Markov chain visited that model. One can then calculate what the posterior probability of model  $i$  would have been if the analysis had been performed using a uniform prior. The posterior probability of model  $i$  is

$$f'(M_i|X) = \frac{f(M_i|X) \times \frac{f'(M_i)}{f(M_i)}}{\sum_j f(M_j|X) \times \frac{f'(M_j)}{f(M_j)}}$$

where  $f'(M|X)$  is the posterior probability and  $f'(M)$  is the prior probability one is interested in calculating and  $f(M|X)$  and  $f(M)$  are the posterior probability and prior probability, respectively, actually used when performing the MCMC analysis.

Application of likelihood ratio tests to a small subset of hierarchically nested models would often lead to different conclusions about the optimal model than were reached here, when considering all possible

time-reversible models. For example, consider the following three nested model comparisons: Test 1,  $M_1$  (111111) vs.  $M_{15}$  (121121); Test 2,  $M_{15}$  vs.  $M_{40}$  (121131); and Test 3,  $M_{40}$  vs.  $M_{203}$  (123456). Here, the likelihood ratio test statistic is asymptotically distributed as a  $\chi^2$  with one (tests 1 and 2) or three (test 3) degrees of freedom. At the 5% level, the null (simple) model is rejected all of the time for Test 1, half of the time for Test 2, and 13 of the 16 times for Test 3. Importantly, the most complicated model considered ( $M_{203}$ ) would often be chosen as the best model when looking at only a small subset of all possible models. A similar conclusion would be reached if one only considered the subset of models that have one, two, or six substitution types. Phylogeneticists often concentrate on only these models, probably because they are easily implemented using the program PAUP\* (Swofford, 2002) and are the only possible models in older versions of the program MrBayes (Huelsenbeck and Ronquist, 2001).

The models chosen as best in the 16 DNA sequence alignments examined here were all similar in one respect: virtually all of them allowed the transition rates to differ from the transversion rates. That is to say, one rarely saw a model chosen that constrained a transition rate to be the same as a transversion rate (e.g.,  $M_{60}$  where the substitution pattern is 123313). In fact, with only two exceptions, every model in the 95% credible sets of models allowed the transition rates to be different from the transversion rates. The exceptions included the HIV-*env* and the vertebrate  $\beta$ -globin alignments, where the best models included  $M_{25}$  (112212),  $M_{60}$  (123313),  $M_{64}$  (112232),  $M_{100}$  (112312), and  $M_{171}$  (112343), among others. Clearly, in almost all of the sequence alignments examined in this study, the predominant pattern is that transitions occur at a different rate than transversions, and this fact determines which models are chosen as best; the best models all appear to be minor variants of the model proposed by Kimura (1980) and Hasegawa et al. (1984, 1985).

The models currently used in molecular phylogenetics are quite complicated, and are likely to become more so as biologists add parameters to the models. For example, one strategy that is used to allow a phylogenetic model to better explain a DNA sequence alignment is to partition the data, perhaps according to codon or some other biologically relevant criterion, and allow the parameters of the model to be independently estimated for each partition (Ronquist and Huelsenbeck, 2003). This strategy, while it allows the phylogenetic model to better explain the data, also adds many parameters, and it is unclear how the biologist is to sort out which partitioning scheme is best; is it best to constrain a parameter, such as the gamma shape parameter that describes among site rate variation, to be the same for all of the data partitions or should one allow the parameter to be different? If it is to be potentially different, should the parameter be independently estimated for all partitions, or should this parameter be constrained to be the same for some but not all of the partitions? Questions such as these are likely to become more prevalent in the future as models become more complicated, and biologists become more concerned with using any knowledge about a gene to tailor the assumptions of the analysis to specific gene regions. The automated procedure described in this paper has the advantage that it allows the screening of a large number of possible models when performing a phylogenetic analysis. In principal, one could set up a very complicated general model, such as one that allows parameters to be independently estimated for each data partition, and then allow the Markov chain Monte Carlo procedure to explore the different possible sub-models in proportion to their probability. Formally, the inference of phylogeny would not depend upon any specific model, but would be integrated over uncertainty in the model parameters. Moreover, one would learn how the pattern of substitution differs among data partitions from the output of such an analysis.

## Acknowledgments

J.P.H. was supported by NSF grants MCB-0075404 and DEB-0075406. B.L. was supported by NIH grant R01 GM68950-01 and NSF grant DEB-0075406. M.E.F. was supported by ITR grant CACTTOL-22035A.

## LITERATURE CITED

- ADACHI, J., and M. HASEGAWA. 1995. Phylogeny of whales: Dependence of the inference on species sampling. *Mol. Biol. Evol.* **12**:177–179.
- AKAIKE, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.). Akademiai Kiado, Budapest.
- ALFARO, M. E., and S. J. ARNOLD. 2001. Molecular systematics and evolution of *Regina* and the thamnophiine snakes. *Mol. Phylo. Evol.* **21**:408–423.
- ARNASON, U., A. GULLBERG, and A. JANKE. 1997. Phylogenetic analyses of mitochondrial DNA suggest a sister group relationship between Xenartha (Edentata) and Ferungulates. *Mol. Biol. Evol.* **14**:762–768.
- BARNS, S. M., C. F. DELWICHE, J. D. PALMER, and N. R. PACE. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. USA* **93**:9188–9193.
- BELL, E. T. 1934. Exponential Numbers. *Amer. Math. Monthly* **41**:411–419.
- BROWER, A. V. Z. 2000. Phylogenetic relationships among the Nymphalidae (Lepidoptera) inferred from partial sequences of the wingless gene. *Proc. R. Soc. Lond. B* **267**:1201–1211.
- COX, D. R. 1962. Further results on tests of families of alternate hypotheses. *J. R. Stat. Soc. B* **24**:406–424.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: A justification. *Evolution* **38**:16–24.
- GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**:182–198.
- GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- GREEN, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**:711–732.
- HAFNER, M. S., P. D. SUDMAN, F. X. VILLABLANCA, T. A. SPRADLING, J. W. DEMASTES, and S. A. NADLER. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265**:1087–1090.
- HARSHMAN, J., C. J. HUDDLESTON, J. P. BOLLBACK, T. J. PARSONS, and M. J. BRAUN. 2003. True and False Gharials: A Nuclear Gene Phylogeny of Crocodylia. *Syst. Biol.* **52**:386–402.
- HASEGAWA, M., T. YANO, and H. KISHINO. 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. *Proc. Japan Acad. Ser. B* **60**:95–98.
- HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97–109.
- HAYASAKA, K., T. GOJOBORI, and S. HORAI. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.* **5**:626–644.

- HUELSENBECK, J. H., and B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276**:227–232.
- HUELSENBECK, J. P., and F. RONQUIST. 2001 MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755.
- HUELSENBECK, J. P., B. LARGET, R. E. MILLER, and F. RONQUIST. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* **51**:673–688.
- HURVICH, C. M., and C-L. TSAI. 1989. Regression and time series model selection in small samples. *Biometrika* **76**:297–307.
- JACKMAN, T. R., A. LARSON, K. DE QUEIROZ, and J. B. LOSOS. 1999. Phylogenetic relationships and tempo of early diversification in *Anolis* lizards. *Syst. Biol.* **48**:254–285.
- JEFFREYS, H. 1961. *Theory of Probability*. Oxford University Press, Oxford.
- JUKES, T. H., and C. R. CANTOR 1969. Evolution of protein molecules. Pages 21–123 in *Mammalian Protein Metabolism* (H. N. Munro, ed.). Academic Press, New York.
- KASS, R. E., and A. E. RAFTERY. 1995. Bayes factors and model uncertainty. *J. Am. Stat. Assoc.* **90**:773–795
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- KIMURA, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci., USA* **78**:454–458.
- LARGET, B., and D. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**:750–759.
- LAVINE, M., and M. J. SCHERVISH. 1999. Bayes factors: What they are and what they are not. *Am. Stat.* **53**:119–122.
- LEBRETON, J-D., K. P. BURNHAM, J. CLOBERT, and D. R. ANDERSON. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Mono.* **62**:67–118.
- LOSOS, J. B., T. R. JACKMAN, A. LARSON, K. DE QUEIROZ, and L. RODRÍGUEZ-SCHETTINO. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. *Science* **279**:2115–2118.
- MATHEWS, S., and M. J. DONOGHUE. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**:947–950.
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. W. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**:1087–1091.
- MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Mol. Biol. Evol.* **11**:715–724.
- POSADA, D. 2003. Using Modeltest and PAUP\* to select a model of nucleotide substitution. Pages ? in *Current Protocols in Bioinformatics* (A. D. Baxevanis, et al., eds.). John Wiley & Sons, Inc.

- POSADA, D., and K. A. CRANDALL. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* **14**:817–818.
- RAFTERY, A. E. 1996. Hypothesis testing and model selection. Pages 163–187 in *Markov chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, D. J. Spiegelhalter, eds.). Chapman & Hall, London.
- RONQUIST, F., and J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572–1574.
- SCHÖNIGER, M., and A. VON HAESLER. 1994. A stochastic model and the evolution of autocorrelated DNA sequences. *Mol. Phyl. Evol.* **3**:240–247.
- SCHWARZ, G. 1978. 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**:1001–1013.
- STREELMAN, J. T., M. E. ALFARO, M. W. WESTNEAT, D. R. BELLWOOD, and S. A. KARL. 2002. Evolutionary history of the parrotfishes: Biogeography, ecomorphology, and comparative diversity. *Evolution* **56**:961–971.
- SWOFFORD, D. L. 2002. PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pages 57–86 in *Lectures in Mathematics in the Life Sciences*, vol. 17.
- TIERNEY, L. 1994. Markov chains for exploring posterior distributions. *Ann. Statist.* **22**:1701–1762.
- VAN DEN BUSSCHE, R. A., R. J. BAKER, J. P. HUELSENBECK, and D. M. HILLIS. 1998. Base compositional bias and phylogenetic analyses: A test of the “flying DNA” hypothesis. *Molecular Phylogenetics and Evolution* **13**:408–416.
- VERDINELLI, I., and L. WASSERMAN. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.* **90**:614–618.
- YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**:306–314.
- YANG, Z., R. NIELSEN, N. GOLDMAN, and A. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.

**Table 1**  
**All possible time-reversible models of DNA substitution**

$K$	Models				
1	$M_1 = 111111$				
2	$M_2 = 122222$	$M_3 = 121111$	$M_4 = 112111$	$M_5 = 111211$	$M_6 = 111121$
	$M_7 = 111112$	$M_8 = 112222$	$M_9 = 121222$	$M_{10} = 122122$	$M_{11} = 122212$
	$M_{12} = 122221$	$M_{13} = 122111$	$M_{14} = 121211$	$M_{15} = 121121$	$M_{16} = 121112$
	$M_{17} = 112211$	$M_{18} = 112121$	$M_{19} = 112112$	$M_{20} = 111221$	$M_{21} = 111212$
	$M_{22} = 111122$	$M_{23} = 111222$	$M_{24} = 112122$	$M_{25} = 112212$	$M_{26} = 112221$
	$M_{27} = 121122$	$M_{28} = 121212$	$M_{29} = 121221$	$M_{30} = 122112$	$M_{31} = 122121$
	$M_{32} = 122211$				
3	$M_{33} = 123333$	$M_{34} = 123222$	$M_{35} = 122322$	$M_{36} = 122232$	$M_{37} = 122223$
	$M_{38} = 123111$	$M_{39} = 121311$	$M_{40} = 121131$	$M_{41} = 121113$	$M_{42} = 112311$
	$M_{43} = 112131$	$M_{44} = 112113$	$M_{45} = 111231$	$M_{46} = 111213$	$M_{47} = 111123$
	$M_{48} = 123333$	$M_{49} = 123233$	$M_{50} = 123323$	$M_{51} = 123332$	$M_{52} = 123322$
	$M_{53} = 123232$	$M_{54} = 123223$	$M_{55} = 122332$	$M_{56} = 122323$	$M_{57} = 122233$
	$M_{58} = 121333$	$M_{59} = 123133$	$M_{60} = 123313$	$M_{61} = 123331$	$M_{62} = 112333$
	$M_{63} = 112322$	$M_{64} = 112232$	$M_{65} = 112223$	$M_{66} = 123122$	$M_{67} = 123212$
	$M_{68} = 123221$	$M_{69} = 121322$	$M_{70} = 121232$	$M_{71} = 121223$	$M_{72} = 122312$
	$M_{73} = 122321$	$M_{74} = 122132$	$M_{75} = 122123$	$M_{76} = 122231$	$M_{77} = 122213$
	$M_{78} = 123311$	$M_{79} = 123131$	$M_{80} = 123113$	$M_{81} = 121331$	$M_{82} = 121313$
	$M_{83} = 121133$	$M_{84} = 123211$	$M_{85} = 123121$	$M_{86} = 123112$	$M_{87} = 122311$
	$M_{88} = 122131$	$M_{89} = 122113$	$M_{90} = 121321$	$M_{91} = 121312$	$M_{92} = 121231$
	$M_{93} = 121213$	$M_{94} = 121132$	$M_{95} = 121123$	$M_{96} = 112331$	$M_{97} = 112313$
	$M_{98} = 112133$	$M_{99} = 112321$	$M_{100} = 112312$	$M_{101} = 112231$	$M_{102} = 112213$
	$M_{103} = 112132$	$M_{104} = 112123$	$M_{105} = 111233$	$M_{106} = 111232$	$M_{107} = 111223$
	$M_{108} = 112233$	$M_{109} = 112323$	$M_{110} = 112332$	$M_{111} = 121233$	$M_{112} = 121323$
	$M_{113} = 121332$	$M_{114} = 122133$	$M_{115} = 122313$	$M_{116} = 122331$	$M_{117} = 123123$
	$M_{118} = 123132$	$M_{119} = 123213$	$M_{120} = 123231$	$M_{121} = 123312$	$M_{122} = 123321$
4	$M_{123} = 123444$	$M_{124} = 123433$	$M_{125} = 123343$	$M_{126} = 123334$	$M_{127} = 123422$
	$M_{128} = 123242$	$M_{129} = 123224$	$M_{130} = 122342$	$M_{131} = 122324$	$M_{132} = 122234$
	$M_{133} = 123411$	$M_{134} = 123141$	$M_{135} = 123114$	$M_{136} = 121341$	$M_{137} = 121314$
	$M_{138} = 121134$	$M_{139} = 112341$	$M_{140} = 112314$	$M_{141} = 112134$	$M_{142} = 111234$
	$M_{143} = 123344$	$M_{144} = 123434$	$M_{145} = 123443$	$M_{146} = 123244$	$M_{147} = 123424$
	$M_{148} = 123442$	$M_{149} = 122344$	$M_{150} = 122343$	$M_{151} = 122334$	$M_{152} = 123423$
	$M_{153} = 123432$	$M_{154} = 123243$	$M_{155} = 123234$	$M_{156} = 123342$	$M_{157} = 123324$
	$M_{158} = 123144$	$M_{159} = 123414$	$M_{160} = 123441$	$M_{161} = 121344$	$M_{162} = 121343$
	$M_{163} = 121334$	$M_{164} = 123413$	$M_{165} = 123431$	$M_{166} = 123143$	$M_{167} = 123134$
	$M_{168} = 123341$	$M_{169} = 123314$	$M_{170} = 112344$	$M_{171} = 112343$	$M_{172} = 112334$
	$M_{173} = 112342$	$M_{174} = 112324$	$M_{175} = 112234$	$M_{176} = 123412$	$M_{177} = 123421$
	$M_{178} = 123142$	$M_{179} = 123124$	$M_{180} = 123241$	$M_{181} = 123214$	$M_{182} = 121342$
	$M_{183} = 121324$	$M_{184} = 121234$	$M_{185} = 122341$	$M_{186} = 122314$	$M_{187} = 122134$
5	$M_{188} = 123455$	$M_{189} = 123454$	$M_{190} = 123445$	$M_{191} = 123453$	$M_{192} = 123435$
	$M_{193} = 123345$	$M_{194} = 123452$	$M_{195} = 123425$	$M_{196} = 123245$	$M_{197} = 122345$
	$M_{198} = 123451$	$M_{199} = 123415$	$M_{200} = 123145$	$M_{201} = 121345$	$M_{202} = 112345$
6	$M_{203} = 123456$				

NOTE.— $K$  is the number of substitution types. The named models are  $M_1$ ,  $M_{15}$ ,  $M_{40}$ ,  $M_{122}$ ,  $M_{168}$ ,  $M_{195}$ , and  $M_{203}$ .

**Table 2**  
**The DNA sequence alignments examined in this study**

Name	Gene	# Taxa	# Sites	Reference
Angiosperms	<i>phyA</i> & <i>phyC</i>	46	1104	Matthews and Donoghue (1999)
Archaea	rRNA	64	1620	Barns et al. (1996)
Bats	<i>IRBP</i>	13	1255	Van Den Bussche et al. (1998)
Butterflies	<i>wingless</i>	106	378	Brower (2000)
Crocodiles	<i>c-myc</i>	68	818	Harshman et al. (2003)
Gophers	mtDNA: <i>COI</i>	15	379	Hafner et al. (1994)
HIV-I	<i>env</i>	13	273	Yang et al. (2000)
HIV-I	<i>pol</i>	23	2841	Yang et al. (2000)
Lice	mtDNA: <i>COI</i>	17	379	Hafner et al. (1994)
Lizards	mtDNA: <i>ND2</i> + 5 tRNAs	30	1456	Losos et al. (1998), Jackman et al. (1999)
Mammals	mtDNA	23	9741	Arnason et al. (1997)
Parrotfish	12S, 16S, <i>cyt b</i> , <i>Timo-4C4</i>	18	1689	Streelman et al. (2002)
Primates	partial mtDNA: 3 tRNA's, <i>ND2</i> , <i>ND5</i>	12	898	Hayasaka et al. (1988)
Vertebrates	$\beta$ -globin	17	432	Yang et al. (2000)
Water snakes	mtDNA: <i>cyt b</i> , <i>ND2</i> , 12S rDNA	34	2866	Alfaro and Arnold (2001)
Whales	mtDNA: <i>cyt b</i>	31	1140	Adachi and Hasegawa (1995)

**Table 3**  
**The best models for each data set**

Name	AIC	BIC	PP	BF	95% Credible Set of Models
Angiosperms	189	189	189 (0.55)	248.8	(189, 193, 125, 147)
Archaea	198	198	168 (0.53)	230.1	(168, 198, 193)
Bats	50	50	112 (0.32)	95.0	(112, 50, 162, 147, 125, 152, 90, 183, 157, 122, 15, 189)
Butterflies	125	193	136 (0.32)	94.8	(136, 162, 112, 168, 40, 90, 125, 201, 191, 198, 183, 152)
Crocodiles	134	134	125 (0.28)	77.4	(125, 40, 166, 168, 134, 168, 189, 191, 162, 193)
Gophers	162	162	112 (0.28)	77.5	(112, 162, 15, 50, 40, 189, 125, 147, 95, 90, 138, 201, 183, 136, 117, 152, 122, 191)
HIV-1 ( <i>env</i> )	25	25	25 (0.29)	83.0	(25, 60, 50, 64, 100, 125, 102, 97, 164, 169, 152, 159, 173, 157, 175, 147, 171, 191, 193, 189, 140, 117)
HIV-1 ( <i>pol</i> )	157	52	50 (0.62)	335.2	(50, 125, 157, 152, 147, 193)
Lice	15	15	15 (0.56)	260.0	(15, 40, 117, 90, 50, 122, 136, 95, 166, 112, 125)
Lizards	193	203	193 (0.70)	481.1	(193, 138, 200, 203)
Mammals	203	203	193 (0.64)	364.3	(193, 203)
Parrotfish	189	189	162 (0.56)	258.0	(162, 189, 201)
Primates	112	112	15 (0.31)	91.0	(15, 40, 112, 95, 138, 162, 90, 136, 50, 125, 168, 122, 166, 117, 134)
Vertebrates	125	125	125 (0.21)	52.3	(125, 40, 168, 64, 134, 189, 166, 193, 191, 162, 136, 171, 198, 138, 50, 175, 173)
Water snakes	191	191	166 (0.55)	242.9	(166, 191, 117, 152, 134, 200, 198, 177)
Whales	162	162	15 (0.60)	300.1	(15, 40, 117, 95, 85, 122, 112, 90, 134, 50, 166)

NOTE.—AIC, the best model using the Akaike Information Criterion; BIC, the best model using the Bayesian Information Criterion; PP, the model with the highest posterior probability, with its corresponding probability; BF, the Bayes factor for the best model.

**Table 4**  
**The Bayes factors for  $K$ , the number of substitution types**

Name	$K$					
	1	2	3	4	5	6
Angiosperms	—	—	0.01	0.56	34.6	10.3
Archaea	—	—	—	2.42	9.64	6.72
Bats	—	0.07	2.59	0.79	0.52	0.43
Butterflies	—	0.01	0.37	3.72	1.84	0.84
Crocodiles	—	—	0.41	4.17	1.23	0.38
Gophers	—	0.54	1.01	1.32	1.06	0.61
HIV-1 ( <i>env</i> )	—	2.31	0.95	0.62	0.59	0.60
HIV-1 ( <i>pol</i> )	—	—	2.19	0.94	0.73	0.40
Lice	—	7.14	0.68	0.19	0.04	0.02
Lizards	—	—	0.01	0.37	48.3	11.0
Mammals	—	—	—	—	22.6	112.0
Parrotfish	—	—	—	2.73	8.58	6.48
Primates	—	2.50	1.09	0.55	0.24	0.09
Vertebrates	—	0.04	0.48	2.66	2.28	1.56
Water snakes	—	—	0.11	4.53	3.58	2.46
Whales	—	8.24	0.59	0.18	0.05	0.01

NOTE.—Some entries in the table are blank because the Bayes factor is too small to be accurately measured when running the MCMC algorithm with a uniform prior on models. They can be accurately calculated, but only by giving models with small  $K$  more prior weight, and then calculating what the posterior probability would have been if a uniform prior had been assumed.

FIG 1.—The posterior probability of  $K$ , the number of substitution types, for each DNA sequence alignment. The first figure shows the prior probability distribution of  $K$ , calculated as the number of models with  $K$  substitution parameters divided by the total number of substitution models (203).

