Estimating the Rate of Evolution of the Rate of Molecular Evolution

Jeffrey L. Thorne, Hirohisa Kishino,† and Ian S. Painter**

*Program in Statistical Genetics, Statistics Department, North Carolina State University; and †Department of Social and International Relations, University of Tokyo

A simple model for the evolution of the rate of molecular evolution is presented. With a Bayesian approach, this model can serve as the basis for estimating dates of important evolutionary events even in the absence of the assumption of constant rates among evolutionary lineages. The method can be used in conjunction with any of the widely used models for nucleotide substitution or amino acid replacement. It is illustrated by analyzing a data set of *rbcL* protein sequences.

Introduction

If the rate of sequence evolution is the same in different evolutionary lineages, then comparison of homologous gene sequences will facilitate chronological dating of speciation and gene duplication events. In addition to shedding light on the pattern of historical evolutionary events, the possibility that different lineages change at about the same rate, if confirmed, would be an important tool for the characterization of evolutionary processes. The molecular clock hypothesis of Zuckerkandl and Pauling (1965) has therefore generated careful scrutiny (e.g., Kimura 1983; Ohta 1987; Gillespie 1991).

A variety of statistical tests have been proposed that evaluate whether a particular data set is consistent with a null hypothesis of a molecular clock (e.g., Langley and Fitch 1974; Felsenstein 1981; Wu and Li 1985; Tajima 1993). Applications of these tests to specific data sets have shown that the molecular-clock hypothesis can often be rejected (e.g., Gaut et al. 1992). It might not be surprising, for example, that rates of evolution differ between mammalian and viral lineages. On the other hand, it does seem that there is a correlation of evolutionary rates among closely related evolutionary lineages. This correlation is biologically plausible because factors that may be responsible for divergence of rates among lineages (e.g., population size, generation time, fidelity of DNA replication) may themselves be quite similar among closely related lineages. It may be the gradual divergence of these factors that is responsible for the gradual divergence of evolutionary rates among lineages.

If the divergence among factors that impact evolutionary rates were better understood, dating of evolutionary events from comparisons of homologous sequences could be performed even without an assumption that the rates of evolution for different lineages are exactly equal. Similarly, inferences made about evolutionary processes would be more accurate. Much more work regarding the identification and characterization of factors affecting evolutionary rates of lineages still needs to be done, but it is clear that the divergence of many of these factors will be lineage-specific rather than gene-specific. This is why reliance upon a molecular clock is likely to be unwarranted even for the analysis of data sets with sequences from many different loci.

When the goal is to date evolutionary events with data that clearly violate a molecular clock, the two most readily available options are both unsatisfactory. One of these options is to ignore evidence that data are inconsistent with the clock and to employ a method for dating under the incorrect assumption that the clock is valid. The other option is to ignore data that are inconsistent with a molecular clock and thereby discard information pertinent to chronological dates that these data may contain. A more satisfactory choice lies between these two undesirable extremes. We describe one approach for extracting dating information from data sets that violate the molecular clock assumption.

Method

We borrow ideas from the extensive research by Gillespie (1991) on, as he terms it, ''the rate of evolution of the rate of evolution.'' If evolutionary events (e.g., nucleotide substitutions) occur independently, then the number of evolutionary events that occur on a branch existing from time 0 to time T and having rate $R(t)$ at time t follows a Poisson distribution with mean

$$B(T) = \int_{t=0}^{T} R(t)\, dt \qquad (1)$$

(Gillespie 1991). We will refer to $B(T)$ as a branch length.

A difficulty is that $R(t)$ cannot be directly observed. One way to overcome this problem is to adopt the restrictive molecular clock assumption of a constant rate with respect to time. Another is to avoid an explicit model of how the rate changes with time and instead to estimate the integral in equation (1) separately for each branch. This has become the usual procedure for maximum-likelihood reconstruction of phylogenies (e.g., Felsenstein 1981). Because no attempt is made to model how the rate changes with time, the usual procedure is not directly applicable to chronological dating.

Statistical and Computational Issues

As the process of molecular evolution becomes better understood, the models for statistically describing

Abbreviation: r.t.u., relative time unit.

Key words: molecular clock, phylogeny, Markov chain Monte Carlo, Metropolis-Hastings algorithm.

Address for correspondence and reprints: Jeffrey L. Thorne, Program in Statistical Genetics, Statistics Department, Box 8203, North Carolina State University, Raleigh, North Carolina 27695-8203. E-mail: thorne@statgen.ncsu.edu.

Mol. Biol. Evol. 15(12):1647–1657. 1998

this process will inevitably become more complex. A cost of the adoption of complicated and parameter-rich models is that the variance of parameter estimates rises as the number of parameters increases. At the extreme, a model contains so many parameters that the number of parameters exceeds the number of available data points, and problems of statistical identifiability ensue. Our own attempts to explicitly model how evolutionary rates might change over time and then apply maximum-likelihood techniques have not been successful due to issues of both computational tractability and statistical identifiability.

Computational concerns become particularly important when the goal is to apply maximum-likelihood techniques to a situation in which the rates of molecular evolution are treated as random variables. In such a case, evaluation of the likelihood requires integration over all trajectories of rates of molecular evolution that are possible with a specific phylogeny. In standard applications of maximum likelihood to phylogeny reconstruction, the pruning algorithm of Felsenstein (1981) makes maximum likelihood feasible. In these cases, the sites of sequences are assumed to evolve independently, and the likelihood for a data set is simply the product of the individual site likelihoods. When it is necessary to integrate over rate trajectories that are shared by all sites, the pruning algorithm in its usual form is no longer sufficient to provide computational tractability, because the likelihood becomes an integral of a product. If a model were proposed that had only a few possible rate trajectories, likelihood calculations would be feasible but, because of extreme limitations on the number of possible rate trajectories, the potential value of these models would be reduced.

One way to analyze a complicated model and simultaneously avoid computational difficulties and undesirable statistical behavior is to adopt a Bayesian perspective. Bayesian methods have recently been proposed for phylogeny reconstruction from data sets of both nonaligned sequences (Allison and Wallace 1994) and aligned sequences (Rannala and Yang 1996; Mau and Newton 1997; Yang and Rannala 1997). In Bayesian analyses, a priori knowledge about parameter values is summarized through assignment of probability distributions known as priors. The observed data and the prior distributions are then used to determine probability distributions known as posteriors. The posterior distribution is a probability distribution representing uncertainty about the parameters after observing the data. It is the posterior distribution that serves as the basis for Bayesian inference. Although parameters about which there is little information in the data will have a posterior distribution that is similar to their prior distribution, an analysis that includes these parameters may still be more realistic than one in which they are absent, because their prior (and posterior) distributions will be concentrated around biologically reasonable values. Bayesian inference is thus less subject to problems of overparameterization.

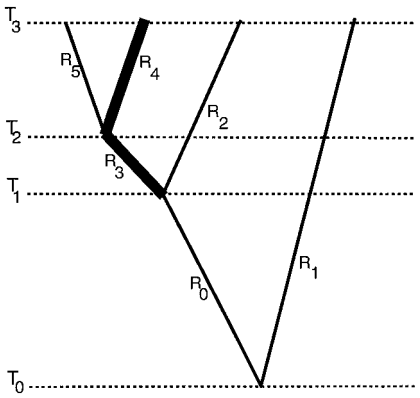

FIG. 1.—Rates and node times on an example tree.

Hierarchical Model of Rate Evolution

Our purpose here is to present a relatively simple model for the stochastic process that governs the change of rate with respect to time. This model specifies a prior distribution of the evolutionary rates along branches. By allowing autocorrelation of rates over evolutionary time, the model has the potential to provide an improved statistical fit to sequence data sets. Our hope is that this can serve as the basis for future, more realistic, models, just as the Jukes-Cantor model of nucleotide substitution (Jukes and Cantor 1969) has served as the basis for many subsequently proposed more realistic models of nucleotide substitution. We explain our approach with reference to amino acid replacements, but other types of evolutionary events, such as nucleotide substitutions or insertions and deletions, could also be studied.

We make the simplifying assumption that the rate of molecular evolution is constant on any particular branch of an evolutionary tree, but we allow rates to differ among branches. The rate of branch i will be denoted R_i. The autocorrelation of rates between an ancestral branch and its direct descendant will depend on the time difference between the midpoints of the ancestral and the descendant branches. For example, the time difference between the thickened ancestral and descendant branches in figure 1 is

$$\frac{T_2 + T_3}{2} - \frac{T_1 + T_2}{2} = \frac{T_3 - T_1}{2}.$$

We assume that the logarithm of the rate on the descendant branch has a normal distribution with a mean equal to the logarithm of the rate on the ancestral branch and with a variance equal to the time difference multiplied by a constant that we will refer to as ν. A high value of ν means there is little rate autocorrelation, and a low value implies strong rate autocorrelation.

By Bayesian convention, a parameter governing a prior distribution is called a hyperparameter. In our model, the value of ν determines the prior distribution for the rates of molecular evolution on different branches given the internal node times. Because the value of the hyperparameter ν can have a strong influence on an analysis, we add another level to our hierarchical model. The additional level in the hierarchy allows flexibility

for the Bayesian analysis by incorporating uncertainty regarding the appropriate value of ν. We add the additional level via an exponential prior distribution $p(\nu)$ for ν. Typically, model levels above this additional level provide little benefit in Bayesian inference (e.g., Carlin and Louis 1996, p. 24).

As well as specifying the relationship between rates for descendant and ancestral branches, the distribution of rates for the two branches on a bifurcating tree that directly emanate from the root need to be included in the model. One of these two branches is selected and referred to as branch 0. Branch 0 will have a rate R_0 that is sampled from some specific prior distribution $p(R_0)$. This prior distribution is assumed to be independent of ν and the internal node times for the tree. We have used an exponential distribution for $p(R_0)$. The logarithm of the rate on the second of the two branches connected to the root is assumed to be sampled from a normal distribution with a mean equal to $\log(R_0)$ and a variance equal to ν multiplied by the mean of the durations of the two branches that are directly connected to the root. For the tree depicted in figure 1, the mean would be $(T_1 - T_0)/2 + (T_3 - T_0)/2$.

The process that we have described for the lognormal changes of rates over time is not stationary. With respect to figure 1, the prior distribution from which R_5 is sampled depends on, but will differ from, the prior distribution from which R_0 is sampled. It is unclear whether this lack of stationarity should be viewed as an advantage or a disadvantage. Alternative models warrant further explanation.

To fix notation, let $R = (R_0, R_1, \ldots, R_k)$ be the rates of molecular evolution on the $k + 1$ branches of the rooted tree, and let T be a vector that specifies the internal node times (including the root). Once the times T and the constant ν are determined, the conditional distribution $p(R \mid T, \nu)$ of the rates of molecular evolution is determined as above. The distribution of R_0 is fixed a priori and is not affected by the values of T and ν.

A binary-splitting branching process, or Yule process (e.g., Karlin and Taylor 1975, pp. 119–123), specifies $p(T)$, the prior probability distribution for T. We have selected $p(T)$ on the basis of simplicity of the Yule process. More complicated and potentially more realistic choices of priors for $p(T)$ (e.g., Yang and Rannala 1997) can be studied in the future so as to reflect effects of lineage extinction and taxonomic sampling.

Our prior is generated from the assumption that the Yule process begins with one lineage splitting into two lineages. Births of new lineages continue according to the Yule process until immediately before the birth event that would result in there being one more lineage than there are tips on the tree of interest. Only one parameter, the birth rate, is needed to define $p(T)$ for this Yule process. For a given number of sequences representing tips on the tree, the birth rate can be used to calculate the expected value of $-T_0$, where T_0 is the time of a common ancestral sequence and the time at the tips of the tree is assumed to be 0. Likewise, the birth rate can be determined from the expected value of $-T_0$.

Although the prior for T is conditional on the birth rate, this fact is omitted from the notation for the sake of clarity. Similarly, the priors for ν and R_0 are exponential, but the parameters upon which these priors depend are not included in the notation.

Posterior Distribution

For a data set X of aligned homologous sequences, the posterior distribution depends on $p(T, R, \nu \mid X)$ through

$$p(T \mid X) = \int_R \int_\nu p(T, R, \nu \mid X)\, d\nu\, dR. \qquad (2)$$

The distribution $p(T, R, \nu \mid X)$ is

$$\begin{aligned} p(T, R, \nu \mid X) &= \frac{p(X, T, R, \nu)}{p(X)} \\ &= \frac{p(X \mid T, R, \nu)p(T, R, \nu)}{p(X)} \\ &= \frac{p(X \mid T, R, \nu)p(R \mid T, \nu)p(T \mid \nu)p(\nu)}{p(X)} \\ &= \frac{p(X \mid T, R)p(R \mid T, \nu)p(T)p(\nu)}{p(X)}. \end{aligned} \qquad (3)$$

The last step is justified by assuming that the value of ν neither provides information about the divergence times T nor, if both the rates R and the divergence times T are known, about the data X. Letting $B = (B_0, \ldots, B_k)$ represent the lengths of the branches on the tree, we have

$$p(T, R, \nu \mid X) = \frac{p(X \mid B)p(R \mid T, \nu)p(T)p(\nu)}{p(X)}, \qquad (4)$$

because $p(X \mid T, R) = p(X \mid B)$.

Metropolis-Hastings Algorithm

Although the numerator of equation (4) can be directly calculated, the denominator $p(X)$ is more difficult to evaluate, because multiple integration over T, R, and ν is required. This fact calls for adoption of the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) to obtain an approximately random sample from $p(T, R, \nu \mid X)$. The Metropolis-Hastings algorithm is a Markov chain Monte Carlo technique that permits construction of a Markov chain on the parameter space (T, R, ν) such that the stationary distribution of the chain is $p(T, R, \nu \mid X)$.

The algorithm begins at some initial state that satisfies $p(T, R, \nu \mid X) > 0$. Given the state (T, R, ν) of the Markov chain, a new state (T', R', ν') is then randomly proposed according to a proposal density denoted by $J(T', R', \nu' \mid T, R, \nu)$. The proposed state is accepted with probability r, where

$$r = \min\left(1, \frac{p(T', R', \nu' \mid X)J(T, R, \nu \mid T', R', \nu')}{p(T, R, \nu \mid X)J(T', R', \nu' \mid T, R, \nu)}\right). \qquad (5)$$

Notice that the $p(X)$ term of equation (4) need not be calculated to determine the above ratio, because

$$\frac{p(T', R', \nu' \mid X)}{p(T, R, \nu \mid X)} = \frac{p(X \mid B')p(R \mid T', \nu')p(T')p(\nu')}{p(X \mid B)p(R \mid T, \nu)p(T)p(\nu)}. \quad (6)$$

If the proposed state is accepted, it becomes the next state of the Markov chain. If it is rejected, the next state of the Markov chain will be identical to the current state. Then, the Metropolis-Hastings algorithm continues by again randomly proposing a state. Repeating many cycles of this procedure of random proposals followed by acceptance or rejection produces a Markov chain with a stationary distribution that is the desired posterior distribution $p(T, R, \nu \mid X)$. Details of our algorithm are in the appendix.

Multivariate Normal Approximation

The rate at which the Markov chain generated by the Metropolis-Hastings algorithm converges to its stationary distribution $p(T, R, \nu \mid X)$ determines whether the approach is computationally feasible. Our implementation requires calculation of $p(X \mid B)$ for many different values of B. In an entire analysis, $p(X \mid B)$ may have to be evaluated millions of times to obtain a suitable approximately random sample from $p(T, R, \nu \mid X)$. The usual maximum-likelihood approach also involves repeatedly evaluating the likelihood $p(X \mid B)$, but the purpose of the repeated evaluations is simply to maximize the likelihood. As a result, the Markov chain Monte Carlo analysis may require many times more evaluations than the maximum-likelihood approach. Therefore, a Bayesian analysis would be computationally prohibitive if calculation of $p(X \mid B)$ was too slow. To address this computational issue, we use an approximation for the likelihood $p(X \mid B)$ that is quick to evaluate.

We approximate the likelihood surface with a multivariate normal distribution. The mean of the multivariate normal distribution is B, the maximum-likelihood estimate of the branch lengths. The covariance matrix of the multivariate normal distribution is estimated from the curvature of the log-likelihood surface (i.e., the inverse of the information matrix; Stuart and Ord 1991, pp. 675–676). A Taylor series expansion (through the quadratic term) of the log-likelihood function around its maximum shows that the likelihood function can be approximated with the density of a multivariate normal distribution multiplied by a constant that does not depend on B (for a related approximation, see Gelman et al. [1995, pp. 94–95]). The value of the constant need not be determined, because our Metropolis-Hastings calculations all involve the ratio $p(X \mid B')/p(X \mid B)$, and the constant is identical for the approximations of the numerator and denominator of this ratio. In fact, we have structured our Markov chain Monte Carlo algorithm so that $p(X \mid B')/p(X \mid B)$ will either be 1 or a ratio of univariate normal densities (see appendix).

In our implementation, we root the tree topology of interest with an outgroup and assume that the topology relating the sequences in our data is known. We then obtain maximum-likelihood estimates of branch lengths for the unrooted topology consisting of the outgroup and the ingroup. The next step is to estimate the covariance matrix for the multivariate normal approximation. Because branches with an inferred length of zero can negatively impact the approximation of the covariance matrix, we treat them specially but omit the technical details here. Because we use the outgroup only to root the topology relating the sequences of interest, the multivariate normal distribution needed by the Markov chain Monte Carlo algorithm is determined by integrating the multivariate normal distribution calculated from the log-likelihood surface over all outgroup branches. The covariance matrix that results is simply the covariance matrix calculated from the log-likelihood surface with all rows and columns corresponding to outgroup branches removed.

Example

Sanderson (1997) recently demonstrated an approach for estimating divergence times with a data set of *rbcL* DNA sequences. We elected to follow this example by illustrating our method with a data set of 31 amino acid sequences from the *rbcL* chloroplast gene. Amino acid sequences rather than DNA sequences were selected to highlight the fact that our method can be implemented in conjunction with complex and comparatively realistic models of sequence evolution.

Alignment of the *rbcL* sequences was straightforward, because few insertions and deletions seem to have occurred since divergence of the proteins from their common ancestral sequence. The topology relating the sequences was fixed so that it is compatible with the one inferred by Sanderson (1997). A *Marchantia* sequence served as the outgroup and allowed the tree relating the remaining 30 ingroup sequences to be rooted. A model of amino acid replacement that attempts to incorporate the impact of protein secondary structure and solvent accessibility on the process of amino acid replacement (Goldman, Thorne, and Jones 1998) was assumed. Branch lengths $\hat{B}$ were estimated wuth this model (see fig. 2).

Prior Specification

In addition to approximating $p(X|B)$ by finding $\hat{B}$ and estimating its covariance matrix, prior distributions need to be specified. The issue of prior distribution specification arises frequently in Bayesian applications. A formal Bayesian approach entails specifying the prior without the assistance of the data. A pragmatic approach is to adopt the empirical Bayesian strategy of estimating the hyperparameters that govern the prior distributions from the data. By not treating the hyperparameters as random variables, this empirical Bayesian approach violates the philosophy that attracts many to the Bayesian framework. However, our incentives for incorporating the prior distributions into our analysis are mainly to avoid overparameterization and achieve computational tractability.

One way to roughly estimate the hyperparameter for the prior distribution of the root depth is to set the mean of the prior distribution equal to the root depth estimated with maximum likelihood and the assumption

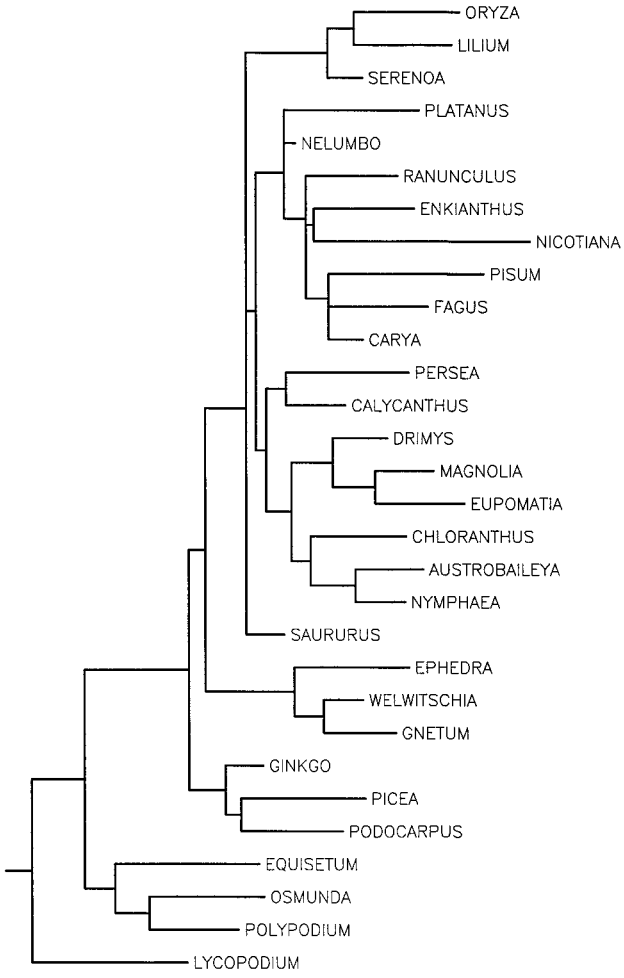

FIG. 2.—Branch lengths estimated by the amino acid replacement model for the rooted ingroup of *rbcL* sequences.

of a perfect molecular clock. The PAML package (Yang 1997) was used to analyze the 30 *rbcL* ingroup sequences under a hypothesis of a molecular clock. This analysis was done by assuming the Jones-Taylor-Thornton model (Jones, Taylor, and Thornton 1992) and constraining the topology to be identical to the one we assumed. The resulting estimated depth of the ingroup root was 6.03 amino acid replacements per 100 sites. The depth of the ingroup root was measured in terms of expected amino acid replacements per 100 sites rather than in terms of years, because fossil evidence was not used to calibrate the molecular clock.

We define 1 ''relative time unit'' (r.t.u.) as the expected amount of time for one amino acid replacement event to occur per 100 sites given some constant rate of replacement per year. The value of this rate need not be specified. It is used as a reference so that all other rates can be measured relative to it. With our model, rates vary among branches, and the expected number of amino acid replacements between root and tip will vary among tips. Therefore, if the depth of the root is 6 r.t.u., the path from the root to a particular tip may include mainly branches with a high rate of change and the expected number of replacements along the path may be more than 6 per 100 sites. Similarly, the expected number of replacements along a path from a root of depth 6 r.t.u. to a particular tip may be less than 6 per 100 sites.

Based on the results of the PAML analysis, we decided to set the prior distribution for the Yule process so that the expected depth $-T_0$ of the root was 6 r.t.u. The prior distribution for ν was exponential, with mean 1/6. The prior distribution for R_0 was exponential, with mean 1 replacement per 100 sites per r.t.u. We also investigated the effects of other prior distributions for ν and the root depth.

Implementation

To allow the Markov chain to reach stationarity, the Markov chain Monte Carlo algorithm completed 100,000 initial cycles before the state of the Markov chain was sampled. Thereafter, the Markov chain was sampled every 1,000 cycles until a total of 1,000 samples were collected. A single run required approximately 83 min of CPU time on a workstation with one 300-MHz UltraSPARC-II microprocessor.

To explore whether our procedure was generating an approximately random sample from the posterior distribution, some runs were performed without evaluating $p(X \mid B)$ and instead assuming that this likelihood was a constant that does not depend on the branch lengths B. If $p(X \mid B)$ and $p(X \mid B')$ are left out of the acceptance formula (see eqs. 5 and 6), the stationary distribution of the resulting Markov chain is just the prior distribution. Our samples did resemble samples that would be observed when directly sampling from the prior.

As another diagnostic for convergence of the Markov chain, we started the algorithm at different initial states. If the procedure is converging, samples that were obtained from different initial states should yield approximations similar to those for the posterior. This is what was observed. It should be noted that positive correlations between the parameter values of consecutive samples were obtained from our procedure. Although the procedure does seem to yield a reasonably accurate description of the posterior, improvements that increase the ''mixing'' rate of the Markov chain would reduce the computational burden.

Results

A rooted tree that summarizes our Bayesian analysis is shown in figure 3. Except for a few general points, we will not discuss the results of this analysis in much detail here, because our purpose is to demonstrate the feasibility of the method rather than to contribute to the literature on *rbcL* evolution.

In conventional maximum-likelihood phylogeny reconstruction, rates and times are not separately estimated. Instead, they enter the likelihood as a product. With the method described here, there is some ability to separately estimate rates of molecular evolution and node times, but our experience is that this ability is rather limited if the prior distribution for node times is relatively diffuse. This is reflected in table 1 by the high

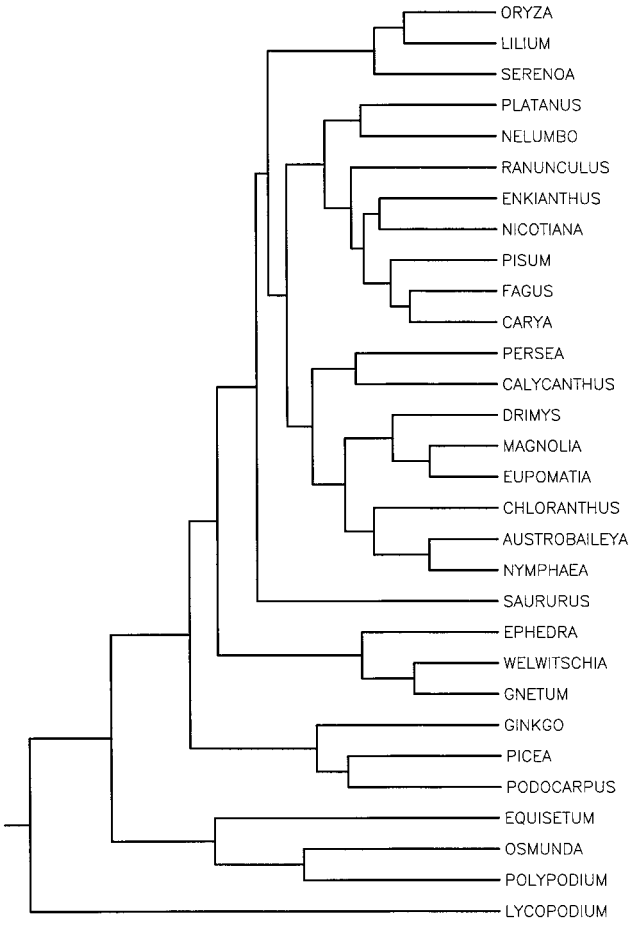

FIG. 3.—An *rbcL* tree obtained from the Bayesian analysis when the prior distribution had an expected value of 6 r.t.u. for R_0 and an expected value of 1/6 for ν. The proportion of the depth for each internal node relative to the root was calculated for each sample from the Markov chain Monte Carlo algorithm. The sample means of these proportions were used to draw the tree.

sensitivity of the posterior expectation of $-T_0$ to the prior distribution of $-T_0$. Although the posterior distribution of $-T_0$ is quite sensitive to its prior distribution, posterior distributions of the ratios of the depths of particular nodes to the root depth $-T_0$ are robust to the prior distribution (table 1). By normalizing rates and times of each sample generated by the Metropolis-Hastings algorithm so that the average rate on the sample tree is 1 and so that branch lengths are identical before and after normalization, we find that the normalized values of $-T_0$ are also robust to the prior distribution (table 2). The normalized values of ν seem to be more robust to the prior distribution than are the nonnormalized values (table 3).

The posterior means of the normalized times to the root are close to but greater than the root depth of 6.03 replacements per 100 sites estimated by the PAML program. The difference may be partially attributable to our analysis allowing evolutionary rates to evolve, but this explanation need not be invoked. The difference may also be explained by the fact that PAML ignores alignment columns with gaps, whereas our programs treat

Table 1
Estimated Posterior Means of the Root Depth and the Relative Depths of Two Internal Nodes for Four Combinations of Prior Distributions of ν and $-T_0$

		$E(-T_0) = 6$	$E(-T_0) = 18$
$E(\nu) = 1/6$	$-T_0$	4.98 (1.18)	14.14 (3.55)
	Conifer	0.39 (0.11)	0.39 (0.12)
	Eudicot	0.37 (0.09)	0.37 (0.09)
$E(\nu) = 1/18$ ...	$-T_0$	5.15 (1.23)	14.50 (3.49)
	Conifer	0.39 (0.11)	0.40 (0.11)
	Eudicot	0.37 (0.08)	0.37 (0.10)

NOTE.—Entries labeled "$-T_0$" are nonnormalized root depths. Entries labeled "conifer" are the proportions of depths of the nodes representing the most recent common ancestors of the conifers (i.e., *Ginkgo, Picea, Podocarpus*) relative to the root. Entries labeled "eudicot" are the proportions of depths of the nodes representing the most recent common ancestors of the eudicot angiosperms (i.e., *Platanus, Nelumbo, Ranunculus, Enkianthus, Nicotiana, Pisum, Fagus, Carya*) relative to the root. Estimated standard deviations are shown in parentheses.

gaps as missing data. Another possible source of the small difference is that the model of amino acid replacement assumed in the Bayesian analysis allows rate heterogeneity among sites that is associated with structural environment, but the Jones-Taylor-Thornton model does not.

The fossil record implies that the existence of the conifer clade probably predates the origin of the eudicot angiosperm clade (see Sanderson 1997). Our analysis finds the ages of these two clades to be more similar than they probably are (table 1 and fig. 3). This similarity may be attributable to the estimation of the branch lengths on the unrooted tree. It may also be explained by the prior distributions assumed in our analysis. Preliminary simulations show no obvious bias in the estimates of node depths relative to root depths (data not shown).

Discussion and Conclusions

Information concerning dates of evolutionary events can be extracted from sequence data sets even when the assumption of globally constant rates of evolution is unwarranted. Others have come to the same conclusion (e.g., Penny, Murray-McIntosh, and Hendy 1998) but have chosen different strategies for extracting this information. The method that is most similar to our own is probably the "nonparametric" technique outlined by Sanderson (1997). This technique shares with our approach the idea that autocorrelation of rates can be exploited to extract chronological information. In some sense, the technique of Sanderson and our highly parametric technique are at opposite ends of a statistical

Table 2
Posterior Means and Standard Deviations of Normalized Root Depths ($-T_0$)

	$E(-T_0) = 6$	$E(-T_0) = 18$
$E(\nu) = 1/6$	6.21 (1.10)	6.21 (1.12)
$E(\nu) = 1/18$	6.07 (1.06)	6.17 (1.11)

Table 3
Posterior Means and Standard Deviations of Nonnormalized and Normalized Values of ν for Four Combinations of Prior Distributions of ν and $-T_0$

		$E(-T_0) = 6$	$E(-T_0) = 18$
$E(\nu) = 1/6$ ...	Nonnormalized	0.18 (0.09)	0.07 (0.04)
	Normalized	0.14 (0.07)	0.16 (0.07)
$E(\nu) = 1/18$ ..	Nonnormalized	0.12 (0.05)	0.06 (0.03)
	Normalized	0.10 (0.04)	0.14 (0.06)

continuum. Rather than explicitly modeling the evolution of the rate of evolution, the Sanderson method extracts chronological information by optimizing what seems to be a reasonable criterion. As with our method, it allows different rates on different branches of the tree and assumes that rates are constant on individual branches. Unlike our method, it assumes that branch length estimates are the true branch lengths. With the assumptions underlying the Sanderson method, the rate on a branch is equal to the estimated branch length divided by the duration of the branch. The Sanderson technique infers internal node times by minimizing the sum over all ancestral and descendant branches of some function that penalizes a change in rates between the ancestral and descendant branches. This penalty function may take the form of the squared difference between ancestral and descendant rates, for example. Because both highly parametric and less parametric approaches enjoy great success in the field of statistics, it is premature to make general conclusions regarding the merits of our approach and that of Sanderson. Both warrant further exploration.

A promising feature of the Sanderson technique that we intend to implement is the inclusion of constraints. It would be straightforward to directly incorporate fossil evidence that a particular node of the tree has an age that equals or surpasses some minimum. Constraints have the potential to improve date estimation throughout the tree and will also influence the posterior distribution of ν. Consideration of constraints may greatly reduce the sensitivity of the posterior distributions of $-T_0$ and ν to their priors.

A point made by J. Kim and communicated by Sanderson (1997) is that methods of phylogeny reconstruction themselves could capitalize on the idea that closely related evolutionary lineages are likely to evolve at similar rates. A consequence of this rate autocorrelation is that branches that are nearby in a tree will have correlated lengths. Widely used methods of phylogeny reconstruction ignore this potential information.

An approach for reconstructing phylogenies that includes a model of evolution of the rate of evolution would benefit not only from incorporation of the tendency for branch lengths to be correlated due to rate autocorrelation structure, but also from the information that the time structure of a tree can provide. Even in the absence of rate autocorrelation, branch lengths will be correlated due to this time structure. For instance, in figure 1, the branch with rate R_4 and the branch with rate R_5 evolved for identical amounts of time. If the rates experienced by these branches were independent, the branch lengths would be correlated due to their existence for a common amount of time.

It is valid to attach a slightly different interpretation to the probabilistic structure of our model. Instead of having constant rates of molecular evolution on a branch and different rates of molecular evolution on different branches, the model can be viewed as describing the average rate of molecular evolution on branches and how this average rate differs among branches. To reduce computation, we do not explicitly model variation of rates of molecular evolution within a branch. A branch attached to a slowly evolving ancestral branch and two quickly evolving descendant branches probably experienced a higher rate of evolution near its end than near its beginning, but this is not reflected by our model. We cannot quantify the loss of information due to this inadequacy of our model, but we suspect that the loss is typically small.

The fact that different genes evolve at different rates is well established. The rate of molecular evolution probably evolves at different rates for different genes as well. With our model, this issue can be addressed by comparing the posterior distributions of ν for different genes.

Three general possibilities for changes in the rate of molecular evolution are depicted in figure 4. Figure 4*A* exhibits no autocorrelation; the rate of molecular evolution in one time interval is independent of the rate of molecular evolution in the next time interval. Figure 4*B* and *C* both show a positive correlation between the rates in successive intervals. In figure 4*B,* the logarithm of the rates drifts around a mean value of 0. In other words, figure 4*B* shows a stationary process. In figure 4*C,* the process is not stationary. Instead, figure 4*C* was generated according to a Brownian motion process. Therefore, the absolute value of the difference between the logarithm of the rate at time 0 and the logarithm of the rate at time t will tend on average to increase as t increases. In fact, this difference will tend toward infinity as t approaches infinity.

The model for change of the rate of molecular evolution that we have described and implemented here is approximately a Brownian motion process that operates on the logarithm of the rates of molecular evolution. However, the actual process affecting the rate of molecular evolution may behave more like the processes indicated by figure 4*A* or *B.* In the future, a specific hypothesis that the actual process falls into a particular one of these three categories could be evaluated in a Bayesian framework.

The method we have described accounts for uncertainty of branch length estimates as well as variation of rates due to evolution. By accounting for these two types of variation in an integrated framework, improved estimates of evolutionary dates and more accurate quantification of uncertainty is possible. One strength of our approach is that it can be used with all widely used models of nucleotide substitution and amino acid replacement. We are optimistic that generalizations and

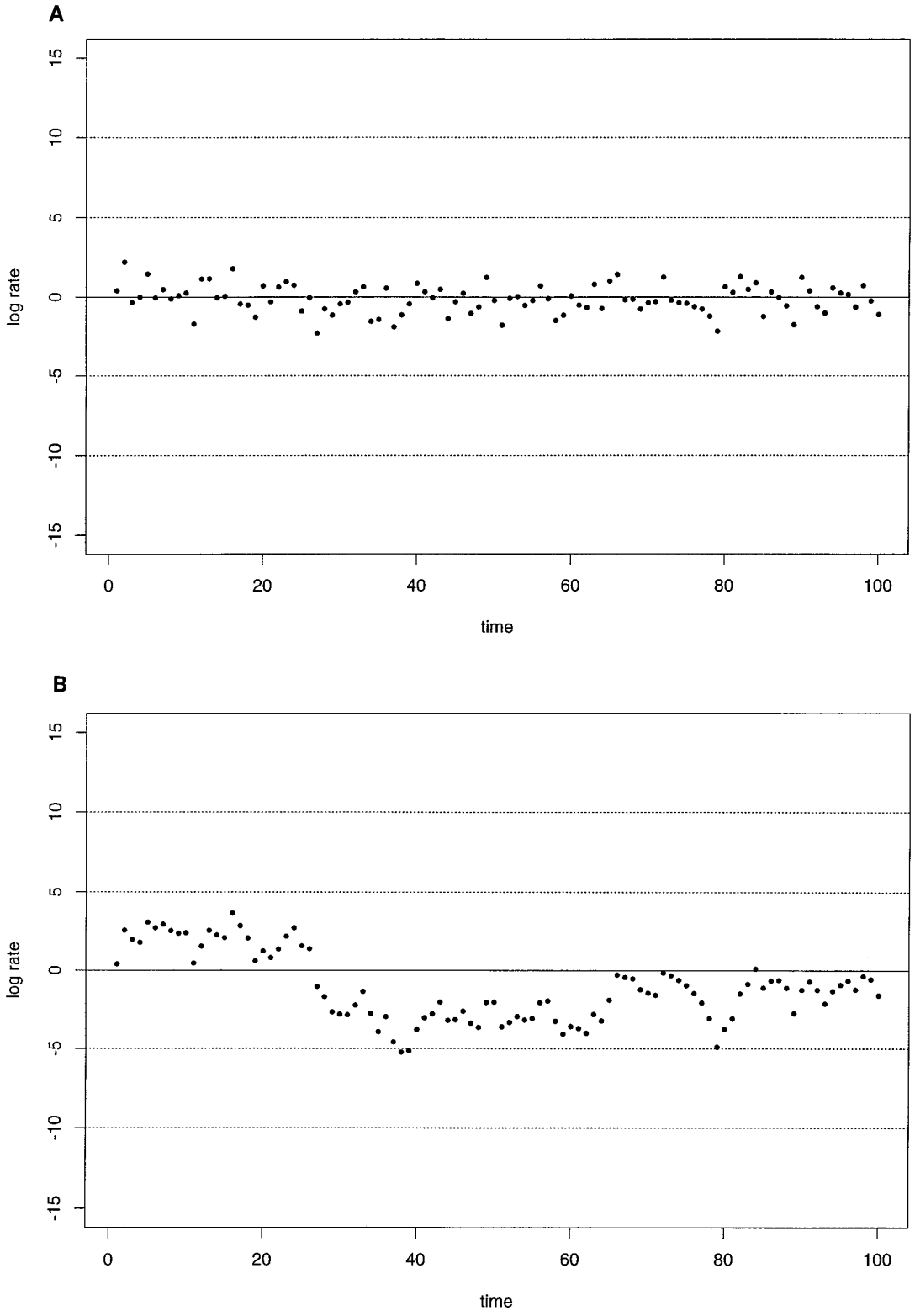

FIG. 4.—Relationship between the logarithm of the rate of molecular evolution and time. *A*, A process with no autocorrelation. *B*, A stationary process with positive autocorrelation. *C*, A nonstationary process with positive autocorrelation.

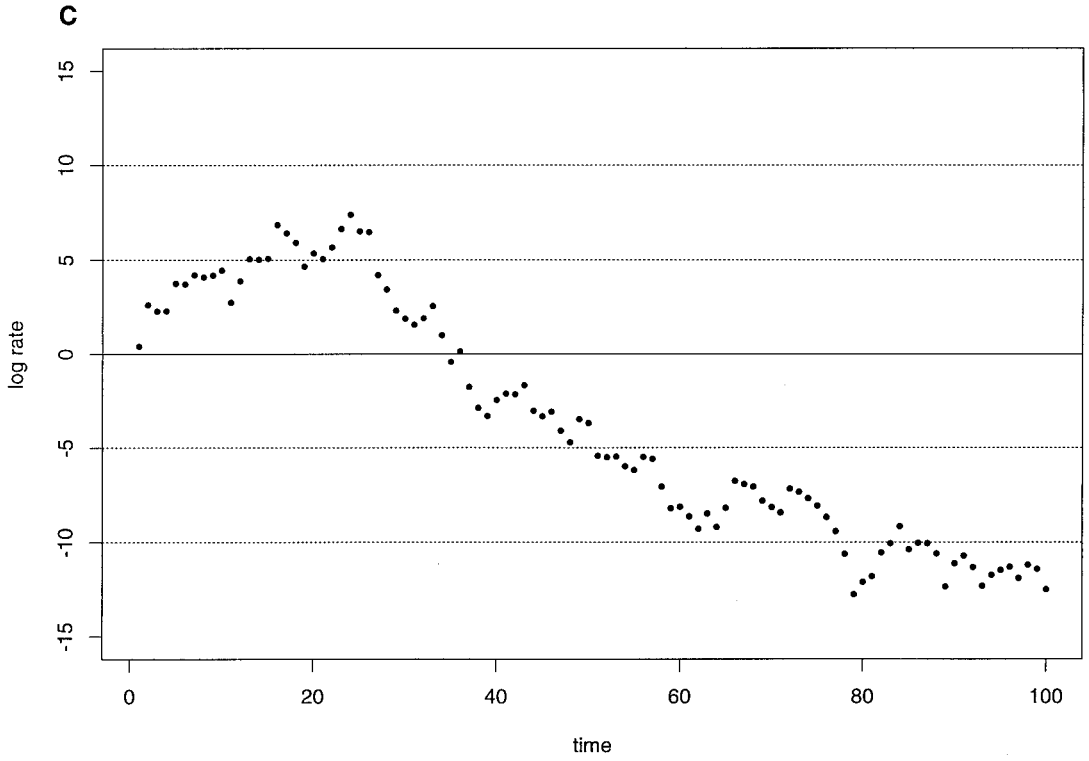

FIG. 4 *(Continued)*

refinements of this approach will prove worthwhile for understanding both evolutionary history and the process of evolution.

Acknowledgments

We benefited from careful comments by the Associate Editor and two reviewers. Jennifer Shoemaker and Bret Larget provided helpful discussions. We thank M. Sanderson for sharing his manuscript with us prior to its publication. I.S.P. and J.L.T. were supported by National Institutes of Health grant P01-GM45344. Software written in the C language that implements the techniques described here is available from J.L.T.

APPENDIX

The Markov Chain Monte Carlo Algorithm

In our implementation, the Metropolis-Hastings algorithm actually cycles through a series of proposal steps. The resulting Markov chain is irreducible and aperiodic. Each individual step involves a proposed state that differs from the current state of the Markov chain by the value of only one or a few parameters.

Proposal Step for ν

In one step of the cycle, a state (T', R', ν') is proposed that differs from the current state (T, R, ν) only by the value of ν. The proposed value ν' is generated by first randomly sampling a value U from a uniform distribution on the (0, 1) interval and then setting

$$\nu' = \nu e^{H_1(U-0.5)},$$

where H_1 is a constant with a prespecified value. With this rule for proposing a new value of ν, equation (5) simplifies to

$$r = \min\left(1, \frac{p(R' \mid T', \nu')p(\nu')\nu'}{p(R \mid T, \nu)p(\nu)\nu}\right). \quad (7)$$

Internal Node Time Proposal Steps

The next part of the cycle for our algorithm involves proposing states for which all but one internal node time is the same for the current and proposed states. In this part of the cycle, a new time for each internal node i is proposed exactly once and then this part of the cycle is exited. The internal node i will have parental node p, eldest-child node e, and youngest-child node y. The rate on a branch will be indexed according to the node at which the branch ends. Therefore, the rate on the branch that ends at node i will be denoted R_i, whereas R_e and R_y, respectively, denote the rates on the branches that end at nodes e and y.

If node i is not the root, its proposed time T_i' should be greater than the time T_p of its parental node p and less than the time T_e of its eldest-child node e. Our current implementation samples T_i' from the uniform distribution on the interval from T_p to T_e. It would be possible to concentrate the distribution of our proposed time T_i' closer to the current time T_i, but we have not yet needed to implement this refinement. So that $p(X \mid B)$ and $p(X \mid B')$ are not involved in our calculations of r, we adjust proposed rates on all branches connected to node i so that all branch lengths are the same for the proposed and current states of the Markov chain. If a

new time for the node currently at time T_i were being proposed, we would set

$$R'_i = \frac{R_i(T_i - T_p)}{(T'_i - T'_p)},$$

$$R'_e = \frac{R_e(T_e - T_i)}{(T'_e - T'_i)},$$

and

$$R'_y = \frac{R_y(T_y - T_i)}{(T'_y - T'_i)}.$$

In order to determine the ratio of the proposal densities (see eq. 5), it helps to reparameterize so that the proposed state has only one parameter with a value different from that of the current state. Without reparameterization, the proposed and current states have different values of R_i, R_y, R_e, and T_i. Reparameterization from a parameter space (R_i, R_y, R_e, T_i) to a parameter space (B_i, B_y, B_e, T_i), where B_i, B_e, and B_y, respectively, represent the lengths of branches that end at nodes i, e, and y, leads to only the value of T_i being different between the proposed and current values of the reparameterized parameter space. Because T'_i is being sampled from a uniform distribution, the proposal densities in the reparameterized space are equal, and their ratio is 1. In the original parameter space, adjustment for the transformation of the random variables (i.e., parameters) means that the proposed state should be accepted with probability

$$r = \min\left(1, \frac{p(R' \mid T', \nu')p(T')(T_i - T_p)(T_e - T_i)(T_y - T_i)}{p(R \mid T, \nu)p(T)(T'_i - T'_p)(T'_e - T'_i)(T'_y - T'_i)}\right). \quad (8)$$

To propose a new time T'_0 for the root node of the ingroup, it must be ensured that the proposed time of this root node is less than the time T_e of its eldest-child node. This is accomplished by setting

$$T'_0 = T_e - (T_e - T_0)e^{H_2(U-0.5)},$$

where U is a uniform random variable on the interval (0, 1) and H_2 is the value of a prespecified constant. By also adjusting proposed rates for branches that are connected to the root such that the current and proposed states of the Markov chain have the same branch lengths, calculations are again simplified. With a reparameterization strategy similar to the one for proposing new times of internal nodes that are not the root, it can be shown that r simplifies to

$$r = \min\left(1, \frac{p(R' \mid T', \nu')p(T')(T_y - T_0)}{p(R \mid T, \nu)p(T)(T'_y - T'_0)}\right), \quad (9)$$

where y is the youngest-child node of the root.

Rate Proposal Steps

For each of the $k + 1$ branches on the ingroup, we have a proposal step for suggesting a new rate R'_i ($i \in \{0, \ldots, k\}$) from the current state R_i. This is accomplished by sampling a value U from a uniform distribution on (0, 1) and using a prespecified constant H_3 to get

$$R'_i = R_i e^{H_3(U-0.5)}.$$

Unlike for the previous proposal steps, this type of step does have proposed branch lengths that differ from the current branch lengths. In this case,

$$r = \min\left(1, \frac{p(X \mid B')p(R' \mid T', \nu')R'_i}{p(X \mid B)p(R \mid T, \nu)R_i}\right). \quad (10)$$

The only part of our Metropolis-Hastings cycle in which the proposed and current states have different branch lengths is the step corresponding to equation (10). Because only one branch length will differ between B and B', the ratio $p(X \mid B')/p(X \mid B)$ simplifies to become a ratio of univariate normal densities. For all other steps of the Metropolis-Hastings cycle, the ratio is 1, because the numerator and denominator of the ratio $p(X \mid B')/p(X \mid B)$ are equal.

"Mixing" Step

To improve convergence of the Markov chain, we have found it worthwhile to add one additional step per cycle. With this step, all proposed node times differ from the current node times by a factor of M. The value of M is obtained by

$$M = e^{H_4(U-0.5)},$$

where U is a uniform random variable on the interval (0, 1), and H_4 is a prespecified constant. For all node times T_i, the proposed state has

$$T'_i = MT_i.$$

The proposed rate for all branches is determined by

$$R'_i = \frac{1}{M}R_i.$$

Under this scheme, the current and proposed states have identical branch lengths. For a reparameterization strategy similar to the one outlined for proposing node times, it can be shown that the probability of accepting the proposed state should be

$$r = \min\left(1, \frac{p(R' \mid T', \nu')p(T')}{p(R \mid T, \nu)p(T)}\frac{1}{M^I}\right), \quad (11)$$

where I is the number of internal nodes (including the root) on the bifurcating tree.

LITERATURE CITED

ALLISON, L., and C. S. WALLACE. 1994. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. J. Mol. Evol. **39**:418–430.

CARLIN, B. P., and T. A. LOUIS. 1996. Bayes and empirical Bayes methods for data analysis. Chapman and Hall, London.

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

GAUT, B. S., S. V. MUSE, W. D. CLARK, and M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J. Mol. Evol. **35**:292–303.

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

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.

GOLDMAN, N., J. L. THORNE, and D. T. JONES. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445–458.

HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**:97–109.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–32 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KARLIN, S., and H. M. TAYLOR. 1975. A first course in stochastic processes. 2nd edition. Academic Press, San Diego.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

LANGLEY, C. H., and W. M. FITCH. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. **3**:161–177.

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

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. **21**: 1087–1092.

OHTA, T. 1987. Very slightly deleterious mutations and the molecular clock. J. Mol. Evol. **26**:1–6

PENNY, D., R. P. MURRAY-MCINTOSH, and M. D. HENDY. 1998. Estimating times of divergence with a change of rate: the orangutan/African ape divergence. Mol. Biol. Evol. **15**:608–610.

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

SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. **14**:1218–1232.

STUART, A., and J. K. ORD. 1991. Kendall's advanced theory of statistics. Vol. 2, 5th edition. Oxford University Press, New York.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

YANG, Z. 1997. Phylogenetic analysis by maximum likelihood (PAML). Version 1.3. University of California, Berkeley.

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

ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York.

STANLEY A. SAWYER, reviewing editor

Accepted September 2, 1998