# Model Averaging and Bayes Factor Calculation of Relaxed Molecular Clocks in Bayesian Phylogenetics

Wai Lok Sibon Li[1,2,3,4] and Alexei J. Drummond*[,1,2,3,5]

[1]Computational Evolution Group, University of Auckland, Auckland, New Zealand

[2]Bioinformatics Institute, University of Auckland, Auckland, New Zealand

[3]Department of Computer Science, University of Auckland, Auckland, New Zealand

[4]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles

[5]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

***Corresponding author:** E-mail: alexei@cs.auckland.ac.nz.

**Associate editor:** Rasmus Nielsen

## Abstract

We describe a procedure for model averaging of relaxed molecular clock models in Bayesian phylogenetics. Our approach allows us to model the distribution of rates of substitution across branches, averaged over a set of models, rather than conditioned on a single model. We implement this procedure and test it on simulated data to show that our method can accurately recover the true underlying distribution of rates. We applied the method to a set of alignments taken from a data set of 12 mammalian species and uncovered evidence that lognormally distributed rates better describe this data set than do exponentially distributed rates. Additionally, our implementation of model averaging permits accurate calculation of the Bayes factor(s) between two or more relaxed molecular clock models. Finally, we introduce a new computational approach for sampling rates of substitution across branches that improves the convergence of our Markov chain Monte Carlo algorithms in this context. Our methods are implemented under the BEAST 1.6 software package, available at http://beast-mcmc.googlecode.com.

**Key words:** Bayesian phylogenetics, model selection, Bayes factor, BEAST, model averaging, relaxed molecular clock.

## Introduction

For many years, the phylogenetic community took the notion of branch lengths on a tree as a representation of distance between the species in units of substitutions per site. However, a more biologically relevant way to consider the branch lengths is to treat the distances as the product of divergence time from the common ancestor and the rate at which the substitutions occurred. Partitioning the genetic distances into divergence times and rates allows one to reconstruct the temporal aspect of evolutionary history and dissect the processes involved (Martin and Palumbi 1993; Adachi and Hasegawa 1995; Gu 1998; Glazko and Nei 2003; Bunce et al. 2009). The classical approach for rate/divergence time estimation is to force the rates to conform to a "molecular clock," which assumes that the rates are equal across all branches on a tree (Zuckerkandl and Pauling 1965). In reality, rates of substitution differ across species as a result of variation in mutation rates, metabolic rates, generation times, population size and structure, and selection, among other things. Strict molecular clocks are therefore generally confined to analyses within a species or among a few closely related species.

As a result, the idea of relaxed molecular clocks has been developed, where the rates of substitution are permitted to vary across branches of the tree (Sanderson 1997; Rambaut and Bromham 1998; Thorne et al. 1998). Relaxed molecular clock models have in recent years been accepted into the broader field of phylogenetics. This is mainly due to the biological relevance of these models as it is well established that rates of substitution naturally vary across species and lineages (Wu and Li 1985; Britten 1986; Gaut et al. 1992). It has also been demonstrated that the use of relaxed molecular clock models can, in some circumstances, improve the accuracy of phylogenetic estimation (Drummond et al. 2006). Hence, relaxed phylogenetic methods are not only expected to improve estimation of divergence times but even the accuracy of estimated tree topologies.

Like any other statistical modeling technique, relaxed molecular clock methods suffer from problems of model misspecification and uncertainty. Model misspecification is a deep-rooted problem that plagues a range of applications of mathematics across the sciences and can cause errors and bias in the resulting analysis. In Bayesian phylogenetics, as with any statistical inference task, a sensible balance between practicality and parameter richness is required. A good model is not necessarily the most parameter rich but instead is a model that captures the essential features of the hypothesis being tested without introducing unnecessary error, bias, and overfitting. In a complex process such as molecular evolution, the model will always be misspecified in the sense that all evolutionary models are severe simplifications of

reality. Our aim therefore is to choose a model or a set of models that are 1) able to test the hypothesis and 2) are most plausible given the data at hand. There are two general approaches to evaluating data in light of alternative models: model averaging and model selection. Model averaging allows the data to be evaluated by a weighted average over a set of models. The benefit of model averaging is that uncertainty in models can be incorporated into the analysis. Also, in some cases where multiple models appear to fit the data well, inferences from these models can be averaged over. In the case of a nested family of models, model averaging can also be used to investigate the importance of different parameters in explaining the data. Results inferred by model averaging are not based on or biased toward a single model, but rather the data itself determines which model or set of models are most probable.

In Bayesian statistics, a common approach to model averaging is stochastic sampling of the model space with Markov chain Monte Carlo (MCMC) (Godsill 2001). Reversible jump MCMC (rjMCMC; Clyde 1999; Hoeting et al. 1999) is well known and commonly employed but is just one of a class of "composite model" formulations of model averaging within MCMC (Godsill 2001), which allows the MCMC to jump between spaces of varying dimensions (Green 1995). However, reversible jump can be difficult to implement in some contexts. Another composite model formulation is Bayesian stochastic search variable selection. Although it is less computationally efficient than rjMCMC, it is easier to implement and has already found several applications in Bayesian phylogenetics (Gray et al. 2009; Lemey et al. 2009; Wu and Drummond 2011).

Arguably, the most appropriate technique for selecting between two models in a Bayesian setting is the calculation of the Bayes factor (BF) (Kass and Raftery 1995). BFs can be used to evaluate evidence for one model over another. Though certain heuristics have been proposed (Newton and Raftery 1994), accurate calculation of BFs is most easily achieved by the same computational techniques as used for model averaging in an MCMC framework.

In Bayesian phylogenetics, model selection has been previously implemented for substitution models (Huelsenbeck et al. 2004; Gowri-Shankar and Rattray 2007), the rate of nucleotide change (Suchard et al. 2001), and site heterotachy (Pagel and Meade 2008). The application of model averaging and model selection to relaxed molecular clock models has not yet been examined (but see, Drummond and Suchard 2010). In recent years, the standard approach has been to approximate the BF with estimated marginal likelihoods obtained from two independent MCMC analyses of the same data using different modeling assumptions. Software packages such as BEAST (Drummond and Rambaut 2007) provide a posterior sample of the likelihood, which can be used to estimate the marginal likelihood through computing the harmonic mean of the posterior sample. This approach can be interpreted as an approximation of the marginal likelihood using importance sampling, where the posterior distribution is the importance distribution and the prior distribution is the target distribution (Newton and Raftery 1994). However, this approximation is known to often provide extremely poor estimates of the marginal likelihood (Beerli and Palczewski 2010) as the posterior distribution is often not a good importance distribution for the prior distribution. This is especially the case when there is a lot of data because then the posterior typically has much smaller variance than the prior.

In this paper, we outline a strategy for model averaging of relaxed molecular clock models under phylogenetic inference with Bayesian MCMC. Consequently, such model averaging allows for accurate calculation of BFs for model selection. Instead of rjMCMC, our method employs a simple composite model formulation (Godsill 2001). We show that our method can improve phylogenetic estimation compared to using a single model when the underlying distribution of rates is unknown. We also demonstrate that by using our method, we can accurately estimate the BFs needed to perform model selection.

Finally, we propose a new algorithm for sampling the rate values on the distribution in the MCMC and also investigate an alternative distribution for rates across branches on a tree.

## Materials and Methods

### Sampling Rates as Quantiles

We first outline a procedure to sample rates of substitution in an MCMC by representing the rate on each branch by its corresponding probability in the cumulative probability distribution of the branch-rate distribution model. Under phylogenetic analysis with relaxed clock models, each branch on the tree is assigned a separate rate. In a Bayesian framework, each of these rate parameters is sampled in the MCMC. The conventional approach to sample rates of substitution is to draw rates from a distribution on a continuous scale (Rannala and Yang 2007). However, when the parameters of the rate distribution are also treated as random variables (i.e., mean and standard deviation [SD] parameters of the log-normal distribution [LN]), then the standard parameterization can be difficult to produce efficient proposal kernels for because of the strong correlation between the rate values on individual branches and the parent distribution parameters.

We propose a more computationally convenient strategy: sample values of $q \in (0, 1)$ rather than the actual rates, $r$, on individual branches. $q$ can then be interpreted as a rate using the inverse cumulative distribution function (iCDF) of the branch-rate distribution. The rate of the $j$th branch, $r_j$, can be defined in terms of its corresponding probability in the CDF ($q_j$):

$$r_j = F_\omega^{-1}(q_j),$$

where $F$ is the CDF for the relaxed clock model with parameters $\omega$. Each value of $\mathbf{q} = \{q_1, q_2, \ldots, q_{2n-2}\}$ (where $n$ is the number of taxa) can be estimated by MCMC. In comparison with directly sampling the rates, sampling the $\mathbf{q}$ values allows the MCMC to independently sample the parent parameters $\omega$ and the rate parameters $\mathbf{q}$ separately while still getting excellent convergence of the Markov chain.

Additionally, since the quantile function parameterization automatically draws rates from the prior defined by the parent distribution, no per-branch terms associated with the rate model need to be added to the calculation of the prior density, so long as the prior on each element of **q** is unit uniform.

## Model Averaging

In the previous section, we provided a means of sampling the rate of substitution on each branch as cumulative probability values. Effectively, a value of $q$ describes the rate of a branch relative to the other branches. As a result, it is possible to change the underlying distribution of the rates without altering the ordering of the rates and without changing the probability of the rates given the rate distribution. Given a set of values **q**, rate values can be obtained for any parametric distribution for which the iCDF can be easily computed.

Using MCMC, we can sample the underlying distribution itself. The sampling mechanism is based on the "standard model selection" parameterization of the composite model space *sensu* Godsill (2001) as follows. We define an indicator variable, $i \in \{1, 2, \ldots, N\}$, where $N$ is the total number of branch-rate distributions to be considered. Each value of $i$ refers to a different branch-rate distribution, $F_i$ with a set of associated distribution parameters, $\omega_i$ which models the underlying distribution of rates across all branches. Accordingly, $i$ along with all the distribution parameters $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ are sampled in the MCMC.

The probability of the sequence data can be computed given the tree, **q**, and $F_i$. For instance, for a model-averaged relaxed molecular clock model where $F_1$ and $F_2$ specify the lognormal and exponential distributions, respectively, the probability of rate $r$ on a branch $P(r|\Omega)$ is given as

$$P(r|\Omega) = \begin{cases} \frac{1}{r\sqrt{2\pi \ln\left(1+\frac{\sigma^2_{\mathrm{LN}}}{\mu^2_{\mathrm{LN}}}\right)}} \exp\left[\frac{-\left(\ln(r/\mu_{\mathrm{LN}}) - \frac{1}{2}\ln\left(1+\frac{\sigma^2_{\mathrm{LN}}}{\mu^2_{\mathrm{LN}}}\right)\right)^2}{2\ln\left(1+\frac{\sigma^2_{\mathrm{LN}}}{\mu^2_{\mathrm{LN}}}\right)}\right] & \text{if } i = 1 \\ \frac{\exp\left[\frac{-r}{\mu_E}\right]}{\mu_E} & \text{if } i = 2 \end{cases},$$

where $\Omega = \{[\mu_{\mathrm{LN}}, \sigma_{\mathrm{LN}}], [\mu_E]\}$. Consider a Markov chain at time $t$, with a state of $\theta_t = \{i=j, \mathbf{q}, \Omega, g\}$. If a new branch-rate distribution is proposed, $F_k$ (i.e., $\theta'_t = \{i=k, \mathbf{q}, \Omega, g\}$), it will be accepted, with probability

$$\alpha = \min\left(1, \frac{\Pr(D|F_k^{-1}(\mathbf{q}), g, \Omega)p(i=k)}{\Pr(D|F_j^{-1}(\mathbf{q}), g, \Omega)p(i=j)}\right),$$

where $p(i = j)$ and $p(i = k)$ are the prior probabilities of models $j$ and $k$, respectively. Consequently, the proportion of samples that have a particular value of the indicator variable $i$ will estimate the posterior probability of the corresponding branch-rate model. Also, the resulting posterior distribution of trees will be a model-averaged posterior distribution, weighted by the probabilities of the branch-rate models considered.

In our implementation of the method, we use a uniform prior on $i$ (i.e., $p(i = j) = p(i = k) = 1/N$), where we assume that there is no prior knowledge as to which model is preferred; thus, the prior probability of each model is equal, and the ratio of the posterior probabilities of the two models is equivalent to the BF.

## Inverse Gaussian Distribution Model

Besides examining model averaging, the use of the inverse Gaussian (IG) distribution as a model for the distribution of rates across branches is investigated. The suggestion of using alternative distributions stems from Kitazoe et al. (2007), who found that alternative models of rate distribution were more suitable for modeling the rate heterogeneity across branches in mammalian mitochondrial proteins. Overall, the probability distribution function of the IG is similar to the lognormal when the coefficient of variation is low (less than 1). By evaluating the skew and kurtosis of the density function of the IG, we found that when the coefficient of variation is high (i.e., $\sigma$ is relatively large compared with $\mu$), the LN has a much sharper and less symmetric distribution (data not shown). The density function of the IG therefore has a longer tail, meaning that decrease in the upper tail is slower, relative to other similar distributions. The IG distribution is therefore more liberal in allowing for relatively faster rates of evolution within the tree. This property may be suitable for data sets where there are "rogue taxa" with exceptionally fast rates. An example of such data sets is mammalian data where the rodent lineages have accelerated rates of substitution (Wu and Li 1985; Britten 1986; Martin and Palumbi 1993; Li et al. 1996).

## Algorithm Implementation

The models and relaxed clock implementations were written in Java 1.5 and are part of the BEAST (Drummond and Rambaut 2007) software package.

## Relaxed Clock Model Priors

For our analysis, we compare three distributions that can be used to model the variation in rate of substitution across branches: the LN, the exponential distribution ($E$), and the IG distribution. $E$ and LN were already implemented in BEAST (Drummond and Rambaut 2007) and are commonly used to model the rates across branches.

The shape of the IG distribution is determined by two parameters: the mean, $\mu$, and the shape parameter, $\lambda$. It should be noted that under our implementation, IG was parameterized with the SD parameter $\sigma$ rather than with its standard distribution parameter $\lambda$. The motivation lies within the relationship between $\sigma$ and $\lambda$ in the IG, which is

$$\sigma = \left(\frac{\mu^3}{\lambda}\right)^{\frac{1}{2}}$$

As $\sigma$ and $\lambda$ are inversely related to the IG, the hyperprior for the MCMC that is naturally imposed on the distribution parameter, if it was parameterized with $\lambda$, will be an inverse of the natural hyperprior on $\sigma$ in LN. Therefore, sampling $\sigma$ in IG was performed to improve consistency of priors across the models compared.

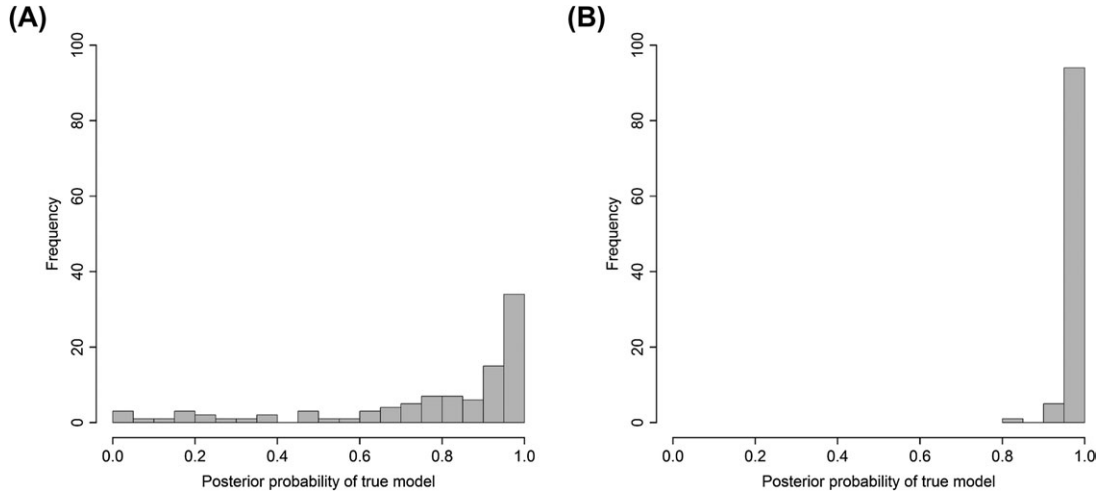


FIG. 1. Histograms showing the distribution of posterior probabilities from using our model-averaged model on the simulated data set. (*A*) The posterior probabilities of the *E* distribution when the rates were simulated under $D_E$. (*B*) The posterior probabilities of the LN distribution when the rates were simulated under $D_{LN}$.

### Other MCMC Priors

A Yule pure birth process was used as a prior on the speciation process (Yule 1924). For analyses that used the Hasegawa-Kishino-Yano (HKY) nucleotide substitution model (Hasegawa et al. 1985), a $1/x$ prior was placed on the transition–transversion parameter, $\kappa$. In analyses where the general time-reversible (GTR) model (Tavaré 1986) was used, a $1/x$ prior was placed on the relative rate parameters.

A list of the proposal kernels used in the MCMC process is listed in the supplementary material (Supplementary Material online).

## Results

### Application of Model Averaging

#### *Simulated Data*

We decided to benchmark our model-averaged MCMC in terms of how well it could recover the true underlying distribution of the rates. We used a balanced tree of 32 taxa plus an outgroup to simulate sequence alignments. The divergence times on each branch were all set to 5 time units, except the outgroup branch which had a length of 30 to make the tree ultrametric. For each of the branches on a tree, we assigned a rate of substitution drawn from either an exponential distribution with a mean of 0.005 ($D_E$) or a LN with a mean of 0.005 and variance = 0.004 ($S^2$ parameter of 0.5) ($D_{LN}$). The rates assigned to the branches on the simulated trees are uncorrelated rates, so that for each branch, the rate is drawn independently from the distribution, rather than autocorrelated rates, where the rate of a branch is dependent on the rate of its parent branch and divergence time from the parent. One hundred realizations of rates were simulated under each of the two models $D_E$ and $D_{LN}$. Alignments of 1,000 nts in length were generated from each of the 200 trees using Seq-gen (Rambaut and Grassly 1997), under a Hasegawa-Kishino-Yano (1985) nucleotide substitution model with gamma-distributed rate heterogeneity across sites (Yang 1994) (HKY + $\Gamma$ model) with a transition–transversion ratio of 3.0 and shape parameter of 0.5.

Each alignment was analyzed with BEAST using a model average of LN and *E* ($M_{LN,E}$). The mean for both distributions was fixed at 0.005, but the SD parameter of LN was estimated by the MCMC. The tree topologies were constrained to the true tree topology but divergence times were estimated. A HKY + $\Gamma$ nucleotide site model was used in the analysis, with model parameters estimated. The analyses were each run for 50,000,000 steps with the initial 5,000,000 steps discarded as the burn-in. The convergence of the chains was verified by checking that all effective sample sizes (ESS) were greater than 100.

The results of this analysis demonstrate that our model-averaging method generally yields a higher posterior probability for the true underlying model that the rates were drawn from figure 1). For the 100 alignments with rates drawn from $D_E$, 83 had a higher posterior probability for *E* than for LN and the mean posterior probability of the *E* model was 0.77. For the $D_{LN}$ alignments, all 100 of the runs had a higher posterior probability for LN and the mean posterior for the LN model was 0.999. Of the 100 runs where the rates of substitution were drawn from $D_E$, 97 of the analyses contained the true distribution in the 95% credible set of models. For alignments with rates of substitution drawn from $D_{LN}$, all 100 of the runs contained the true distribution in the 95% credible set. In this specific setup, it appears that the model-averaging technique is better able to predict the underlying rate distribution when the true distribution is LN than it is when the true distribution is *E*.

The BF for each of the runs was calculated and used to interpret the support, under the criteria outlined in Kass and Raftery (1995), for the true underlying branch-rate model of the data (supplementary material fig. S1*A*, Supplementary Material online). In most cases, there is support

for the true underlying branch-rate distribution as simulated in the data. For $D_E$, 67% showed positive support for the data being *E* distributed, whereas only 17% showed some degree of negative support. For $D_{LN}$, all the analyses had strong or more support for the LN model, 71% of which showed very strong evidence for the data being LN distributed (supplementary material fig. S1*B*, Supplementary Material online).

These results indicate that our model-averaging method is capable of retrieving the true underlying distribution of the rates of substitution. In particular, the statistical power of this method is demonstrated by the fact that both sample distributions $D_E$ and $D_{LN}$ have comparable variances (0.005 and 0.004, respectively). Hence, even when we are drawing rates from two fairly similar distributions, our method is able to differentiate between them.

We compared our model-averaging technique with the prevailing method in the literature of computing the ratio of the harmonic mean estimator of marginal likelihoods. The rates on the trees obtained by model averaging and model selection were compared with the simulated rates on the true trees. For the estimation using marginal likelihood, two schemes were used: 1) the posterior rate estimate, averaged between the rates of *E* and LN, and 2) the posterior point estimates of rate from the model (either *E* or LN) that is supported by approximate BF. The root mean squared (RMS) deviation of the estimated rates from the true rates was calculated.

For $D_{LN}$, the mean RMS values for the data set were 0.00144, 0.00146, and 0.00147, respectively, for 1) $M_{LN,E}$, 2) model averaging with harmonic mean estimator of marginal likelihoods, and 3) model selection with harmonic mean estimator of marginal likelihoods. These differences were statistically significant for all pairwise comparisons ($P$ value $< 0.0001$; nonparametric *t*-tests). For $D_E$, a much smaller difference was seen between the three methods, with mean RMS values of 0.002130, 0.002133, and 0.002130; a statistically significant difference was observed only between (2) and (3) above ($P$ value $< 0.01$). The results demonstrate that our single-chain method of model averaging more accurately recovers the true tree than estimation using marginal likelihood approximations and performs equally or better than any multichain model selection process available in BEAST.

### *Mammalian Data*

We used the OrthoMaM data set v4.0 (Ranwez et al. 2007), which contains alignments of orthologous genomic markers shared between placental mammals. All coding sequences markers that shared orthology across the 25 species in OrthoMaM were obtained for a total of 1,056 alignments. For this analysis, we wanted to obtain a set of alignments where the true species topology between the sequences is well established and uncontroversial. We trimmed taxa from the 1,056 alignments to contain 12 mammalian species: *Canis familiaris*, *Felis catus*, *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus abelii*, *Macaca mulatta*, *Microcebus murinus*, *Otolemur garnettii*, *Mus musculus*, *Rattus norvegicus*, *Ochotona princeps*, and *Oryctolagus cuniculus*. The phylogeny between these species is well supported by the literature (Yoder 1997; Novacek 2001; Reyes et al. 2004; Bashir et al. 2005; Steiper and Young 2006; Prasad et al. 2008) and is provided in the supplementary material fig. S2 (Supplementary Material online).

**Table 1.** A Summary of Rate Distribution Priors Used in the Relaxed Clock Models of the Mammalian Data Set.

| Distribution | Parameter | Prior Boundaries of Parameter |
|---|---|---|
| Exponential (*E*) | Rate, $\lambda$ | 1.0 ($\sigma$ = 1.0) |
| Lognormal (LN) | SD, $\sigma$ | 0.0–10.0 |
| IG | SD, $\sigma$ | 0.0–10.0 |

The data set was analyzed using MCMC runs with a model average of the LN and *E* distributions ($M_{LN,E}$). We assumed a GTR model of nucleotide substitution (Tavaré 1986) on the data with gamma-distributed heterogeneity across sites (Yang 1994) (GTR + $\Gamma$). For this analysis, the tree topology was not constrained. For each of the 1,056 sequences, an MCMC algorithm was run for a chain length of 50 million steps, with the first 10% of the run treated as burn-in and discarded. If all parameters of interest had not converged after the run, the chain was rerun for longer until the ESSs were above 100. However, for some of the genes, we were unable to obtain convergence in the MCMC chains even after 200 million steps. With the given data, models, and proposal kernels, the MCMC runs could not be made to converge within the practical running time. In cases where this occurred in one or more of the models being compared, the genes were excluded from the analysis. We acknowledge this as a potential source of bias, though this occurred for only 54/1,056 alignments (5.1% of the $M_{LN,E}$ analyses).

As no time calibration data were available, a mean rate, $\mu$, of 1.0 was used for all distributions. Hence, we examined the relative rates of the branches across a tree rather than the absolute rates of substitution. The priors used for each of the branch-rate distributions are shown in table 1.

One question we wanted to answer was whether the use of model averaging improved the quality of phylogenetic estimation. As there is no known accurate time calibration data available for this data set, it would be difficult to benchmark our method in terms of rate estimation. However, the ability of the phylogenetic analysis to recover the true tree topology of the taxa is also a good measure of model performance.

We compared our analysis with $M_{LN,E}$ on the mammalian data set to the same analysis with three other settings: using a relaxed clock model with only a LN ($M_{LN}$), an exponential distribution ($M_E$), and with a strict molecular clock model ($M_{SC}$). Table 2 shows a summary of the statistics of the analysis using each of the models. Results indicated that $M_{LN,E}$ and $M_{LN}$ estimated the true tree topology significantly more often than $M_E$ ($P$ value $<$ 0.0001; nonparametric *t*-tests). The fact that the results of any of the three relaxed molecular clock models only captured the true tree on average ≈85% of the time in the 95% credible set suggests that there is some degree

**Table 2.** Statistics Related to the Estimation of Topology and Rate for the Mammalian Data Set.

| | $M_{LN,E}$ | $M_{LN}$ | $M_E$ | $M_{SC}$ |
|---|---|---|---|---|
| **Proportion of times true tree is point estimate** | **0.520** | **0.533** | **0.475** | **0.530** |
| **Average posterior probability of true tree** | **0.417** | **0.427** | **0.372** | **0.498** |
| **Average number of unique trees in 95% credible set** | **102.331** | **93.303** | **118.404** | **6.772** |
| **Proportion of times true tree appears in 95% credible set** | **0.850** | **0.852** | **0.847** | **0.729** |

NOTE.—Nine hundred and fifty-three genes were used for this analysis.

of model misspecification, though the model misspecification is not only limited to the molecular clock but also contributed by misspecification of other aspects of the evolutionary model. We would not expect the evolutionary models used to capture all characteristics of the evolutionary process. Model averaging does not improve the estimation of tree topology in this data set, when compared with the better of the two models ($M_{LN}$), but does significantly improve performance of the point estimate when compared with the $M_E$ model. This suggests that model averaging can protect against poor inference when the correct model is not known. Although $M_{SC}$ often chooses the correct true tree as the point estimate, it fails to contain the true tree within the 95% credible set significantly more often ($P$ value $< 0.0001$). When $M_{SC}$ contains the true tree in the 95% credible set, the corresponding mean estimate of the $\sigma$ parameter in $M_{LN}$ is 0.869, whereas its mean is 2.080 in the remainder. These two means are significantly different ($P < 0.0001$; using a nonparametric test). This shows that the $M_{SC}$ model is not robust to data that is not clock like but probably performs well on the large fraction of relatively clock-like alignments in this data set. The data analyzed here contained only a small number of taxa, so further empirical studies are needed to confirm these results and such studies should include specification of more than two branch-rate distributions.

We further examined the biological relevance of different parametric distributions as models of rate heterogeneity in real data. We observed the relative posterior probabilities of each of the two distributions, LN and $E$, across the analyses of the mammalian genes. From figure 2A, we can see that in a majority of the genes LN is preferred over $E$. The mean posterior probability of the LN model was 0.701 (hence mean posterior of $E = 0.299$). In 150 of the 1002 genes we compared, the $M_{LN,E}$ model had a posterior probability of over 0.95 for one of the models; 149 of these genes showed strong preference for the LN model, whereas only one showed strong preference for the $E$ model. Our results suggest that on average, the LN better models the rates of substitution across branches this in mammalian data set.

Figure 3 shows three gene trees that model averaging shows, respectively, 1) strong support for rates that are lognormally distributed, 2) exponentially distributed, and 3) no support for either hypotheses. From the trees, it can be seen that the rates vary substantially between the trees that are supported by each model, thus justifying the need of different models, even when the same set of taxa are considered. Upon inspection,

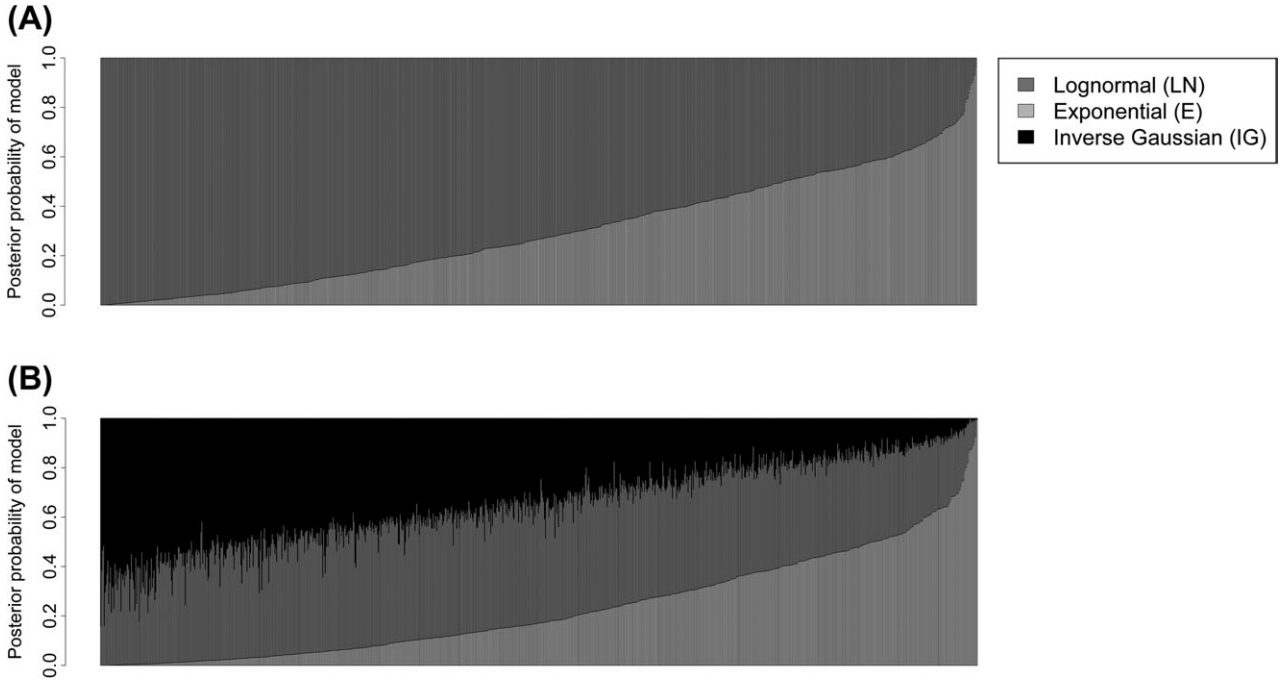


FIG. 2. Bar plots showing the distribution of posterior probabilities of each distribution when applying model-averaged models (*A*) $M_{LN,E}$ ($n = 1002$) and (*B*) $M_{LN,IG,E}$ ($n = 1008$) on the mammalian data set. Data are sorted by the posterior probability of $E$.

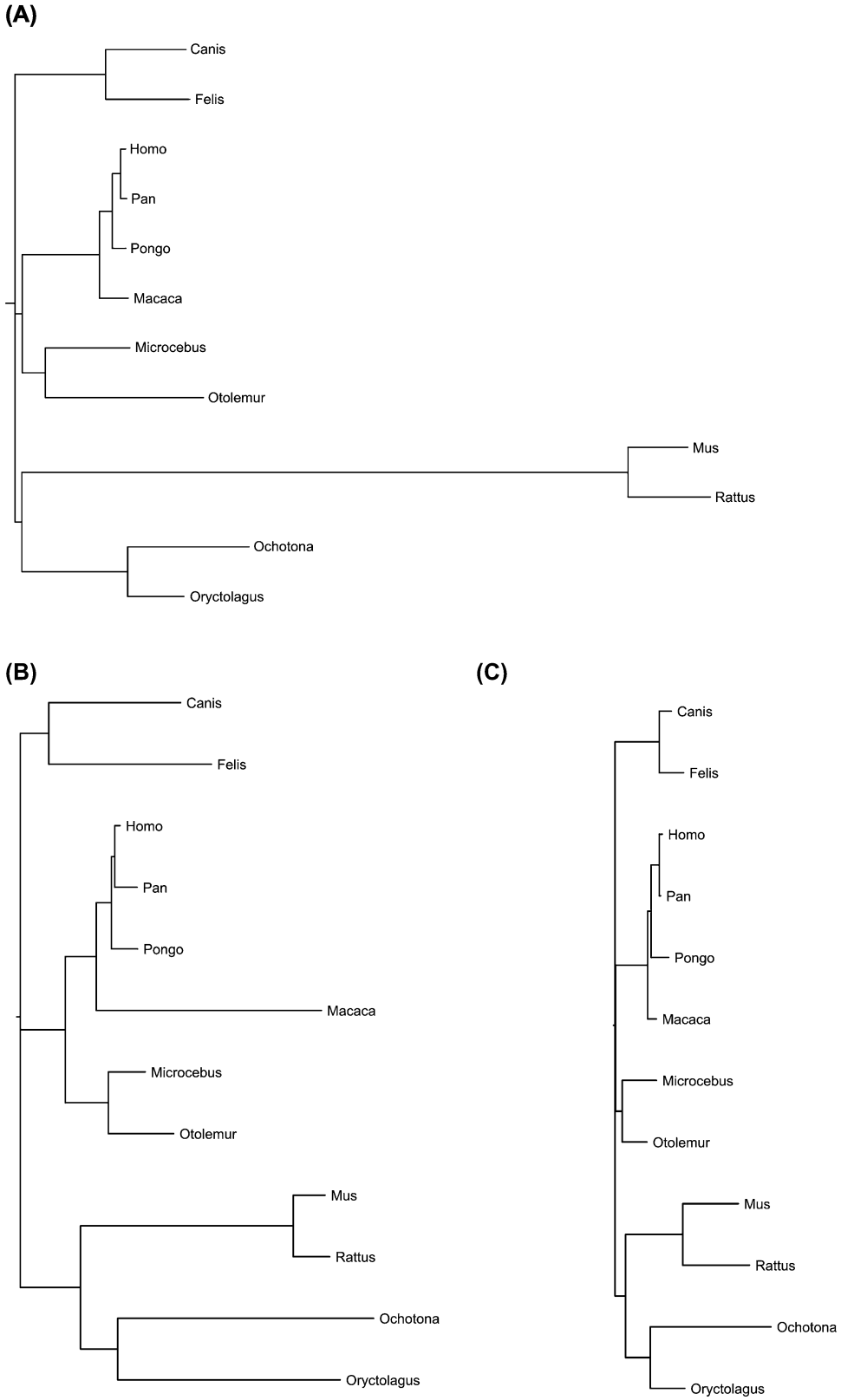


FIG. 3. Point estimates of example gene trees from the $M_{LN,E}$ analysis, showing the difference in trees between data which supports (*A*) the LN model, (*B*) the *E* model, and (*C*) neither model.

the top 15 trees that showed the highest support for *E* were unable to capture the true tree topology with the point estimate and thus could not be used in this comparison; this is in contrast to LN, where only one of the top 15 trees supporting the model chose an incorrect point estimate. This could be because selection of the

*E* model coincides with genes that do not have good phylogenetic signal or because the rate variation in genes where the *E* model is preferred is still poorly modeled by the exponential distribution, leading to poor phylogenetic inference.

We then expanded on this analysis to a comparison of more than two models. We ran the model-averaging analyses on the mammalian data again, except this time including the IG distribution, along with the LN and *E* ($M_{LN,IG,E}$) (fig. 2*B*). The mean posterior probabilities of the LN, IG, and *E* models were 0.421, 0.347, and 0.231, respectively, indicating that there was a preference of the LN model over the other two models. The exponential model had a posterior probability >0.95 in one of the 1,008 genes. On the other hand, in 237 of the genes, *E* was not contained in the 95% credible set. However, there were no genes that contained only LN or only IG in the 95% credible set, but rather when the support for the *E* distribution was low, the 95% credible set contained a combination of posterior samples supporting either LN or IG. This suggests that the characteristics of LN and IG are more similar to each other than they are to *E*; hence, the model averaging could not distinguish between them.

## Indirect Assessment of BF Computations

A simple consistency check of BF computations involves computing the BF between two models, with or without a third model in the mixture. If the implementation is correct and the MCMC is run sufficiently long, then the BF computed for a pair of models should be the same whether a third model is present in the set of models to average. We used this fact to test the consistency of our BF estimates. For the $M_{LN,E}$ and $M_{LN,IG,E}$ models that we ran, the log ratio of the posterior probabilities between LN and *E*: $P(M = \text{LN} \mid D)/P(M = E \mid D)$ was calculated. Because our priors were uniform, the ratio of the posteriors is equivalent to the BF (Kass and Raftery 1995). This comparison was carried out as it is known that IG has similarities to LN (Takagi et al. 1997), and thus, the covariance between them is likely to be lower than each of their covariances to *E*. Comparisons between the BF for each gene are shown in figure 4. Even with the introduction of a third distribution in the $M_{LN,IG,E}$ model, the BFs between the LN and *E* distributions were mostly very similar. The coefficient of determination between the log BFs of the two models was 0.932, indicating a very strong correlation (*P* value < 0.001). It is clear from manual inspection of the plot that variability in the estimate of the BF increases with the log BF. This is to be expected as log BF $< -3$ or log BF $> 3$ represents strong support for one model over the other, meaning that the less probable model is seldom sampled, and the relative error of the estimate will be greater.

## BF Calculation

We then compared the values of the BFs as computed with our model averaging against the approximated value of the BF as defined by Newton and Raftery (1994). For simplicity, the two methods of BF calculation were compared with the

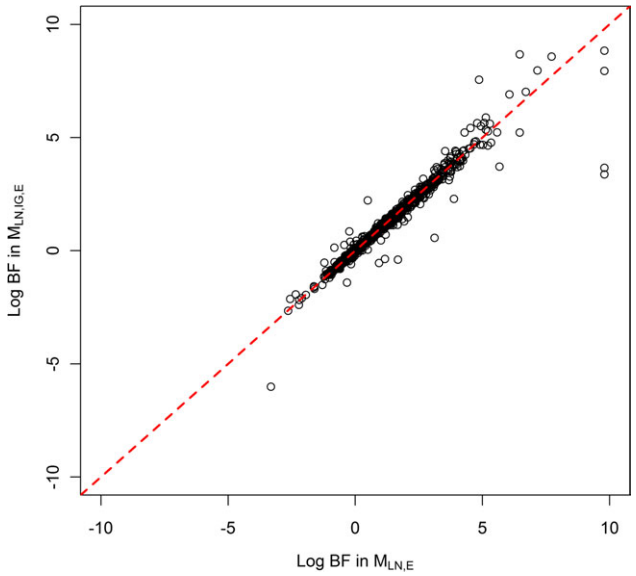


**Fig. 4.** Scatter plot of the log BFs of $F_{LN}$ against $F_E$ in the $M_{LN,E}$ model against the same value in the $M_{LN,IG,E}$ model ($n$= 883). Where the posterior probability of a model was 0, we assigned a probability of 0.5/9001 (≈0.00005; 9001 is the total number of sampled trees), which is used as a minimum value.

$M_{LN,E}$ model. As Newton and Raftery's method is an approximation, we can assume any significant differences to our values to be due to sampling error in the importance sampling method. As can be seen in figure 5, the values calculated via approximation appeared to be relatively conservative, whereas the BFs calculated by model averaging tend to have much larger variation than estimated values (SDs of 1.890 and 0.882, respectively). We removed the presence of outliers by computing leverage coefficients and removing values that were considered significant (Hoaglin and Welsch 1978; Sokal and Rohlf 1995). The coefficient of

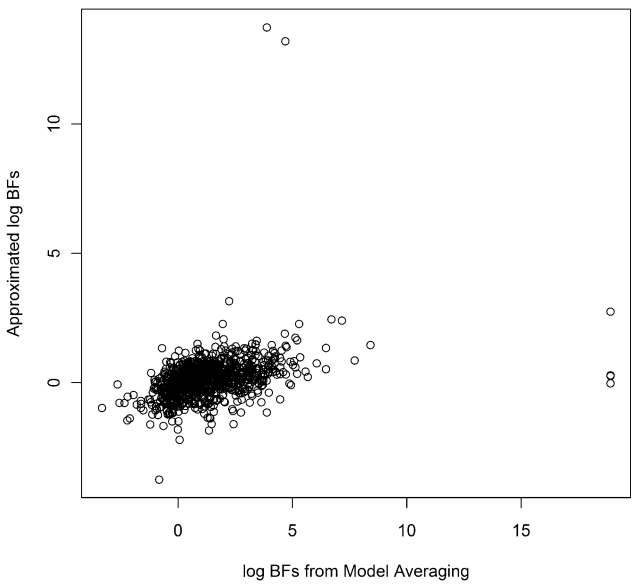


**Fig. 5.** Scatter plot of BF values for the mammalian data calculated by using model averaging versus using an approximation with importance sampling ($n$=961).

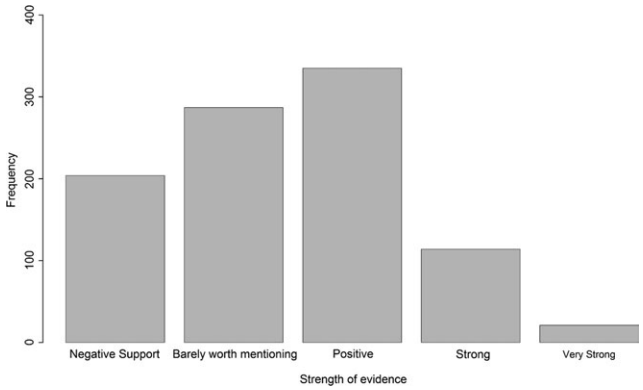


FIG. 6. Interpretation of the BF for support of the $F_{LN}$ model in the mammalian data set ($n=961$). The support is categorized as follows: negative support (BF<1), barely worth mentioning (1<BF<3), positive (3<BF<20), strong (20<BF<150), and very strong (BF>150).

determination calculated from the outlier-omitted data was 0.133, indicating a lack of correlation between the logged values of the two BF computations. This suggests that the calculation of BFs with importance sampling provides inaccurate estimates that do not reflect the actual BF, as has been previously shown in other Bayesian phylogenetics contexts (Beerli and Palczewski 2010).

The supports of the BFs for the lognormal model across the genes in terms of hypothesis testing are shown in figure 6. Seven hundred and fifty-seven of the 961 genes compared indicated support for LN, 470 of which showed positive or more support and 21 of which had very strong evidence against the data being exponentially distributed. For support of the E model, 204 genes showed some level of support, 21 of which showed positive or more support for the model, whereas none of these genes provided very strong evidence against the data being lognormally distributed. What these interpretations essentially show is a quantification of the posterior probabilities that were calculated, which can be used as an assessment for model selection. This demonstrates that calculation of BFs can be used for model choice based on established statistical frameworks for hypothesis testing.

This analysis was extended to a pairwise comparison of LN and IG to distinguish between the two more similar models. In 98 of the 1,008 genes compared, positive support was present for the LN model. Although LN had a higher average posterior than IG and more frequently provided a better fit, in 12 of the genes compared, positive support was shown for the IG model over LN. This suggests that in some cases, the IG model provided a better fit to the data than any other model. However, the overall lack of support for either model suggests that for this data set, and with limited taxa, the two models cannot be distinguished and that parameterization of both models is unnecessary.

## Discussion

In this paper, we have introduced four ideas relevant to Bayesian evolutionary analysis using relaxed molecular clocks: 1) model averaging of branch-rate models for relaxed clocks, which consequently permitted; 2) BF calculation for model selection under a Bayesian phylogenetic framework; 3) the sampling of rates on a branch as cumulative probabilities; and 4) the use of the IG distribution for modeling the distribution of rates across branches.

We believe that our model-averaging technique is a positive step toward a phylogenetic analysis scheme that removes the burden of model selection from the user, especially for aspects of the model that are nuisance and for which strong prior knowledge is not available. In phylogenetic estimation, a large source of bias and error comes from model misspecification. Our method can help eliminate such errors. By allowing the data to select an appropriate branch-rate model or a set of models to represent their characteristics, there is less room for error from manual parameterization of models.

The method in its current state poses a few issues in terms of application to standard Bayesian phylogenetic analysis. One issue involves convergence assessment for the distribution parameters of the model. Under our current implementation, values of each distribution parameter are sampled regardless of whether the distributions they correspond to are involved in the likelihood at that step in the MCMC. However, when a particular model is not in use at particular steps, the convergence of its distribution parameters during those steps is irrelevant to the actual convergence of the parameter. The calculated ESS therefore does not reflect the actual ESS, and conventional assessment of convergence cannot be applied here. It should be noted that this problem not only appears with our method but, as pointed out by Green (1995), is also a problem in rjMCMC. Methods to solve similar convergence assessment issues in rjMCMC have been published (Brooks and Giudici 1999; Castelloe and Zimmerman 2002), which can be modified for use in our scheme. Also, it is a feasible extension to common convergence assessment programs such as Tracer (Rambaut and Drummond 2007) to calculate ESS values conditional on the value of the indicator variable. These modifications are implemented in Tracer version 1.6 (soon to be released).

In addition, improvements to the mixing of the MCMC can be achieved by making adjustments to the proposal kernels. The way in which new parameter values are proposed, as well as proposals for sampling rates as quantile values, can be modified to produce more efficient convergence of the algorithm.

An obvious extension to our model-averaging technique would be to allow coparameterization of both uncorrelated relaxed clocks and autocorrelated relaxed clocks (Thorne et al. 1998; Aris-Brosou and Yang 2002). In phylogenetic estimation, it is often of interest as to whether rate of substitution is correlated between parent and child on a tree (Lepage et al. 2007; Ho 2009). Such an extension would allow model selection between uncorrelated and autocorrelated models, which can act as a test of autocorrelation in rates of substitution for a given set of data.

Although the model-averaging method proposed is valuable for finding the most appropriate model(s) given a set of models, its efficacy is dependent on the availability of models to choose from. To date, models proposed for explaining the rate of evolution are rudimentary and mostly only suffice to explain the distribution of substitution rate across the branches of a tree. There is a lack of models that accurately represent the intricate processes of evolution that generates these distributions in substitution rate. By modeling the actual processes of evolution, more accurate models may be possible. An important future step would be to describe more biologically relevant and accurate models of the substitution process.

An alternative algorithm was proposed for computing relaxed clock estimates by sampling the rates as cumulative probabilities. This implementation is equivalent to a full implementation of a relaxed clock but can improve convergence of the MCMC. This proposal is an improvement over the current implementation of discretized branch rates used in BEAST (Drummond et al. 2006), which has been criticized for its shortcomings (Rannala and Yang 2007); specifically, these are: its lack of ability to allow for identical rates (though this was later corrected in a subsequent release); its treatment of similar rates as identical rates; and its inability to accurately estimate branch rates when a small number of rate categories are used. Whether or not in practice this method improves the estimation of rates in comparison with categorical discretization has yet to be determined.

We also proposed the idea of using the IG distribution as an alternative to model rates of substitution across branches. In our analysis of mammalian nuclear genes, we found little evidence to separate the three parametric models compared. This may well be because of the small number of taxa analyzed. The method described here opens the way to larger empirical investigations of the relative merits of different relaxed molecular clock models. The underlying process of rate of substitution is complex, so it is unlikely that there is a single model that is optimal across all data sets. Thus, it is important to take a more liberal approach when choosing the appropriate model and even more crucial to have a wide selection of distributions to choose from. In the BEAST software framework, we have begun to implement an array of positive continuous parametric distributions to be used as models of rate distribution for model averaging of relaxed clocks, such as Gamma and inverse Gamma.

It should be noted that a drawback of the IG distribution is that there is no closed form for its quantile function. Though quantile values can be accurately approximated using Newton–Raphson (Tjalling 1995), it is computationally slower than those with a closed form. Difficulties also exist in approximating extreme values at the tails ends of the distribution. In practice, using this mammalian data set, we have found that the IG quantile calculation slows the entire MCMC by roughly 2-fold.

There are many aspects of modeling the rates of substitution among branches that have yet to be explored, and further developments will improve the shortcomings of the methods presented. Further progress will involve the explicit incorporation of the factors that cause rate variation among lineages, and we anticipate that model-averaging techniques, such as those demonstrated here, will play a role in discovering the factors that are most important in this context.

## Supplementary Material

Supplementary figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The authors would like to thank Allen Rodrigo for ideas contributing to the study and Michael Defoin Platel for work related to the research. W.L.S.L. was supported by Biomatters Ltd. and the Foundation for Research, Science and Technology of New Zealand.

## References

Adachi J, Hasegawa M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J Mol Evol.* 40:622–628.

Aris-Brosou S, Yang Z. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst Biol.* 51:703–714.

Bashir A, Ye C, Price AL, Bafna V. 2005. Orthologous repeats and mammalian phylogenetic inference. *Genome Res.* 15:998–1006.

Beerli P, Palczewski M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185:313–326.

Britten RJ. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231:1393–1398.

Brooks SP, Giudici P. 1999. Convergence assessment for reversible jump MCMC simulations. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian statistics 6. Oxford: Oxford University Press. p. 733–742.

Bunce M, Worthy TH, Phillips MJ, et al. (11 co-authors). 2009. The evolutionary history of the extinct ratite moa and New Zealand Neogene paleogeography. *Proc Natl Acad Sci U S A.* 106:20646–20651.

Castelloe JM, Zimmerman DL. 2002. Convergence assessment for reversible jump MCMC samplers. Technical Report 313. Iowa City (IA): Department of Statistics and Actuarial Science, University of Iowa.

Clyde MA. 1999. Bayesian model averaging and model search strategies. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian statistics 6. Oxford: Oxford University Press. p. 157–185.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Gaut BS, Muse SV, Clark WD, Clegg MT. 1992. Relative rates of nucleotide substitution at the rbcl locus of monocotyledonous plants. *J Mol Evol.* 35:292–303.

Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol.* 20:424–434.
Godsill SJ. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J Comput Graph Stat.* 10:230–248.
Gowri-Shankar V, Rattray M. 2007. A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol Biol Evol.* 24:1286–1299.
Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.
Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732.
Gu X. 1998. Early metazoan divergence was about 830 million years ago. *J Mol Evol.* 47:369–371.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.
Ho SYW. 2009. An examination of phylogenetic models of substitution rate variation among lineages. *Biol Lett.* 5:421–424.
Hoaglin DC, Welsch RE. 1978. The hat matrix in regression and ANOVA. *Am Stat.* 32:17–22.
Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian model averaging: a tutorial. *Stat Sci.* 14:382–401.
Huelsenbeck JP, Larget B, Alfaro ME. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol.* 21:1123–1133.
Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc.* 90: 773–795.
Kitazoe Y, Kishino H, Waddell PJ, Nakajima N, Okabayashi T, Watabe T, Okuhara Y. 2007. Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS One.* 2:e384.
Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 5:e1000520.
Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol.* 24:2669–2680.
Li W-H, Ellsworth DL, Krushkal J, Chang BHJ, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol.* 5:182–187.
Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A.* 90: 4087–4091.
Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J R Stat Soc Series B Stat Methodol.* 56:3–48.
Novacek MJ. 2001. Mammalian phylogeny: genes and supertrees. *Curr Biol.* 11:R573–R575.
Pagel M, Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc B Biol Sci.* 363:3955–3964.
Prasad AB, Allard MW, Program NCS, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol.* 25:1795–1808.
Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. *Mol Biol Evol.* 15:442–448.
Rambaut A, Drummond AJ. 2007. Tracer v1.4. [Internet]. [updated 2009 Jan 23; cited 2009 Oct 20]. Available from: http://beast.bio.ed.ac.uk/Tracer
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.
Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol.* 56:453–466.
Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, Douzery E. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.
Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C. 2004. Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. *Mol Biol Evol.* 21:397–403.
Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol.* 14:1218–1231.
Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research. New York: Freeman and Co.
Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Mol Phylogenet Evol.* 41:384–394.
Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol.* 18:1001–1013.
Takagi K, Kumagai S, Matsunaga I, Kusaka Y. 1997. Application of inverse Gaussian distribution to occupational exposure data. *Ann Occup Hyg.* 41:505–514.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Waterman MS, editor. Some mathematical questions in biology: DNA sequence analysis. Providence (RI): American Mathematical Society. p. 57–86.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15: 1647–1657.
Tjalling JY. 1995. Historical development of the Newton-Raphson method. *SIAM Rev.* 37:531–551.
Wu C-H, Drummond AJ. 2011. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo. *Genetics* 188: 151–164.
Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A.* 82:1741–1745.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.
Yoder AD. 1997. Back to the future: a synthesis of strepsirrhine systematics. *Evol Anthropol.* 6:11–22.
Yule U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos Trans R Soc Lond B Biol Sci.* 213:21–87.
Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.