Bayesian Selection of Relaxed-clock Models: Distinguishing Between Independent and Autocorrelated Rates

Muthukumaran Panchaksaram[1], Lucas Freitas[1,2], and Mario dos Reis,[1,*]

March 11, 2024

1. School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4NS, United Kingdom

2. Department of Genetics, Evolution, Microbiology and Immunology, State University of Campinas, Campinas, São Paulo, Brazil

**E-mail: m.dosreisbarros@qmul.ac.uk*

Abstract

In Bayesian molecular-clock dating of species divergences, rate models are used to construct the prior on the molecular evolutionary rates for branches in the phylogeny, with independent and autocorrelated rate models being commonly used. The two class of models, however, can result in markedly different divergence time estimates for the same dataset, and thus Bayesian model selection appears necessary to select for the best rate model and obtain reliable inferences of divergence times. However, the properties of Bayesian rate model selection are not well understood, in particular when the number of sequence partitions analysed increases and when fossil calibrations are misspecified. Furthermore, Bayesian rate model selection is computationally expensive as it requires calculation of marginal likelihoods by MCMC sampling, and therefore methods that can speed up the model selection procedure without compromising its accuracy are desirable. In this study,

we use a combination of computer simulations and real data analysis to investigate the statistical behavior of Bayesian rate model selection and we also explore approximations of the likelihood to improve computational efficiency in large phylogenomic datasets. Our simulations demonstrate that the posterior probability for the correct rate model converges to one as more molecular sequence partitions are analyzed and when no fossil calibrations are used, as expected due to asymptotic Bayesian model selection theory. Furthermore, we also show the model selection procedure is robust to slight misspecification of fossil calibrations, and reliable inference of the correct rate model is possible in this case. However, we show that when fossil calibrations are seriously misspecified, calculated model probabilities are completely wrong and may converge to one for the wrong rate model. Finally, we demonstrate that approximating the phylogenetic likelihood under an arcsine branch-length transform can dramatically reduce the computational cost of rate model selection without compromising accuracy. We test the approximate procedure on two large phylogenies of primates (372 species) and flowering plants (644 species), replicating results obtained on smaller datasets using exact likelihood. Our findings and methodology can assist users in selecting the optimal rate model for estimating times and rates along the Tree of Life.

Running title: Relaxed clocks

Keywords: molecular-clock dating, marginal likelihood, autocorrelated rates, independent rates, Bayesian analysis

INTRODUCTION

Models of evolutionary rate variation among lineages are important because they are used to construct the rate prior in Bayesian inference of species divergence times (Thorne et al. 1998; Drummond et al. 2006; Rannala and Yang 2007). Roughly, three models (or class of models) are commonly used: The strict clock, the independent-rates, and the autocorrelated-rates models. In the strict clock model, rates among branches in the phylogeny are equal (Zuckerkandl and Pauling 1965), an assumption that appears useful in the analysis of closely related species in which rates among branches are very similar (Yang and Rannala 2006). In the independent-rates model, rates among branches are independent, identical draws from a probability distribution

such as the log-normal or the gamma distributions (Drummond et al. 2006; Rannala and Yang 2007), whereas in the autocorrelated-rates model, rates among branches are generated by an autocorrelated stochastic process such as the geometric Brownian motion process (Thorne et al. 1998; Rannala and Yang 2007) or the Cox-Ross-Ingersoll process (Lepage et al. 2007). A plethora of rate variation models have been developed encompassing the three classes of models or combinations of them (Huelsenbeck et al. 2000; Drummond and Suchard 2010; Heath et al. 2012; Lartillot et al. 2016; Ho and Duchêne 2014).

Choosing an appropriate rate model appears important because the different models can produce different divergence time estimates on the same dataset (e.g., dos Reis et al. 2018), and because different datasets may support different models. For example, Linder et al. (2011) found best support for the independent-rates model in flowering plants, but best support for the autocorrelated-rates model in Primates, when no fossil calibrations were used. Similarly, dos Reis et al. (2018) found a preference for the autocorrelated-rates model in Primates, while Barba-Montoya et al. (2018) found best support for the independent-rates model in flowering plants. Lepage et al. (2007) found autocorrelated rates had the better fit in eukaryotic, vertebrate and mammal datasets (see also McGowen et al. 2020; Álvarez-Carretero et al. 2022). Surprisingly, distinguishing between rate models was found to be difficult in simulated data (Ho et al. 2015), a puzzling result given that support for specific rate models has been strong in the real data analyses. The reasons for this discrepancy are not clear but at least two factors appear to limit our ability to distinguish between rate models: the non-identifiability of times and rates in Bayesian inference and sampling errors in the calculation of marginal likelihoods. Furthermore, real data analyses have been carried out on phylogenies with a few dozen taxa, and it is not clear whether support for a given rate model would still be maintained for larger phylogenies with hundreds of taxa. Analyses on large phylogenies have not been carried out because Bayesian model selection would be too computationally expensive on these datasets (e.g., Álvarez-Carretero et al. 2022), and thus inference technologies that can speed up analysis on large datasets are desirable.

Here we study the statistical properties of Bayesian selection of rate models for inference of divergence times. First, we use computer simulations to study the asymptotic properties of

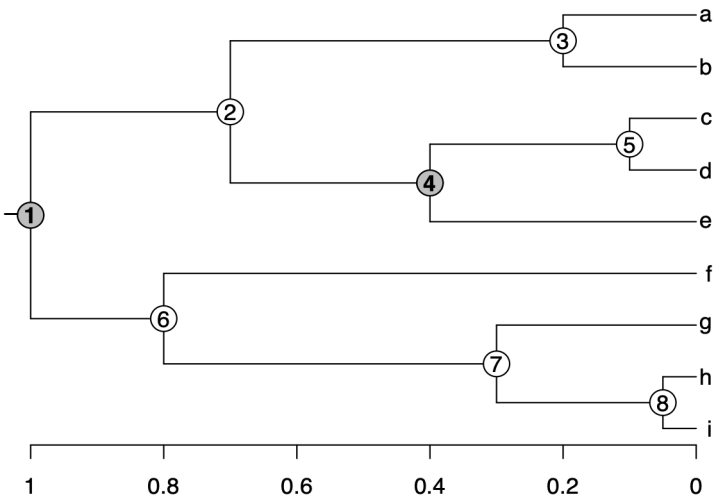
the model selection approach when the amount of data (the number of loci or alignment partitions) increases. We show that, when no fossil calibrations are used or when fossil calibrations are slightly mis-specified, Bayesian inference is well-behaved and Bayesian model selection asymptotically selects the correct rate model. On the other hand, when fossil calibrations are badly misspecified, the model selection approach is unreliable and the wrong model may be selected even with large datasets. We also study the use of likelihood approximations to speed up computation in large empirical datasets. We show that when using an appropriate transform on the branch lengths, so that the likelihood function is well approximated at the tails, the approximate method produces reliable estimates allowing Bayesian rate model selection in datasets with hundreds of taxa.

Simulation Analysis

Simulation strategy

Overview.— We use the bifurcating, rooted phylogeny of Fig. 1, with $s = 9$ species and $n = 2s - 2 = 16$ branches, to simulate molecular alignments under the strict (STR), independent log-normal (ILN) and geometric Brownian motion (GBM) rate models. Let r_i be the molecular evolutionary rate (in units of substitutions per site per time unit) and Δt_i be the time duration of the i-th branch. Simulating under a relaxed-clock model requires sampling n rates from the appropriate statistical distribution and then calculating the length of each branch in substitutions per site: $b_i = r_i \Delta t_i$. The resulting phylogeny, with branch lengths in substitutions per site, is then used to simulate the molecular sequences alignments with the Evolver program from PAML (Yang 2007). Each simulated alignment is then analysed with the computer program MCMCTree (Yang 2007) for MCMC Bayesian inference of node ages and branch rates under each rate model. Marginal likelihoods and posterior probabilities for each rate model under each simulated alignment are then estimated using the stepping-stones method (Xie et al. 2011).

Simulating rates.— We have developed a new R package, `simclock`, to simulate rates on a phylogeny under the three rate models. In the ILN model, we sample the n independent,



True age	No calibration	Slightly misspecified	Badly misspecified
$t_1 = 1.0$	(0.999, 1.001)	(1.0, 1.5)	(0.6, 0.8)
$t_4 = 0.4$	–	(0.3, 0.4)	(0.5, 0.6)

Figure 1: Phylogeny used to simulate nucleotide alignments under rate models. The phylogeny has eight internal nodes, $1, \ldots, 8$, with true ages $t_i = 1, 0.7, 0.2, 0.4, 0.1, 0.8, 0.3$, and 0.05. Bayesian rate model selection analyses are carried out using three fossil calibration set-ups: (1) No calibrations, i.e., the root age, t_1, is fixed to one, which, in MCMCtree, is done by using a narrow uniform distribution between 0.999 and 1.001; (2) slighly misspecified calibrations, in which the true ages, t_1 and t_4, are at the bounds of the corresponding calibration densities; or (3) badly misspecified calibrations, in which the true ages, t_1 and t_4, are not contained withint the corresponding calibration densities. The tree topology and node ages are the same used by Yang and Rannala (2006) in their simulation analyses.

identically distributed rates from the joint log-normal distribution

$$f(r_1,\ldots,r_n) = \prod_{i=1}^{n} \frac{1}{r_i\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(r_i/\mu)+\sigma^2/2)^2}{2\sigma^2}\right], 0 < r_i, \mu, \sigma^2 < \infty, \tag{1}$$

where μ is the mean rate and σ^2 is the log-variance, which controls the departure from the strict clock. When σ^2 is close to zero, the branch rates are close to μ, whereas when σ^2 is large, there will be large rate variation among branches and the molecular clock will be seriously violated.

To simulate rates under the GBM model, we use the GBM density specification by Rannala and Yang (2007). Let i, j be two daughter branches and a their parent branch, let $y_i = \log r_i$ and $\mathbf{y} = (y_i, y_j)$. Then y_i and y_j have a joint normal distribution conditioned on y_a (i.e., r_i and r_j have a joint log-normal distribution conditioned on r_a), with density

$$f(\mathbf{y} \mid y_a) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\eta)^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}-\eta)\right], \tag{2}$$

where

$$\eta = \begin{bmatrix} y_a - (\Delta t_a + \Delta t_i)\sigma^2/4 \\ y_a - (\Delta t_a + \Delta t_j)\sigma^2/4 \end{bmatrix}, \text{ and } \Sigma = \frac{\sigma^2}{2}\begin{bmatrix} \Delta t_a + \Delta t_i & \Delta t_a \\ \Delta t_a & \Delta t_a + \Delta t_j \end{bmatrix}, \tag{3}$$

are the mean vector and covariance matrix respectively. Note that $T_i = (\Delta t_a + \Delta t_i)/2$ is the time span between the midpoint of the parent branch a and the midpoint of its daughter branch i (see Fig. 1 in Rannala and Yang 2007). Thus, $\sigma^2 T_i$ is the amount of accumulated variance for the log-rate of the daughter branch. When $\sigma^2 T_i$ is close to zero, the daughter rate is close to the parent rate, whereas when $\sigma^2 T_i$ is large, the daughter rate diffuses away from that of the parent, and the molecular clock is seriously violated. Note the two branches around the phylogeny's root do not have a parent branch. In this case we set $y_a = \log\mu$ and $t_a = 0$. Thus, μ is the expected rate for the phylogeny. To sample the n rates, we start at the root and sample rates for its two daughter branches, $r_i = \exp y_i$ and $r_j = \exp y_j$, using Eq. (2). We then traverse the remaining nodes in the phylogeny and sample daughter rates after their parent rates have been sampled, resulting in the n required rates.

For the STR model, $r_i = \mu$, for all branches. Note the STR model is a special case of both ILN and GBM when $\sigma^2 = 0$.

Simulating alignments.— In relaxed-clock dating, when both the number of sites and the number of loci (i.e. the number of alignment partitions) analysed approach infinity, the posterior of divergence times converges to a limiting distribution in which the uncertainties of node ages are proportional to the age of the root (Rannala and Yang 2007). Rannala and Yang (2007) show that five loci already show good convergence to the limiting distribution, with 30 loci leading to excellent convergence (Fig. 4 in their paper). Thus, to study the asymptotic properties of Bayesian rate model selection, we simulate alignments with $L = 1, 2$, and 5 loci, with each locus having 1,000 sites. This set up is the largest we can analyse in reasonable time given expensive MCMC calculation of marginal model likelihoods but with reasonable convergence to the limiting distribution.

Let A be an alignment to be simulated, partitioned into L loci. Each locus evolves independently, with its own sets of rates for branches. For each locus, we sample a value for the locus rate, μ_h, $h = 1, \ldots, L$, from the gamma distribution $\Gamma(\alpha = 2/L, \beta = 2/L)$, which has mean $\alpha/\beta = 1$ and variance $\alpha/\beta^2 = L/2$. This dependency of the gamma density on L is used for compatibility with the gamma-Dirichlet prior used in MCMCtree (dos Reis et al. 2014). Next, for each locus, we sample a value for the log-rate variance parameter, σ_h^2, from the gamma distribution $\Gamma(2/L, 20/L)$, with mean 0.1 and variance $L/200$. This sampling density leads to values of σ_h^2 that are small and close to 0.1, thus the resulting simulated rates are similar to each other, and the phylogeny is clock-like (CL). We label simulations using this sampling density as ILN-CL or GBM-CL. Alternatively, we may also sample σ_h^2 from the gamma density $\Gamma(2/L, 2/L)$ with mean 1 and variance $L/2$. This leads, in many simulations, to values of σ_h^2 that are very large, with simulated rates being very different to each other, resulting in phylogenies in which the clock is seriously violated (SV). We label simulations using this sampling density as ILN-SV or GBM-SV. Once the values of $\mu = (\mu_1, \ldots, \mu_L)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_L^2)$ have been set, we use function `simclock::relaxed.tree` to simulate branch rates for each locus. The resulting locus trees, with branches in substitutions per site, are then used with the Evolver program from PAML (Yang 2007) to simulate L loci with 1,000 sites each under the Jukes and Cantor (1969) substitution model.

Table 1: Summary of rate model simulations

Rate model[a]	Sampling densities		Number of simulated alignments		
	μ	σ^2	$L = 1$	$L = 2$	$L = 5$
STR	$\Gamma(2/L, 2/L)$	-	1,000	1,000	1,000
ILN-CL	$\Gamma(2/L, 2/L)$	$\Gamma(2/L, 20/L)$	1,000	1,000	1,000
ILN-SV	$\Gamma(2/L, 2/L)$	$\Gamma(2/L, 2/L)$	1,000	1,000	1,000
GBM-CL	$\Gamma(2/L, 2/L)$	$\Gamma(2/L, 20/L)$	1,000	1,000	1,000
GBM-SV	$\Gamma(2/L, 2/L)$	$\Gamma(2/L, 2/L)$	1,000	1,000	1,000

a. STR: Strick clock, ILN: Independent log-normal; GBM: Geometric Brownian motion. In models labelled CL simulated rates among branches are clock-like, and thus similar to each other. In models labelled SV, simulated rates among branches tend to be very dissimilar, and thus the molecular clock is seriously violated.

In total we simulated alignments from five rate model configurations (STR, ILN-CL, ILN-SV, GBM-CL, GBM-SV) and three locus configurations $L = 1, 2$, and 5, for a total of $5 \times 3 = 15$ simulation configurations. For each configuration, 1,000 alignments were simulated, resulting in 15,000 simulated alignments. Table 1 summarises the simulations.

Bayesian rate model selection and rate prior.— Let M be a rate model (one of STR, ILN or GBM). The posterior of divergence times (node ages), given the alignment, $A = (A_1, \ldots, A_L)$, and rate model, is

$$f(\mathbf{t}, \mathbf{r}, \mu, \sigma^2 \mid A, M) = \frac{f(\mathbf{t}) f(\mu) f(\sigma^2) \prod_h f(\mathbf{r}_h \mid \mathbf{t}, M) f(A_h \mid \mathbf{t}, \mathbf{r}_h)}{f(A \mid M)}, \tag{4}$$

where $f(\mathbf{t})$ is the prior on times, $f(\mu) f(\sigma^2) \prod_h f(\mathbf{r}_h \mid \mathbf{t}, M)$ is the prior on the rates and rate model parameters, $f(A_h | \mathbf{t}, \mathbf{r}_h)$ is the likelihood for the h-th locus in the sequence alignment, and

$$f(A \mid M) = \int f(\mathbf{t}) f(\mu) f(\sigma^2) \prod_h f(\mathbf{r}_h \mid \mathbf{t}, M) f(A_h \mid \mathbf{t}, \mathbf{r}_h) \, \mathrm{d}\mu \mathrm{d}\sigma^2 \mathrm{d}\mathbf{t} \mathrm{d}\mathbf{r} \tag{5}$$

is the marginal-likelihood of the data, which depends on the rate model used to construct the rate prior. Selecting the best rate model for a given data set A, requires computing its marginal likelihood three times, once for each rate model. The posterior model probability is then

$$\Pr(M_i | A) = \frac{\Pr(M_i) f(A \mid M_i)}{\sum_j \Pr(M_j) f(A \mid M_j)}, \tag{6}$$

where $\Pr(M_i)$ is the prior model probability. Here we use $\Pr(M_i) = 1/3$. We make a few observations about construction of the rate prior:

1. For the STR model, $f(\mathbf{r}_h \mid \mathbf{t}, M) = 1$ and $f(\sigma^2 \mid M) = 1$.

2. For the ILN model, $f(\mathbf{r}_h \mid \mathbf{t}) = f(\mathbf{r}_h)$ is calculated using the joint log-normal density of Eq. (1).

3. For the GBM model, $f(\mathbf{r}_h \mid \mathbf{t})$ is calculated using the log-normal density associated with Eq. (2),

$$f(r_i, r_j \mid r_a) = f(y \mid y_a)|J|, \tag{7}$$

   where $|J|$ is the Jacobian of the $(r_i, r_j) \rightarrow (y_i, y_j)$ transform (Rannala and Yang 2007). The joint prior on all rates is then

$$f(\mathbf{r} \mid \mathbf{t}) = \prod_a f(r_i, r_j \mid r_a), \tag{8}$$

   where the product is over the $s - 1$ internal nodes in the phylogeny.

4. The joint prior $f(\mu)f(\sigma^2)$ over loci is calculated using the gamma-Dirichlet density (dos Reis et al., 2014). This prior uses a gamma density on the mean across locus rates, $\bar{\mu} = \sum \mu_i / L$, and on the mean log variance $\bar{\sigma}^2 = \sum \sigma_i^2 / L$, and then uses a Dirichlet distribution to partition the uncertainty across the loci. For the locus rates, we use $\bar{\mu} \sim \Gamma(2, 2)$ with $\alpha = 1$ in the Dirichlet. For the log-variance, we use $\bar{\sigma}^2 \sim \Gamma(2, 20)$, for clock-like datasets, or $\bar{\sigma}^2 \sim \Gamma(2, 2)$, for the severely violated clock datasets. In both cases $\alpha = 1$ in the Dirichlet is used.

Fossil calibrations.—Calculation of marginal likelihoods and posterior rate model probabilities are carried out under three fossil calibration strategies (Fig. 1):

1. No calibrations: The age of the root is fixed to one, $t_1 = 1$, and no fossil calibrations are applied to any other nodes. The prior on the age of the remaining nodes is given by a birth-death process with parameters $\mu = 1, \lambda = 1, \rho = 0$, resulting in a uniform kernel (Yang and Rannala 2006). In this case the model appears to be fully identifiable (see

discussion), and we expect posterior model probabilities to converge to one for the true model as the number of loci increases.

2. Slightly misspecified calibrations: The root is calibrated with the uniform distribution $(1.0, 1.5)$, which has mean 1.25, thus this calibration suggests a root age that is older than the true age, $t_1 = 1$. On the other hand, node 4 has the uniform calibration $(0.3, 0.4)$, which has mean 0.35 which is younger than the true age $t_4 = 0.4$. Thus the two calibrations are in conflict and push the node ages in different directions, and the true node ages are barely contained within the calibration bounds. The prior on the remaining node ages is set using the same birth-death kernel as above. In this case we expect the conflict between the two fossil calibrations to adversely affect the convergence of posterior model probabilities.

3. Badly misspecified calibrations: The root has the uniform calibration $(0.6, 0.8)$, which has mean 0.7 and is thus very different than the true age, $t_1 = 1$. Node 4 has the uniform calibration $(0.5, 0.6)$ which has mean 0.55 and is also very different to the true age, $t_4 = 0.4$. Here the two calibrations are in big conflict and push the node ages close together. Furthermore, neither calibration contains the true node age. The prior on the remaining node ages is set as above. In this case we expect the strong conflict between the two fossil calibrations to severly affect convergence of posterior model probabilitiets.

In cases 2 and 3, the calibration bounds are soft, that is, there are small tail probabilities, $p_l = p_u = 2.5\%$, extending beyond the lower and upper bounds. This means the posterior distribution of node ages can converge to values outside the calibration bounds, as a means to make the analaysis more robust to misspecification of fossil calibrations (Yang and Rannala 2006).

MCMC sampling.—The marginal likelihood (Eq. 5) cannot be calculated analytically, thus, we use MCMC sampling of power posteriors under the stepping-stones method to calculate the log-marginal likelihood, $\ell = \log f(A \mid M)$ (Xie et al. 2011). The power posterior is given by $q_\beta = f(\mathbf{t}) f(\mu) f(\mathbf{r} \mid \mathbf{t}, M) f(\sigma^2 \mid M) f(A \mid \mathbf{t}, \mathbf{r})^\beta$ with $0 \leq \beta \leq 1$. When $\beta = 0$, the power posterior reduces to the prior, and when $\beta = 1$ to the posterior. Thus, by selecting n values of β between 0 and 1, one can sample likelihood values (using MCMC) from the power posteriors in a path

from the prior to the posterior. The sampled likelihoods are then used to estimate ℓ (see Xie et al. 2011 for details). Here we use $n = 8$ β values, which requires running 8 independent MCMC samples to calculate one marginal likelihood value. We use MCMCTree to obtain the required MCMC samples, and we use the `mcmc3r` R package (dos Reis et al. 2018) to choose the β values and to calculate ℓ and $\Pr(M \mid A)$. The likelihood, $f(A \mid \mathbf{t}, \mathbf{r})$, is calculated under the Jukes and Cantor (1969) model in all cases. We note our analysis strategy required a very large number of MCMC chains: Each one of our 15,000 simulated alignments was analysed under each of 3 rate models, each rate model under three calibration strategies, with each combination requiring 8 β values. This gives a total of $15,000 \times 3 \times 3 \times 8 = 1.08$ million MCMC chains. These chains were run in parallel over several months at QMUL's Apocrita high-performance computing cluster.

Simulation results

Figure 2 summarise the results from the simulation analysis. When the age of the root is fixed to one, $t_1 = 1$, and when no fossil calibraitons are used, posterior model probabilities for the correct rate model converge to one in all cases (white dots, Fig. 2). When fossil calibrations are used, and when these are slightly misspecified, rate model selection is robust to the fossil misspecification and posterior model probabilities still converge to one for the correct model (black dots, Fig. 2). However, in this case, the rate of convergence is slightly worse when the true model is either STR (Fig. 2A-A') or GBM (Fig. 2C-C'), but not for the ILN model (Fig. 2B-B'). On the other hand, when fossil calibrations are badly misspecified, posterior model probabilities converge to wrong values in all cases (asterisks, Fig. 2), with the worst behaviour for the STR model, for which model probabilities converge to zero for five loci even despite this being the correct model. The bad calibrations case is very pathological, as in many cases, a wrong model will have a higher posterior probability compared to the correct model, even when analysing five loci. Because in real data analyses is not possible to guarantee that all fossil calibrations are correct, we suggest workers select the rate model in an analysis without fossil calibrations, as our simulations show this analysis setup shows the best behaviour in all cases (Fig. 2).

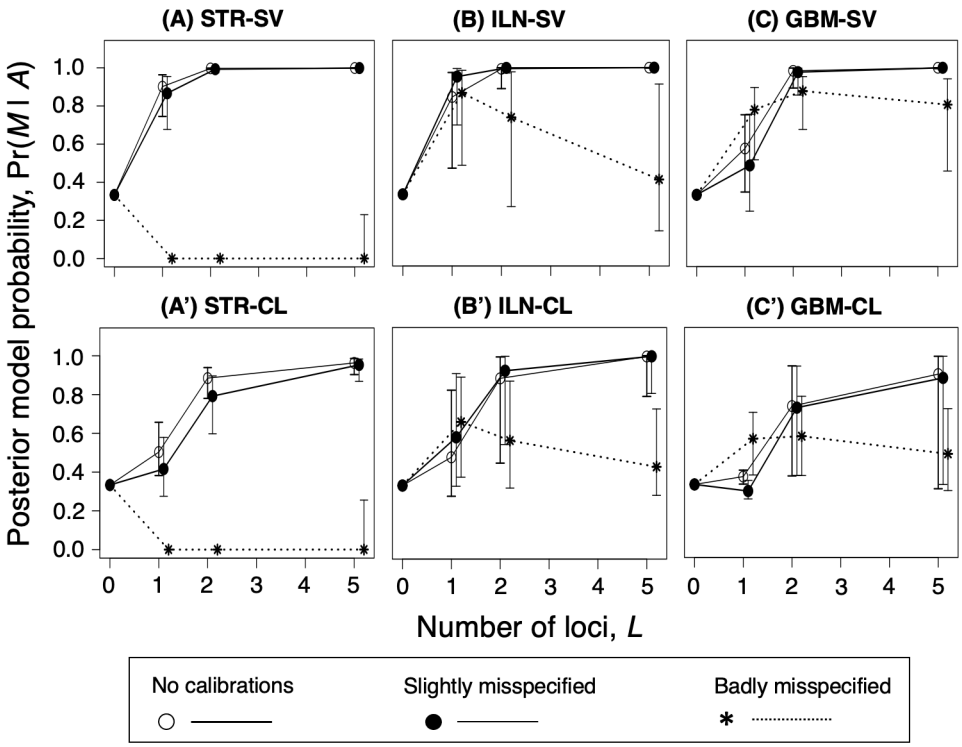


Figure 2: Posterior rate model probabilities for the true model in simulated datasets. The true models are (A, A') STR, (B, B') ILN, and (C, C') GBM. Datasets are analysed under three fossil calibration strategies (no calibration, slightly misspecified, and badly misspecified calibrations). Each dot is the median of $\Pr(M \mid A)$ for the true model across 1,000 simulated datasets, and the vertical bars are the corresponding interquantile range across the simulations. Note that in (A) and (A'), the simulated datasets are the same, but the prior on $\bar{\sigma}^2$ is either the SV or the CL prior when the data are analysed under the ILN and GBM models. For $L = 0$, the dots are plotted at the prior model probability, $\Pr(M) = 1/3$.

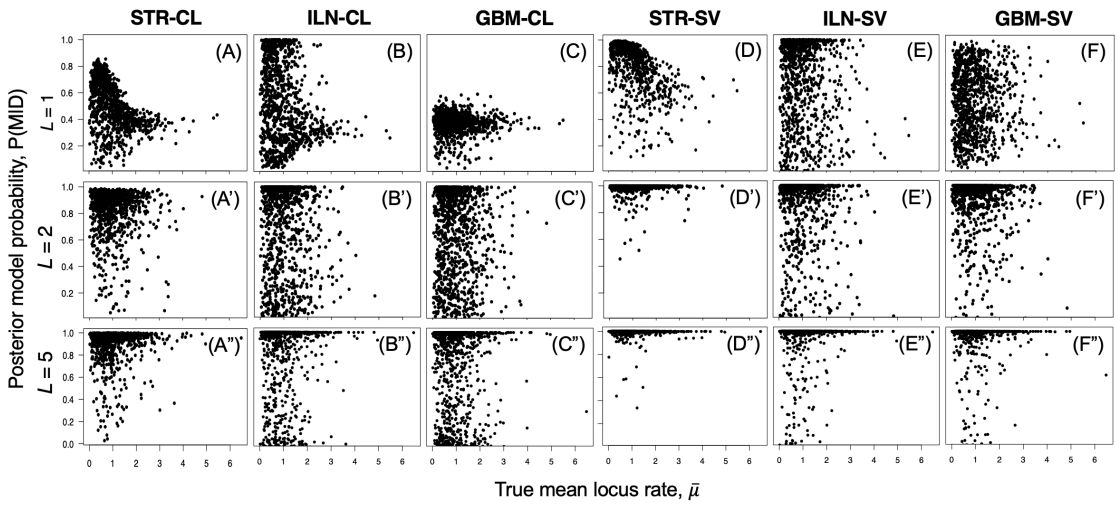


Figure 3: Posterior rate model probabilities for the true model vs the true mean locus rate. Each dot represents the posterior probability of the true model for a single simulated dataset (i.e., each panel contains 1,000 dots corresponding to 1,000 simulations for the particular parameter setting).

Datasets analysed using the clock-like prior, $\bar{\sigma}^2 \sim \Gamma(2, 20)$, showed the slower convergence in posterior probability for the correct model (Fig. 2A'-C'), when compared to the seriously violated clok prior, $\bar{\sigma}^2 \sim \Gamma(2, 2)$ (Fig. 2A-C). The reasons for this are not clear, but we note the clock-like prior is more informative, i.e., it has a smaller variance, than the seriously violated prior. Thus, our results here suggest a diffuse prior on $\bar{\sigma}^2$ should be preferred.

Figures 3 and 4 show posterior model probabilities for the true model vs. true mean locus rate, $\bar{\mu}$, and true mean rate parameter, $\bar{\sigma}^2$, for each simulated dataset. The patterns in the figure are very instructive. For example, it is difficult to ascertain the right model when the true mean rate is very high, and in particular when the number of loci is low. Also, the CL prior makes it hard to distinguish the right model when the true model is GBM, and this effect is reduced as the number of loci is increased. For computational reasons we could not test larger numbers of loci (i.e., $L > 5$) but it is clear from the asymptotic analysis (2) that the ability to ascertain the correct model should converge with probability one to the right model in all cases when no fossil calibrations are used.

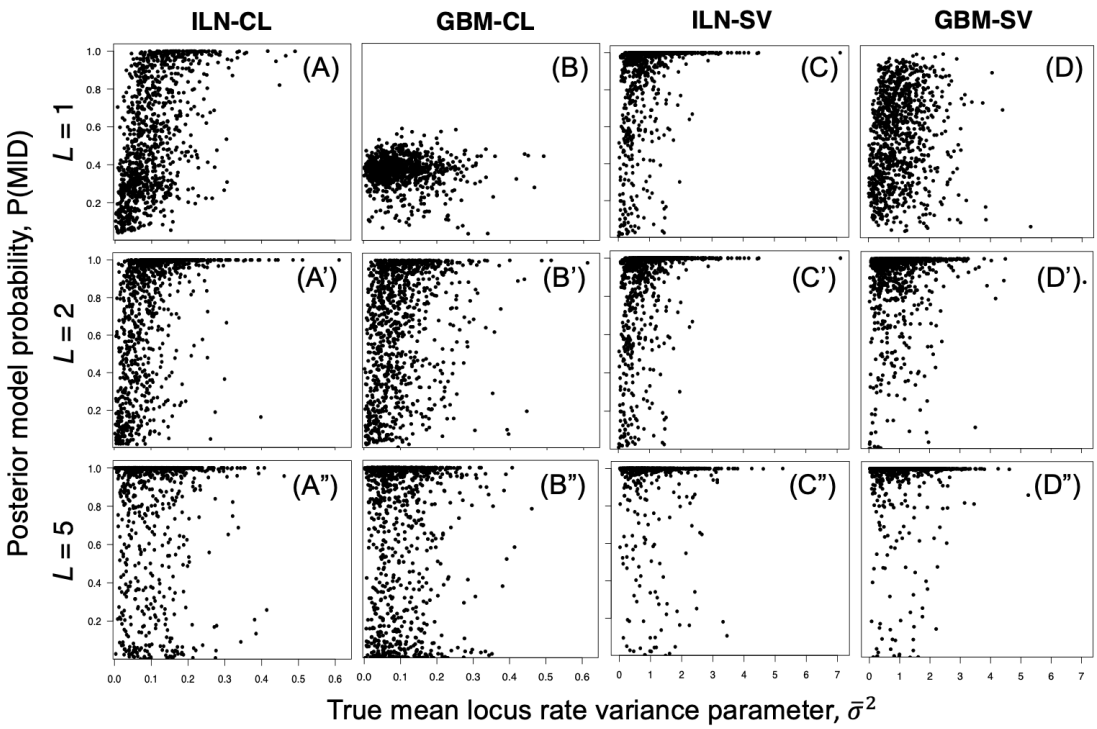


Figure 4: Posterior rate model probabilities for the true model vs the true mean drift parameter. See legend of Fig. 3.

Approximations for Fast Likelihood Computation on Large Phylogenies

Thorne et al. (1998) (see also Seo et al., 2004) suggest approximating the likelihood to speed up computation during MCMC sampling of divergence times, with dos Reis and Yang (2011) suggesting transformation of branch lengths to improve the accuracy of the approximation. The approximation can provide speed-ups of up to 1,000x compared with exact likelihood calculation depending on the dataset (Battistuzzi et al. 2011). However, in the original approximation by Thorne et al. (1998), the approximation error can be substantial at the tails of the likelihood function, with the error worsening as the distance away from likelihood peak increases (Fig. 5). The transforms proposed by dos Reis and Yang (2011) reduce the approximation error at the tails, with the Arcsine-based transform providing the smallest errors (Fig. 5). When conducting MCMC sampling of divergence times under the posterior (Eq. 4), the errors in the likelihood tail appear unimportant because values at the tail are rarely sampled (dos Reis and Yang 2011), and posterior estimates from the approximate method are indistinguisable from the exact method in

large datasets (dos Reis and Yang 2011). On the other hand, tail errors are expected to be important in marginal-likelihood calculation. For example, in the stepping-stones method, tail values will be sampled often when β is close to zero in the power posterior. It is not clear whether tail errors under the branch-length transforms are small enough to obtaine reliable estimates of the marginal likelihoods. Furthermore, when the clock is seriously violated, and when the data are analysed under the clock, the approximate method does not work well because sampled branch-lengths are very far from the likelihood peak (dos Reis and Yang 2011). Here we examine the accuracy of the approximate method when estimating marginal likelihoods for rate models and use the approximate method to select for the best rate model in phylogenies with several hundred taxa.

Accuracy of marginal likelihood estimation under the approximate method.— We use two real datasets. The first are the 12 mitochondrial protein-coding genes from seven ape species analysed by Yang and Rannala (2006). The genes are dividided into three partitions according to codon position, with each partition having 3,331 sites. The second dataset are the 1st and 2nd codon positions from 10 nuclear protein-coding genes from 72 mammal species. This is a subsample of 10 genes out of the 15,268 analysed by Álvarez-Carretero et al. (2022), with the number of sites ranging from 406 to 4,946 among the 10 genes. Thus, the two datasets result in 13 alignments: three alignments for the ape data, and ten alignments for the mammal data. For each alignment, we calculate marginal likelihoods for each rate model, both under the exact and approximate likelihood methods, resulting in $13 \times 3 \times 2 = 78$ marginal likelihood caculations. All 13 alignments are analysed under the HKY+Γ model (Hasegawa et al., 1985; Yang, 1994). We note that in the exact method, the κ and α parameters from the HKY+Γ model are sampled during MCMC, while they are fixed at their maximum likelihood estimates (MLEs) during MCMC sampling with the approximate method (dos Reis and Yang, 2011). This means the marginal likelihood estimates between the exact and approximate approachs cannot be compared directly: the approximate method will result in larger marginal likelihoods because in this case integration over the priors on κ and α does not take place. Instead, we compare the log-Bayes factors between models. Let ℓ_i and ℓ_j be the log-marginal likelihoods for two models ($i \neq j$). The log-Bayes factor, or marginal log-likelihood difference, is $\Delta\ell_{ij} = \ell_j - \ell_i$.

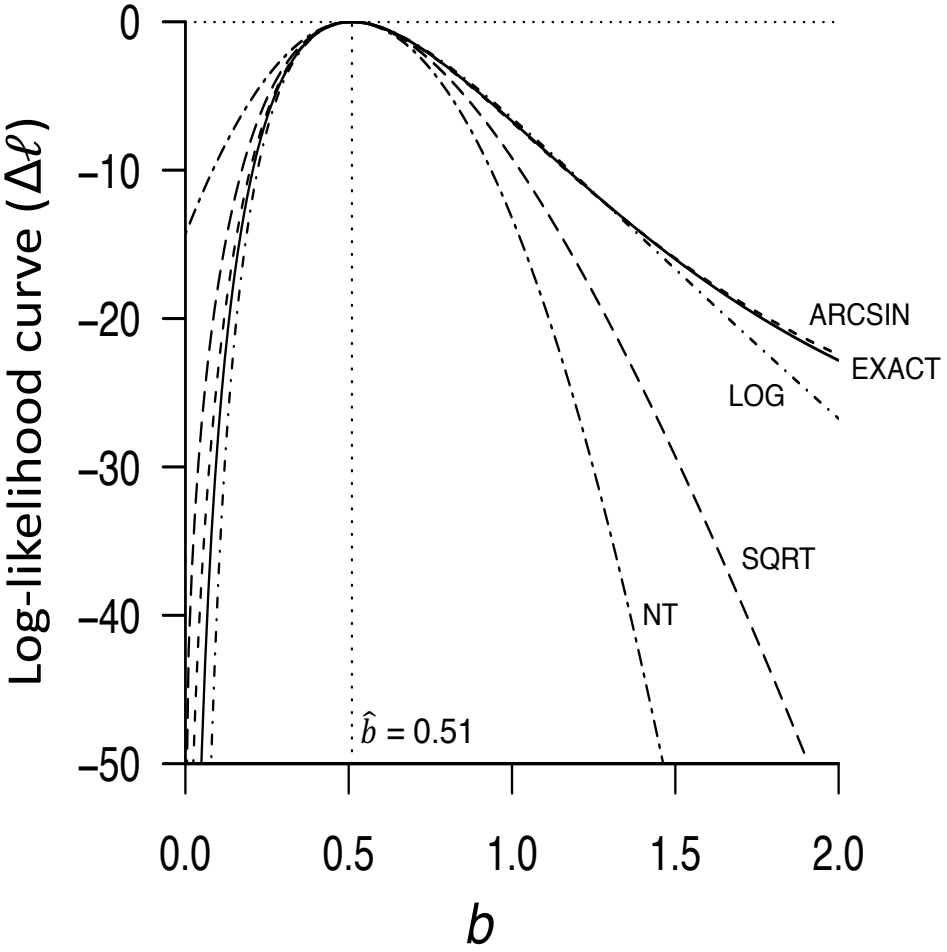


Figure 5: Approximation of the likelihood as a function of the branch length, b. The data are a pairwise alignment with $x = 37$ differences out of $n = 100$ sites, analysed under the JC69 model. The true likelihood is shown as a solid line, with the maximum likelihood value scaled to zero. NT: the original approximation by Thorne et al. (1998) without using transformation of the branch lengths. SQRT: square-root transform. LOG: log transform. ARCSIN: Arcsine-based transform. See dos Reis and Yang (2011) for details. The NT approximation is very poor at the left and right tails of the likelihood. The SQRT and LOG transforms provide reasonable approximations to the left tail, but bad approximations to the right tail. The ARCSIN method provides a reasonable approximation of both, left and right tails. All approaches produced good approximations near the peak of the likelihood (at $\hat{b} = 0.51$).

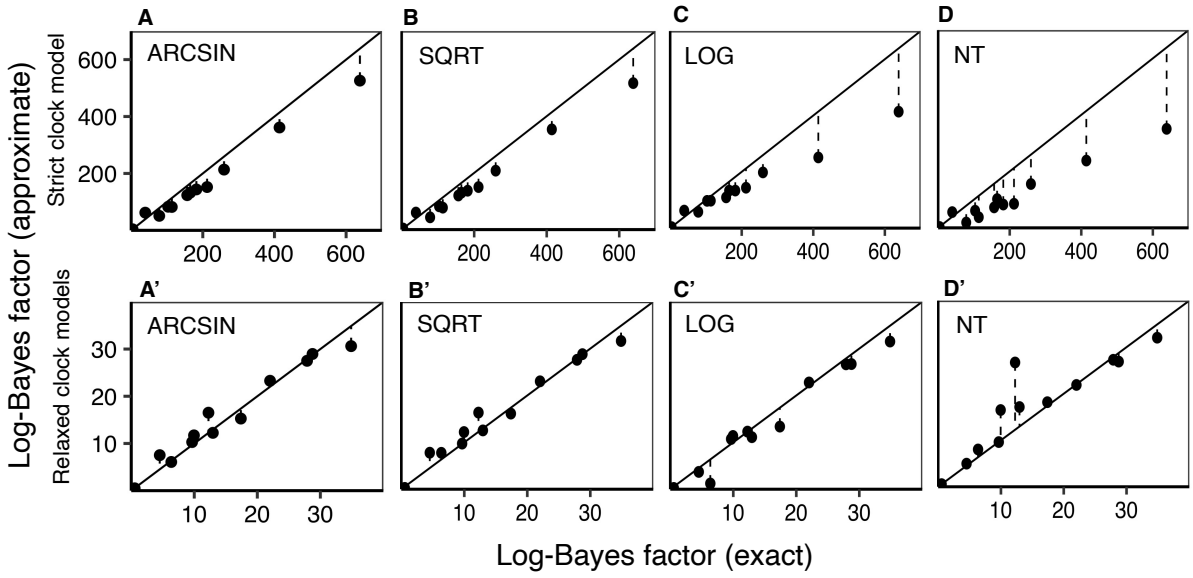


Figure 6: Comparison of log-Bayes factors ($\Delta\ell$) of rate models under exact and approximate likelihood calculation. The data are alignments of ten nuclear genes from 72 mammal species and the three codon pisitions from the mitochondrial proteins from seven ape species (13 aligments in total). (A-D) Comparisons between the GBM and STR rate models for four methods to transform branch lengths (ARCSIN, SQRT, LOG and NT, Fig. 5). (A'-D') Comparisons between the GBM and ILN models for the four transform methods. Each dot is the $\Delta\ell$ value from for a given model comparison on a single alignment.

Figure 6 shows the log-Bayes factors calculated for the 13 alignments and 3 rate models under the exact and approximate approaches. In all 13 alignments, the molecular clock is seriously violated, and the approximate $\Delta\ell$ values calculated in this case are not reliable. For example, the largest discrepancies in $\Delta\ell$ values can be as large as 100 log-units (Fig. 6A-D). On the other hand, when comparing the two relaxed-clock models (ILN vs GBM), the approximations using branch transformations (ARCSINE, SQRT and LOG) perform very well, showing only slight errors in the estimates of $\Delta\ell$ (Fig. 6A'-C'). The approximation without branch-transformation (NT) does show substantial errors in the calculation of $\Delta\ell$ in two comparisons (Fig. 6D'). These results are re-assuring as they demonstrate that approximating the branch lengths under the transform are good enough to approximate $\Delta\ell$ in real datasets. As is evident from Figure 6, the ARCSINE transform provides the best approximation, so we suggest this approximation is used to calculate marginal likelihoods for relaxed-clock models in very large datasets. We do not recommend using the NT method for any comparisons, or any of the approximate methods if comparisons with the strict clock are required.

Application of the approximation to select the rate model in large datasets.— dos Reis et al. (2018) and Barba-Montoya et al. (2018) used marginal likelihood to select the best rate model in Primate and flowering plant phylogenies respectively. Because those authors estimated the likelihood exactly, they were restricted to small datasets in their Bayesian rate model selection analyses: 10 species in dos Reis et al. (2018) and 10 species in Barba-Montoya et al. (2018), a dramatic subsampling of their large phylogenies with 372 and 644 species respectively. Thus, an open question remains, would have those authors selected the same rate model if all species had been analysed? Here we re-analysed the two phylogenies to select for the best rate model using all species and the approximate method. Table 2 shows the results. In the original analysis by dos Reis et al. (2018) the preferred rate model was GBM for the 10 species across five out of six alignment partitions, with the sixth partition, mitochondrial 3rd codon positions, supporting the STR model. This result is confirmed here for all 372 species, except that the 3rd codon positions support the GBM model in the full dataset (Table 2). Barba-Montoya et al. (2018) found best support for the ILN model among 10 species and across three alignment partitions. This result is confirmed here for all 644 species and all partitions (Table 2). Both large datasets appear very informative about the rate model, as the posterior probabilities for the best model are alwasy one in each case, as the marginal likelihood differences across models are very large (Table 2). Selecting the most appropriate rate model in these datasets appears important because the two relaxed rate models result in substantially different evolutonary timelines. For example, in the case of Primates, dos Reis et al. (2018) inferred an age for the crown group of 74.4 Ma (95% Credibility interval: 79.2–70.0 Ma) using the GBM model, while this was substantially older at 90.6 Ma (95% CI: 98.1–83.7 Ma) under the ILN model. Similar discrepancies in time estimates for the two rate models were seen by Barba-Montoya et al. (2018) in their flowering plant analyses. Finally, we note errors in marginal likelihood estimates due to MCMC sampling can be noticeable. For example, for the Primate mitochondrial 3rd codon position partition under the STR model, the estimated standard error (SE) of ℓ was 15.98 log-units (Table 2). This error is unimportant in this case as the log-likelihood difference between models is much larger (over 2,000 log-units between STR and GBM). However, an error of 1 log-unit would be catastrophic if, for example, the true posterior probability for the best model is around 90%.

Table 2: Bayesian rate model selection in large phylogenies using approximate likelihood

Dataset[a]	Partition[b]	$\ell_{\mathrm{STR}} \pm \mathrm{SE}^{c}$	$\ell_{\mathrm{ILN}} \pm \mathrm{SE}^{c}$	$\ell_{\mathrm{GBM}} \pm \mathrm{SE}^{c}$
Primates	Mit 1+2	-1249.87 ± 0.90	-980.28 ± 1.03	$\mathbf{-938.84 \pm 0.89}$
(372 species)	Mit 3	-3240.69 ± 15.98	-1303.65 ± 0.36	$\mathbf{-1187.9 \pm 0.64}$
	Mit RNA	-795.94 ± 0.16	-660.44 ± 0.37	$\mathbf{-638.33 \pm 0.62}$
	Nuclear 1+2	-1123.02 ± 0.57	-788.19 ± 0.57	$\mathbf{-730.76 \pm 0.70}$
	Nuclear 3	-1245.53 ± 0.81	-855.93 ± 0.63	$\mathbf{-773.40 \pm 0.61}$
	Nuclear NC	-1717.58 ± 6.38	-982.28 ± 0.55	$\mathbf{-892.60 \pm 0.84}$
Angiosperms	Plastid 1+2	-29877.16 ± 31.19	$\mathbf{-2877.76 \pm 0.72}$	-2958.10 ± 1.32
(644 species)	Mit 1+2	-4667.87 ± 2.50	$\mathbf{-1607.79 \pm 0.28}$	-1643.91 ± 0.52
	Nuclear RNA	-6838.14 ± 1.98	$\mathbf{-1697.54 \pm 0.25}$	-1763.13 ± 0.42

a. The Primates data is from dos Reis et al. (2018) and the Angiosperm (flowering plants) data from Barba-Montoya et al. (2018). The number of species indicated is the total number in the phylogeny, but the actual number in each partition may differ (see orignal publications for details). b. 1+2: 1st and 2nd codon positions, 3: 3rd codon positions. NC: non-coding. c. The ARCSIN transform and root calibration (0.999, 1.001) are used in all cases. The standard error (SE) for the log-marginal likelihood estimates is calculated using the stationary bootstrap (Politis and Romano, 1994; Álvarez-Carretero et al., 2022). The model with the largest marginal likelihood is shown in boldtype face. The posterior probability for the best model is 1 in all cases, as the marginal likelihood differences are very large.

As a conservative criterium, we recommend the sum of the SEs be much less than the absolute value of the difference in log-marginal likelihoods, i.e., $\mathrm{SE}_i + \mathrm{SE}_j \ll |\ell_i - \ell_j|$. If this is not the case, the SEs can be reduced by increasing the length of the MCMC and/or increasing the number of β points in the power posterior. A fourfold increase in the lenght of the MCMC decreases the SE by half.

Discussion

The posterior probability of the true model is guaranteed to converge to one as the amount of data goes to infinity, if the true model is among the models tested and if certain regularity conditions are met, such as data independence and parameter identifiability (e.g., Yang and Zhu, 2018). In Bayesian molecular clock dating with fossil calibrations, times and rates are non-identifiable and thus the conditions that guarantee convergence to the true model are not met. However, identifiability problems may be solved by finding functions of parameters (e.g., reparameterisations) that are identifiable (Rannala, 2002). In our case, by fixing the age of the root to one, we aim to make the rates and node ages identifiable, because we are reducing the degrees of freedom of the inference problem and thus we estimate relative instead of absolute

node ages. We note it appears no formal proof of identifiability of relative times and rates under relaxed rate models exists, but the evidence from the strict clock model and MCMC simulations with infinite sites in relaxed rate models indicate this should be the case (Rannala and Yang, 2007; Yang and Rannala, 2006). For the strict clock, Yang and Rannala (2006) show that when the number of sites goes to infinity, the posterior distribution of times converges to a one dimensional density and the uncertainty of node ages is proportional to the uncertainty of the root age (Eq. 21 in their paper). The collorary of Yang and Rannala (2006)'s proof is that if the age of the root is fixed to one, then the density must collapse to a point estimate for the infinite data, i.e., the relative node ages are fully identifiable in this case. Under relaxed clocks, the limiting density for an infinite number of loci and infinite number of sites is not available analytically, and thus Rannala and Yang (2007) used MCMC simulation to study the asymptotic behaviour of the distribution. Their results demonstrate the distribution of node ages also approaches a limiting distribution with node uncertainties proportional to the root age uncertainty. Thus, we suspect estimation of relative node ages is also fully identifiable under relaxed clocks, but a formal proof is still required.

Our simulations show that, when we don't use fossil calibrations and estimate relative ages, posterior rate model probabilities asymptotically converge to one for the correct model as expected by theory. Thus, we recommend phylogeneticists interested in selecting the best rate model for their datasets use this approach. As we show in our simulations, rate model selection with fossil calibrations is not reliable. If the fossils are slightly misspecified, posterior model probabilities are well behaved, and converge to one for the correct model, but if they are badly misspecified, posterior model probabilities are spurious and a completely wrong rate model may be selected. In real datasets, we do not know if fossil calibrations are all good, slightly misspecified, or badly misspecified, and thus there is no guarantee we will be choosing the right model. This is a direct consequence of carrying out Bayesian model selection on non-identifiable parameters (absolute times and rates).

A limiting factor in the application of Bayesian rate model selection to real datasets has been the extreme computational expense of calculating the likelihood for large phylogenies during MCMC sampling across power posteriors. This means that relatively few studies have

calculated posterior probabilities for rate models, and in all cases the datasets analysed have been small, with few species and/or few sites (Linder et al., 2011; Lepage et al., 2007; Baele et al., 2012; dos Reis et al., 2018; Barba-Montoya et al., 2018; McGowen et al., 2020; Álvarez-Carretero et al., 2022). In previous studies we had advised against using the approximate method for estimating marginal likelihoods due to concerns about the accuracy of the approximation in the tails of the likelihood function (dos Reis et al., 2018; Álvarez-Carretero et al., 2022). However, here we show that those concerns were too conservative, and the approximations under the ARCSINE and SQRT transform are good enough to provide reliable estimates of posterior model probabilities. Calculation speed-ups under the approximation can be quite dramatic, for example, a dataset that may require one day of MCMC sampling under the approximation, may required nearly three years under the exact method, given the 1000x speed-up reported in a benchmarking study (Battistuzzi et al., 2011). We recommend phylogeneticists interested in testing for rate models in large datasets use the approximation under the ARCSINE transform, as this provides the best approximation for the long tail of the likelihood (Fig. 5). This may be accompanied by sanity tests using the exact method on a smaller subsample of species from the phylogeny. However, we note the approximate method must not be used if the strict clock is among the models being compared or if extremely long branches are present in the phylogeny.

Finally, we would like to point out that the approximate method may have applications beyond rate model seelction. For example, Fourment et al. (2020) explored the use of similar approximations to estimate the marginal likelihood and select the best tree topology among a candidate set of trees (for example, among the set of n trees with highest likelihood from a maximum likelihood analysis). Thus, our results here suggest the approximation currently implemented in MCMCtree can also be used to estimate posterior model probabilities among tree topologies in very large phylogenomic datasets, datasets for which traditional MCMC sampling with tree moves would be computationally prohibitive: We may select the top n topologies from a maximum likelihood analysis with, say with RAxML or IQ-TREE, then use MCMCtree with the approximate method to estimate the corresponding marginal likelihoods for each tree topology and calculate their posterior probilities. This would make Bayesian topology estimation for

phylogenomic datasets with hundreds of taxa and millions of sites within each.

Supplementary Material

The Primate and Angiosperm datasets analysed here are available from their original data repositories at http://dx.doi.org/10.5061/dryad.c020q and https://figshare.com/s/404b70bc39656c2cf57e respectively. The `simclock` R package, to simulate relaxed rate models on phylogenies, is available from `https://github.com/dosreislab/simclock`. The `mcmc3r` R package is available from `https://github.com/dosreislab/mcmc3r`. This package can be used to select β values, and prepare MCMCtree control files, to estimate marginal likelihoods unsing the stepping-stones (Xie et al., 2011) and thermodynamic integration with Gaussian quadrature (Rannala and Yang, 2017) methods. The package also allows calculation of the standard error of marginal likelihood estimates using analytical formula (Xie et al., 2011; dos Reis et al., 2018), and the stationary bootstrap (Álvarez-Carretero et al., 2022).

Conflicts of Interest

The authors declare no conflicts of interest.

References

Álvarez-Carretero, S., A. U. Tamuri, M. Battini, F. F. Nascimento, E. Carlisle, R. J. Asher, Z. Yang, P. C. J. Donoghue, and M. dos Reis. 2022. A species-level timeline of mammal evolution integrating phylogenomic data. Nature 602:263–267.

Baele, G., W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2012. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular biology and evolution 30:239–243.

Barba-Montoya, J., M. Dos Reis, H. Schneider, P. C. Donoghue, and Z. Yang. 2018. Constrain-

ing uncertainty in the timescale of angiosperm evolution and the veracity of a cretaceous terrestrial revolution. New Phytologist 218:819–834.

Battistuzzi, F., P. Billing-Ross, A. Paliwal, and S. Kumar. 2011. Fast and slow implementations of relaxed clock methods show similar patterns of accuracy in estimating divergence times. Mol Biol Evol 28:2439–2442.

dos Reis, M., G. F. Gunnell, J. Barba-Montoya, A. Wilkins, Z. Yang, and A. D. Yoder. 2018. Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: primates as a test case. Systematic Biology 67:594–615.

dos Reis, M. and Z. Yang. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol 28:2161–72.

dos Reis, M., T. Zhu, and Z. Yang. 2014. The impact of the rate prior on bayesian estimation of divergence times with multiple loci. Systematic biology 63:555–565.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88.

Drummond, A. J. and M. A. Suchard. 2010. Bayesian random local clocks, or one rate to rule them all. BMC biology 8:1–12.

Fourment, M., A. F. Magee, C. Whidden, A. Bilge, F. A. Matsen IV, and V. N. Minin. 2020. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. Systematic biology 69:209–220.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–74.

Heath, T. A., M. T. Holder, and J. P. Huelsenbeck. 2012. A dirichlet process prior for estimating lineage-specific substitution rates. Mol Biol Evol 29:939–55.

Ho, S. Y. and S. Duchêne. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular ecology 23:5947–5965.

Ho, S. Y., S. Duchêne, and D. Duchêne. 2015. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. Molecular ecology resources 15:688–696.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879–92.

Jukes, T. H. and C. R. Cantor. 1969. Mammalian protein metabolism chap. Evolution of protein molecules, Pages 21–123. Academic Press, New York.

Lartillot, N., M. J. Phillips, and F. Ronquist. 2016. A mixed relaxed clock model. Philosophical Transactions of the Royal Society B: Biological Sciences 371:20150132.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Mol Biol Evol 24:2669–80.

Linder, M., T. Britton, and B. Sennblad. 2011. Evaluation of Bayesian models of substitution rate evolution–parental guidance versus mutual independence. Syst Biol 60:329–42.

McGowen, M. R., G. Tsagkogeorga, S. Álvarez-Carretero, M. Dos Reis, M. Struebig, R. Deaville, P. D. Jepson, S. Jarman, A. Polanowski, P. A. Morin, et al. 2020. Phylogenomic resolution of the cetacean tree of life using target sequence capture. Systematic Biology 69:479–501.

Politis, D. N. and J. P. Romano. 1994. The stationary bootstrap. Journal of the American Statistical association 89:1303–1313.

Rannala, B. 2002. Identifiability of parameters in MCMC bayesian inference of phylogeny. Systematic biology 51:754–760.

Rannala, B. and Z. Yang. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol 56:453–66.

Rannala, B. and Z. Yang. 2017. Efficient bayesian species tree inference under the multispecies coalescent. Systematic biology 66:823–842.

Seo, T. K., H. Kishino, and J. L. Thorne. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Mol Biol Evol 21:1201–13.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647–57.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic biology 60:150–160.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–14.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–91.

Yang, Z. and B. Rannala. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23:212–26.

Yang, Z. and T. Zhu. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proceedings of the National Academy of Sciences 115:1854–1859.

Zuckerkandl, E. and L. Pauling. 1965. Evolving Genes and Proteins chap. Evolutionary Divergence and Convergence in Proteins, Pages 97–166. Academic Press.