Paleobiology, 48(2), 2022, pp. 340–352
DOI: 10.1017/pab.2021.42

PALEOBIOLOGY
A PUBLICATION OF THE Paleontological SOCIETY

Article

Selecting and averaging relaxed clock models in Bayesian tip dating of Mesozoic birds

*Chi Zhang**

Abstract.—Relaxed clock models are fundamental in Bayesian clock dating, but a single distribution characterizing the clock variation is typically selected. Hence, I developed a new reversible-jump Markov chain Monte Carlo (rjMCMC) algorithm for drawing posterior samples between the independent lognormal (ILN) and independent gamma rates (IGR) clock models. The ability of the rjMCMC algorithm to infer the true model was verified through simulations. I then applied the algorithm to the Mesozoic bird data previously analyzed under the white noise (WN) clock model. In comparison, averaging over the ILN and IGR models provided more reliable estimates of the divergence times and evolutionary rates. The ILN model showed slightly better fit than the IGR model and much better fit than the autocorrelated lognormal (ALN) clock model. When the data were partitioned, different partitions showed heterogeneous model fit for ILN and IGR clocks. The implementation provides a general framework for selecting and averaging relaxed clock models in Bayesian dating analyses.

Chi Zhang. Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, and Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing 100044, China. E-mail: zhangchi@ivpp.ac.cn

Accepted: 9 November 2021
*Corresponding author.

Introduction

Since the proposal of a molecular evolutionary clock by Zuckerkandl and Pauling (1965), the clock assumption has provided a fundamental component to dating evolutionary events. The early hypothesis of a strict molecular clock, in which the evolutionary rate is constant over time and across taxa (Zuckerkandl and Pauling 1965), was soon proven to only hold for closely related taxa. To account for the violation of the molecular clock, several relaxed clock models were proposed (Kishino and Hasegawa 1990; Huelsenbeck et al. 2000; Yoder and Yang 2000; Kishino et al. 2001; Thorne and Kishino 2002; Drummond et al. 2006; Lepage et al. 2007; Rannala and Yang 2007), and have been widely used in phylogenetics for estimating divergence times and evolutionary rates (Heath and Moore 2014; Ho and Duchêne 2014; dos Reis et al. 2016; Ho 2021). In paleobiology, the analogical term for "molecular clock" is "morphological clock" (Lee et al. 2014; Lee 2016; Warnock and Wright 2021), wherein the model describes the pattern of morphological character changes instead of nucleotide or amino acid substitutions, but the mathematical assumptions are essentially unchanged.

Relaxed clock models can be loosely divided into two families. In one family, the evolutionary rate is assumed to be independent among the branches in the tree, and these rates are commonly drawn from independent lognormal (ILN) or independent gamma (IGR) distributions (or exponential distribution, which is a special case of gamma distribution when the shape $\alpha = 1$) (Drummond et al. 2006). Slightly different from the IGR model, the white noise (WN) model (Lepage et al. 2007) assumes the variances of the gamma distributions are not identical but proportional to the branch lengths. In the other family, the

0094-8373/22

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press

evolutionary rate is autocorrelated, that is, the rate of a descendant branch is drawn from a distribution, with the mean depending on the rate of the ancestor's branch. In the autocorrelated lognormal (ALN) clock model (Kishino et al. 2001; Thorne and Kishino 2002), the rates are lognormally distributed, with the variances proportional to the branch lengths.

Typically, the distribution characterizing the overall shape of the clock variation is preselected, for example, as a lognormal or gamma distribution, which is considered flexible enough to capture the variation. Testing the fit of clock models to the data is usually performed separately (Li and Drummond 2012; Baele et al. 2013).

In Bayesian statistics, model selection and model averaging are common techniques for such a purpose. The former estimates the marginal likelihood of each model and uses them to calculate Bayes factors that can ultimately be used to decide which model best fits the data (Kass and Raftery 1995). The latter employs the reversible-jump Markov chain Monte Carlo (rjMCMC) algorithm to move among the alternative models and estimates the posterior probability of each model (Green 1995). Model probabilities are directly provided by the rjMCMC but can also be estimated through the marginal likelihoods. Note that the ratio of posterior model probabilities ($P(M_1 \mid D)/P(M_2 \mid D)$, M for model and D for data) equals the prior odds ($P(M_1)/P(M_2)$) multiplied by the ratio of marginal likelihoods (i.e., the Bayes factor, $P(D \mid M_1)/P(D \mid M_2)$). With no prior preference favoring one model over the other ($P(M_1) = P(M_2)$), the probability of M_1 is $P(M_1 \mid D) = P(D \mid M_1)/[P(D \mid M_1) + P(D \mid M_2)]$ and $P(M_2 \mid D) = 1 - P(M_1 \mid D)$.

In theory, the marginal-likelihood and rjMCMC approaches should produce identical model probabilities. In practice, however, the performance of different estimators may vary. Studies have shown that the harmonic mean estimator has very poor performance and is often biased, while more advanced techniques such as path sampling (PS) and stepping-stone sampling (SS) greatly improve the performance of marginal-likelihood estimation (Lartillot and Philippe 2006; Xie et al. 2011; Baele et al. 2012, 2013). Despite being computationally demanding, PS and SS are being widely used in Bayesian model selection. On the other hand, rjMCMC usually takes much less computational cost to obtain similar or better accuracy in terms of estimating model probabilities and has the advantage of comparing more than two models simultaneously and averaging the uncertainties among them (Baele et al. 2013). The rjMCMC algorithm does require careful design to achieve good mixing, and it is challenging to move among very distinct models.

Here I focus on the uncorrelated relaxed clock models and develop a new rjMCMC algorithm for drawing posterior samples between the ILN and IGR models. Previous studies have attempted averaging over the independent exponential rates and ILN models using a different approach, that is, by mapping the quantiles of the two distributions (Li and Drummond 2012). Such mapping involves recalculating the data likelihood, which is likely computationally expensive. I propose a direct match of the branch rates, which is computationally fast while still maintaining good mixing.

I verify the ability of the new rjMCMC algorithm to infer the true model through simulations. The algorithm is then applied to the morphological data of Mesozoic birds (Zhang and Wang 2019) to average over the relaxed clock models while estimating the divergence times and evolutionary rates. A previous study (Zhang and Wang 2019) suffered from poor convergence and mixing under the WN model when the data were partitioned. Also, the age estimates were not very consistent between the unpartitioned and partitioned analyses, and the credibility intervals of the evolutionary rates appeared too wide. I will show that these aspects are improved by mixing the ILN and IGR clock models.

Methods

The rjMCMC Algorithm.—The evolutionary rate at branch i, c_i, is a product of the mean (base) rate, c, and the relative rate, r_i. The ILN and IGR models differ in the probability distribution of r_i. In the ILN model, r_i follows a lognormal distribution with a mean of 1.0 and variance of σ_L; whereas in the IGR model, r_i

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press

follows a gamma distribution with a mean of 1.0 and variance of σ_G. The similarity of the two models provides a simple mapping of the parameters, that is, direct matching of the branch rates and linear mapping of the variances, $\sigma_L = w\sigma_G$, for jumping between the two models.

The key component of the rjMCMC algorithm is calculating the acceptance probability when moving from one model to the other. This probability contains three multiplied parts: the posterior ratio, the generating-variable density ratio, and the Jacobian determinant for transforming variables (Green 1995). The posterior ratio is the product of the prior ratio and the likelihood ratio. Because the move does not change the tree, including branch lengths, the likelihood ratio is 1.0. The prior ratio involves the gamma rate densities multiplied across all branches over the lognormal rate densities moving from IGR to ILN, and the reciprocal when moving backward. The generating-variable density ratio is also 1.0, as the rates are directly mapped, thus no random variable is generated. Finally, the Jacobian determinant is w moving from IGR to ILN and is $1/w$ moving backward, where w is the ratio of the ILN and IGR clock variances.

Unlike conventional MCMC proposals in which the best efficiency is typically achieved when the acceptance rate is about 30% (Yang and Rodriguez 2013), for cross-model rjMCMC proposals, in principle, one should make the acceptance rate as high as possible (by adjusting w in this case) to maximize the proposal's efficiency (Yang 2014). But the acceptance rate by its nature cannot reach an arbitrarily high value, for example, it cannot be higher than 20% if one model has posterior probability 0.9, as the chain has to stay in that model 90% of the time. In the current implementation, adjusting w in the proposal is the only option to increase efficiency. One could preselect several values of w (typically ranging from 1 to 4) for independent runs and use the results with the highest acceptance ratio. This helps to confirm convergence and consistency among runs but requires more computation. Here I introduce an auto-tuning feature to dynamically adjust w during the rjMCMC to avoid the trial and error for picking the appropriate value.

The starting value of w is set to 2.0, which can be changed by the user. When the rjMCMC proposal has been attempted for a certain number of generations (called a batch, default to 1000), the auto-tuner adjusts w by comparing the acceptance rates of this batch and the previous batch, aiming to achieve higher acceptance rate in the next batch. The adjustment amount is $\delta = \min(0.1,\ 1/\sqrt{n})$, where n is the current batch number. Thus, w increases or decreases by the amount of 0.1 in each of the first 100 batches, then changes more and more gradually in the following batches. When the acceptance rates are both zero after two adjacent batches, w is reset to its initial value to continue, such that the tuning mechanism can avoid getting stuck.

The rjMCMC algorithm was implemented in MrBayes software (Ronquist et al. 2012b) (see "Software Implementation" for more details).

Simulation Study.—To verify the implementation and performance of the rjMCMC algorithm, I simulated evolutionary rate variation along tree branches and ran MrBayes to infer the posterior probability of the clock models under each condition.

Specifically, I first generated 100 trees under the birth–death model with speciation rate $\lambda = 3.0$, extinction rate $\mu = 1.0$, time of the most recent common ancestor $t_{\mathrm{mrca}} = 1.0$, and complete sampling of the lineages, using the TreeSim package in R (Stadler 2011). These trees have numbers of tips ranging from 6 to 75. The branch lengths were measured by the expected number of substitutions per site (character) under the base rate of the clock. Given each tree, the relative branch rates (r_i values) were all drawn from a lognormal distribution with a mean of 1.0 and variance of $\sigma_L = 0.1$, 0.5, 1.0, 2.0, respectively; or from a gamma distribution with a mean of 1.0 and variance of $\sigma_G = 0.1, 0.5, 1.0, 2.0$, respectively. Note when $\sigma_G = 1.0$, the gamma distribution is actually an exponential distribution. I also tested a simple model mismatch, that is, the relative rates were drawn from a normal distribution (truncated at 0) with a mean of 1.0 and variance of $\sigma_N = 1.0$. Eventually, each tree was associated with nine sets of branch rates (100 × 9 combinations in total). The probability densities of these distributions are shown in Figure 1.

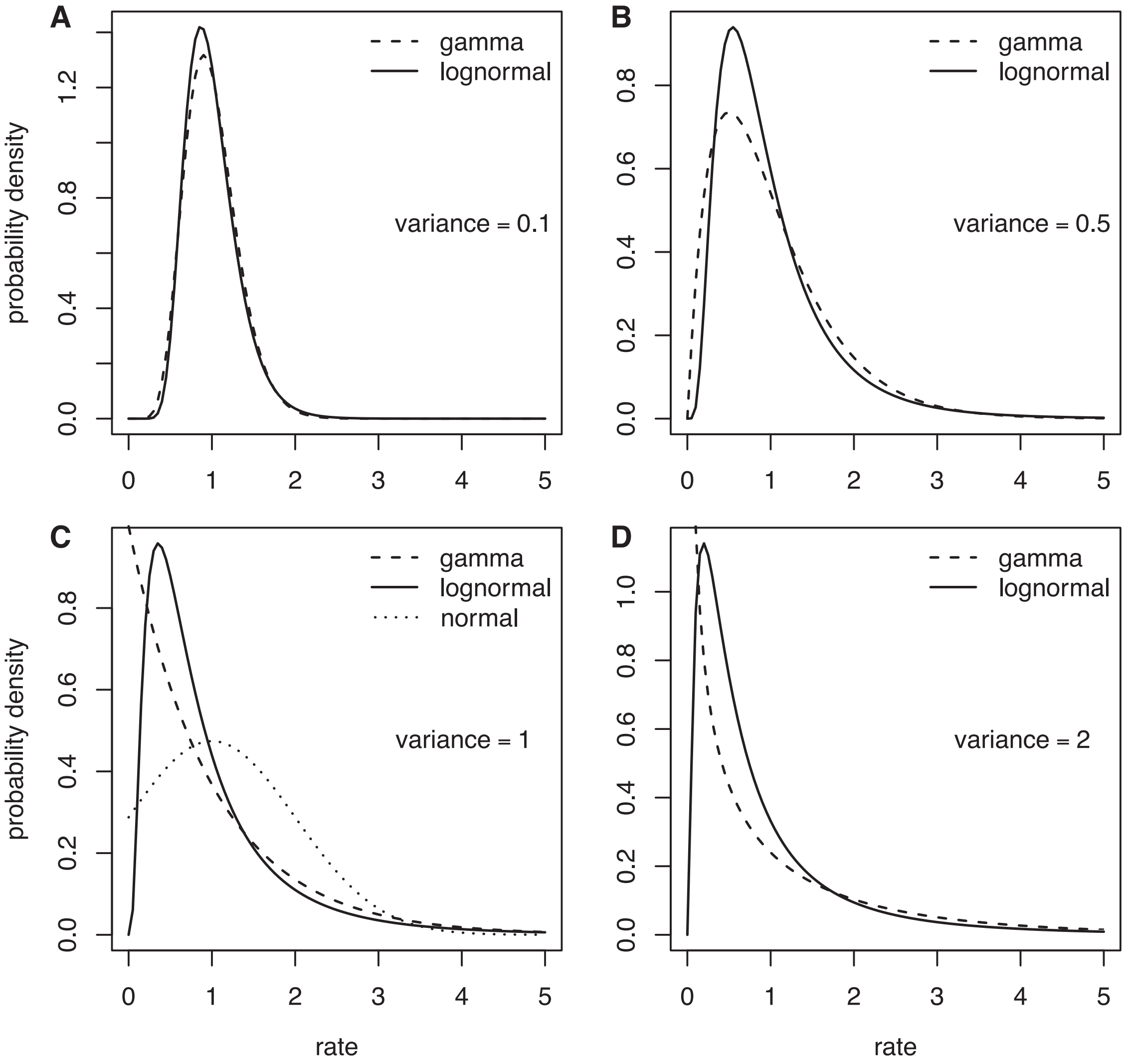

FIGURE 1. Probability densities of gamma and lognormal distributions under mean 1.0 and variance 0.1 (A), 0.5 (B), 1.0 (C), and 2.0 (D), and probability density of $N(1, 1)$ distribution (truncated at 0) (C). Note that the Gamma(1, 1) distribution ($\alpha = 1$) is Exp(1) distribution (C).

I initially fixed the tree topology, branch lengths, and relative rates to the simulated values, so that MrBayes only estimated the probabilities of the clock models and the variance parameter in each model. No data were used at this point (sampling from the prior). The aim was to verify the correctness of the rjMCMC implementation and to avoid introducing uncertainties from irrelevant sources. For each run, the ILN and IGR models were assigned equal prior probabilities (0.5). The base evolutionary rate (c) was assigned a diffuse prior of Exp(1.0) and the prior for the clock variance (σ_L or σ_G) was also Exp(1.0). The tuning parameter w was adjusted through auto-tuning. A single MCMC per replicate was executed for 1 million iterations and sampled every 100 iterations; this setting was determined from preliminary runs. The first 25% of the samples were discarded as burn-in. Convergence was checked across the 100 replicates by examining the traces of parameters and the effective sample sizes (ESS) all higher than 200 (same for the MCMC runs below).

Additionally, I simulated discrete morphological matrices under the Markov *k*-states variable (Mkv) model (Lewis 2001), that is, keeping only the variable characters in the matrices, on each of the 100 trees with each of the nine sets of rates generated earlier. There were 100, 300, and 500 characters, respectively, for each setting (100 × 9 × 3 data sets generated). Each data matrix contained from binary to up to five-state characters with proportions of 0.4, 0.3, 0.2 and 0.1, respectively. This round of simulations was intended to verify the ability of the program to estimate the clock model probabilities along with other parameters (including tree topology) directly from the morphological data. In the inference, I used also the Mkv model consistent with the model used to simulate the data. The prior for the tree was uniform (Ronquist et al. 2012a) and that for the root age was gamma with a mean of 1.0 and a standard deviation of 0.1. The tip ages were fixed to their true ages, and these were treated as part of the data in the tip-dating analyses. The other prior settings were the same as before. The length of the MCMC was extended to 50 million iterations to ensure converge and sufficient ESS.

Apart from the clock models, the evolutionary rates estimated from data are usually of interest. Hence, I calculated the mean squared error to assess the accuracy with which the evolutionary rates are estimated. It is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_i)^2,$$

where r_i is the true rate in the simulation, $\hat{r}_i$ is the estimated value in the inference, and n is the number of branches (rates) in the tree. To match the simulated and estimated rates, the tree topology in the inference was fixed to the one used to simulate the data. In addition to using the rjMCMC to average the ILN and IGR models, I tested using the WN model in the inference that does not match either of the models used to simulate the data.

Mesozoic Birds.—The morphological data of Mesozoic birds (68 species and 280 characters; Zhang and Wang 2019) were used to investigate the performance of the rjMCMC algorithm under the tip-dating framework. The fossilized birth–death (FBD) model (Stadler 2010; Gavryushkina et al. 2014; Heath et al. 2014; Zhang et al. 2016) was used for the tree prior. To account for nonuniform fossil sampling, the sampling rate (ψ) was allowed to vary along time in a piecewise manner with four intervals divided at 145, 100, and 66 Ma. The speciation rate (λ) and extinction rate (μ) were assumed to be constant in the model. The prior for the base evolutionary rate (*c*) was Gamma(2, 100) with a mean of 0.02 (about 2 changes per character per 100 Myr) and a standard deviation of 0.014, which is weakly informative based on the inference results reported by Zhang and Wang (2019). The ILN and IGR clock models were given equal prior probabilities, and the variance parameter (σ_L or σ_G) was assigned Exp(0.5) prior. Two independent runs were executed for 70 million iterations each and sampled every 2000 iterations. The first 25% of the samples were discarded as burn-in, and the remaining samples from the two runs were combined after checking consistency between runs. The efficiency of the rjMCMC algorithm was tested under $w = 1$, 2, 3, and 4, respectively, to compare with auto-tuning w.

The rjMCMC algorithm has two relaxed clock models implemented: the ILN and the IGR. Nevertheless, I also ran an additional analysis under the ALN model to assess the fitness of this model if compared with the analyses run under the two uncorrelated-rates models implemented in the new algorithm. For this model-selection analysis, I computed the marginal likelihoods under these three models using the SS approach (Xie et al. 2011), which I would later use to find out the best-fitting model. The priors for the base evolutionary rate and the variance parameter were the same in these models (i.e., Gamma(2, 100) for *c* and Exp(0.5) for σ). For each clock model, a total of 50 steps were used with 40 million iterations (20,000 samples) within each step. The first step was discarded as initial burn-in. Additionally, the beginning of each step (10 million iterations, 5000 samples) was discarded as burn-in. The sampling was from posterior to prior with the power drawing from a beta(0.4, 1) distribution in the default implementation.

I further applied the rjMCMC algorithm to the characters divided into six anatomical

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press

partitions, as in the previous study, which are skull (53 characters), axial skeleton (36 characters), pectoral girdle and sternum (48 characters), forelimb (65 characters), pelvic girdle (23 characters), and hindlimb (55 characters) (Zhang and Wang 2019). In this case, the program was able to infer the probabilities of the relaxed clock models (ILN vs. IGR) for each partition together with the evolutionary rate variations across partitions. The prior and MCMC settings were the same as in the unpartitioned analysis, except that the chain length was set to 120 million iterations and the first 40% samples were discarded as burn-in. In the previous study (Zhang and Wang 2019), the MCMC was unable to converge using the WN relaxed clock and standard FBD models (e.g., low ESS values and segregated estimates of several key parameters), as a trade-off, fossil ancestors had to be disallowed. In comparison, the inference using the ILN and IGR mixed clock models for the partitioned data were executed under the standard FBD model with fossil ancestors.

Results

Simulation Study.—As the number of characters goes from small ($l = 100$) through medium ($l = 300$) to large ($l = 500$), the data should carry more and more information to inform the evolutionary rate variation and the true model generating the rates. Having the tree and rates fixed in the inference can be viewed as a limiting condition with an infinite number of characters and a very informative prior for the times; thus, the times and rates can be estimated without error. The results indeed agree with this. Under the same clock variance, the probability of the true model increases along with the number of characters (Fig. 2, except for normal distribution). Note that the model used in the inference is not misspecified, that is, the models in the rjMCMC include the true model generating the data.

When the branch rates were generated from a normal distribution, neither ILN nor IGR matched the generating model, but the IGR model dominated the posterior. In the limiting condition when the rates were fixed to the simulated values, the probability of ILN approached zero (Fig. 2D). It appears that the $N(1,\ 1)$ distribution can be better fit by a gamma distribution with variance slightly smaller than 1.0 than any lognormal distribution that has a sharp peak and long tail.

Given the same amount of data (characters), the ability to infer the true clock model depends on the clock variance. When the variance of the branch rates is small (0.1), the shape of the two distributions is quite similar (Fig. 1A), thus the ILN and IGR models fit the data (rates) almost equally well. As a result, the posterior probability of the true model is close to 0.5 (Fig. 2). As the variance increases, the shapes of the two distributions becomes sharply different (Fig. 1B–D), making it easier for the rjMCMC to distinguish the two models (Fig. 2). Interestingly, the lognormal rate variation is slightly more easily distinguished when the variance is small ($\sigma = 0.1$ or 0.5), but the opposite is true when the variance is large ($\sigma = 2.0$).

The evolutionary rates are estimated more accurately when the variance is small (0.1), while the MSEs become one order of magnitude larger when the variance is large (2.0) (Table 1). The estimates are slightly improved by adding more characters, but the effect can be overridden by increased clock variance, for example, the median MSE is larger using 500 characters for variance of 1.0 than using 100 characters for variance of 0.5 (Table 1). When the rates were generated from the normal distribution, the rjMCMC algorithm does not include the true model, but the values of MSE are comparable to those when the rates were generated from gamma or lognormal distributions with variance of 1.0. This type of model mismatch on the rate estimates is less dramatic than using the WN model for inference (Table 1). The WN model has nonidentical variances for the branch rates and thus resulted in worse accuracy when the rates were generated from independent and identical distributions.

Mesozoic Birds.—The time-scaled phylogeny and evolutionary rates inferred are largely consistent with the previous results (Fig. 3, cf. Zhang and Wang 2019: fig. 1). But in detail, averaging over the ILN and IGR models refines both the divergence time and evolutionary rate estimates. As shown in Figure 4, the ages of five key nodes in early stem birds are quite consistent between the unpartitioned and partitioned

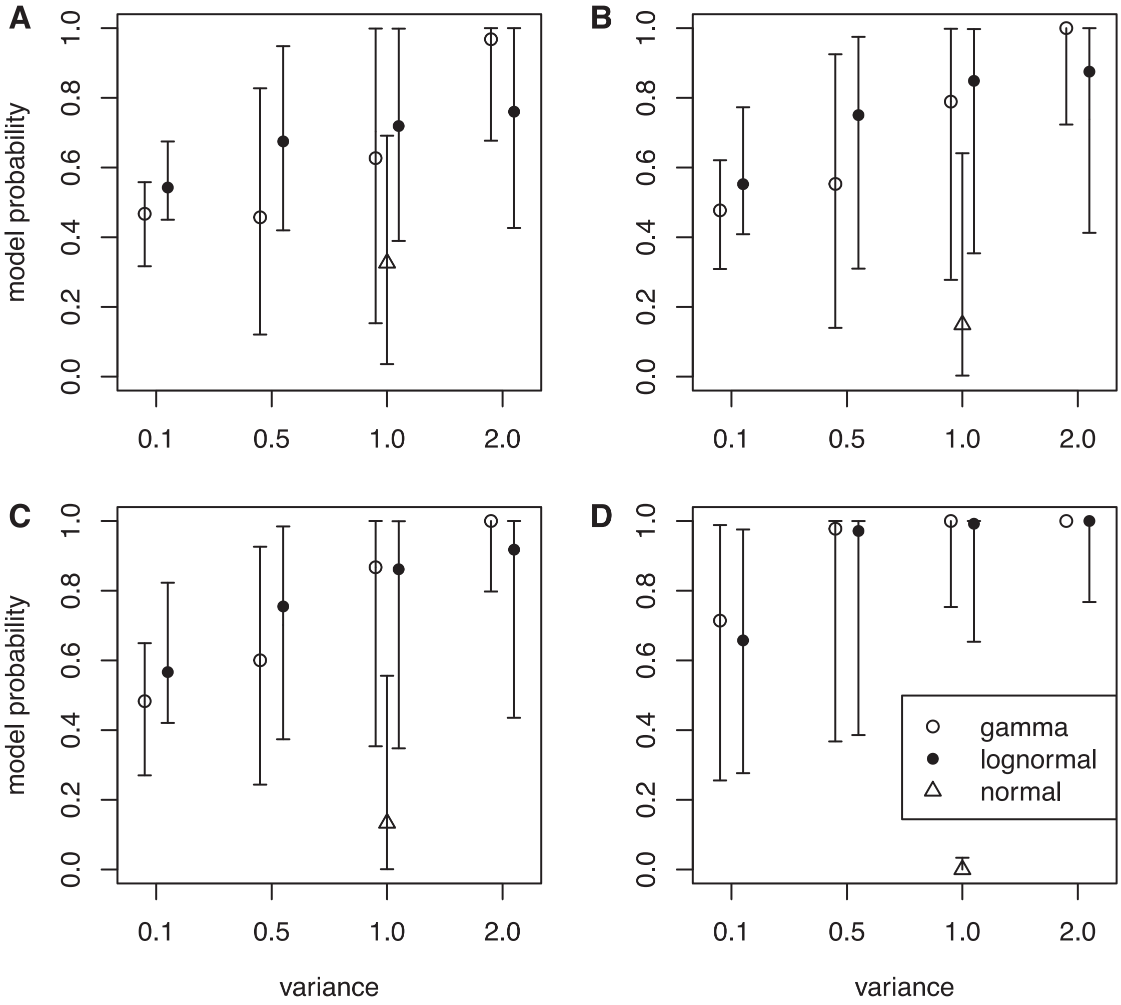

FIGURE 2. Model probabilities estimated by the reversible-jump Markov chain Monte Carlo (rjMCMC) algorithm. Nine sets of rates were simulated under distributions in Fig. 1 given each tree, and for each tree with rates, data matrices with 100 characters (A), 300 characters (B), and 500 characters (C) were simulated for inference. The model probabilities were also inferred when the trees and rates were fixed to the simulated values (no data) (D). When the rate-generating distribution was gamma, the posterior probabilities of the independent gamma rates (IGR) model are shown; when the generating distributions were lognormal and normal, the posterior probabilities of the independent lognormal (ILN) model are shown. The circle is the median, and the error bar denotes the 5th and 95th percentiles summarized across the 100 replicates (trees).

analyses, but previous ones had differences of at least 5 Myr. The variances of the branch rates are also reduced, resulting in narrower credibility intervals for drastic rates than those under the WN model. For example, the 95% highest posterior density (HPD) intervals are (0.08, 7.9) and (0.96, 9.8) at branches subtending Ornithothoraces and Enantiornithes under the mixed model, but are (0.0, 32.5) and (1.49, 30.8) under the WN model. The latter intervals are much wider given the same amount of data, indicating the WN model favors larger variance (or more heterogeneous rates).

The rjMCMC algorithm consistently estimated the posterior probability of the ILN model as $P_{\mathrm{ILN}} = 0.6$ (and $P_{\mathrm{IGR}} = 0.4$). Judging by the acceptance rate of the move, the best efficiency was achieved when w was at 2.0 or 3.0 (Table 2). This is obvious when looking at the estimates of σ_L and σ_G, which are 2.5 (1.1, 4.4) and 0.9 (0.6, 1.3), respectively (mean and 95% HPD interval). The lognormal distribution is

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press

TABLE 1. Median (5th, 95th percentiles) of the mean squared errors (MSE) of relative rates across the 100 replicates (trees). For each tree, there were nine sets of rates (σ_G for gamma variance, σ_L for lognormal variance, and σ_N for normal variance) and three character lengths ($l = 100, 300$, and 500) simulated (100 × 9 × 3 datasets). Each dataset was then analyzed using the mixed independent lognormal (ILN) and independent gamma rates (IGR) clock models and the white noise (WN) clock model.

Variance	$l = 100$	$l = 300$	$l = 500$
		Mixed ILN and IGR clocks	
$\sigma_G = 0.1$	0.08 (0.06, 0.12)	0.07 (0.05, 0.10)	0.06 (0.04, 0.09)
$\sigma_G = 0.5$	0.36 (0.21, 0.52)	0.32 (0.17, 0.47)	0.29 (0.16, 0.44)
$\sigma_G = 1.0$	0.57 (0.33, 0.96)	0.51 (0.29, 0.89)	0.47 (0.24, 0.80)
$\sigma_G = 2.0$	0.99 (0.46, 1.87)	0.90 (0.44, 1.58)	0.85 (0.41, 1.52)
$\sigma_L = 0.1$	0.08 (0.06, 0.12)	0.07 (0.05, 0.11)	0.06 (0.04, 0.09)
$\sigma_L = 0.5$	0.32 (0.19, 0.53)	0.27 (0.15, 0.50)	0.24 (0.14, 0.45)
$\sigma_L = 1.0$	0.54 (0.26, 1.55)	0.45 (0.21, 0.99)	0.42 (0.18, 0.94)
$\sigma_L = 2.0$	0.92 (0.38, 2.81)	0.71 (0.33, 2.00)	0.68 (0.28, 1.95)
$\sigma_N = 1.0$	0.58 (0.41, 0.84)	0.52 (0.38, 0.79)	0.50 (0.34, 0.72)
		WN clock	
$\sigma_G = 0.1$	0.12 (0.07, 0.18)	0.09 (0.06, 0.15)	0.08 (0.04, 0.13)
$\sigma_G = 0.5$	0.47 (0.30, 0.76)	0.42 (0.26, 0.69)	0.40 (0.23, 0.64)
$\sigma_G = 1.0$	0.84 (0.48, 1.77)	0.79 (0.41, 1.39)	0.75 (0.38, 1.31)
$\sigma_G = 2.0$	1.46 (0.62, 3.08)	1.21 (0.56, 2.63)	1.15 (0.53, 2.32)
$\sigma_L = 0.1$	0.11 (0.07, 0.19)	0.09 (0.06, 0.16)	0.08 (0.04, 0.13)
$\sigma_L = 0.5$	0.41 (0.25, 0.70)	0.35 (0.19, 0.68)	0.32 (0.17, 0.65)
$\sigma_L = 1.0$	0.77 (0.38, 1.65)	0.70 (0.33, 1.25)	0.66 (0.30, 1.16)
$\sigma_L = 2.0$	1.22 (0.51, 2.72)	1.14 (0.45, 2.23)	1.10 (0.41, 2.11)
$\sigma_N = 1.0$	0.79 (0.47, 1.71)	0.71 (0.39, 1.36)	0.65 (0.38, 1.29)

more skewed than the gamma distribution under the estimated variances, meaning that it can better capture some extreme rates. Smaller and larger values of w (0.5 and 5) were also tried but resulted in much poorer convergence and mixing; thus, the results were discarded. With auto-tuning enabled, the values of w fluctuated mostly between 2.0 and 3.0 during the run, and the efficiency was similar to when w was fixed to 2.0 or 3.0 (Table 2).

The rjMCMC algorithm does not include an autocorrelated relaxed clock model (e.g., the ALN model). In fact, it is quite challenging to do that (see "Discussion"). On the other hand, it is straightforward to compare these models using marginal likelihoods or Bayes factors. The SS ran about 20 times slower than the rjMCMC (70 million vs. 2 billion iterations). The marginal likelihoods (natural logarithms) were −4788.1 and −4788.9 in two runs under the ILN model and −4789.2 and −4789.9 under the IGR model. Given equal prior probabilities for the two models, the posterior probability of the ILN model, $P_{ILN} = M_{ILN}/(M_{ILN} + M_{IGR})$, ranges from 0.57 to 0.86. The SS had slightly worse convergence under the ALN model, resulting in marginal likelihoods of −4882.3 and −4884.8 in two runs. It has been shown that convergence is much more difficult to reach under the ALN model due to the lack of power to detect the autocorrelation (Ho et al. 2015). Nevertheless, the natural logarithm of the Bayes factor (difference of logarithmic marginal likelihoods) is much larger than 10 for IGR against ALN, indicating IGR fits the data much better than ALN (Kass and Raftery 1995). The independent-rates models (ILN or IGR) allow drastic rate variation among adjacent branches, whereas the ALN model has a smoothing effect and thus has less tolerance for drastic changes at adjacent branches. A study of Hymenoptera (Ronquist et al. 2012a), for example, also detected that the independent rates (WN) fit the data better than the autocorrelated rates (ALN) for the same reason.

When the characters were partitioned according to six anatomical regions, the relative rates at five focal branches among these regions were consistent with previous estimates in general, with slightly narrower credibility intervals (Fig. 5, cf. Zhang and Wang 2019: fig. 2). Moreover, the rjMCMC algorithm also estimated the posterior probabilities of the clock models for each partition (Table 3). The third partition, which contains pectoral girdle and sternum characters, has the highest probability for the

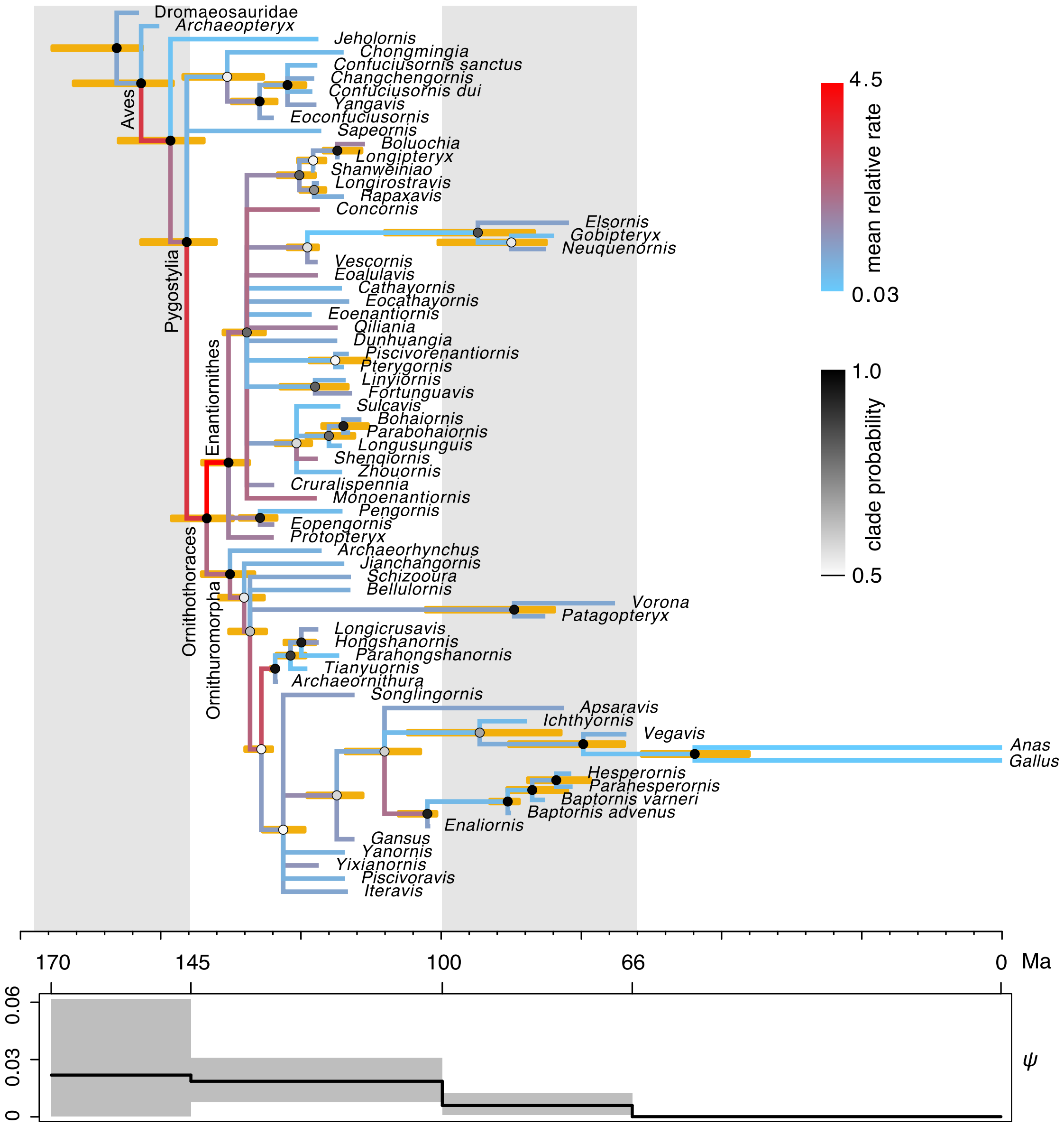

FIGURE 3. Dated phylogeny of Mesozoic birds. The topology shown is the majority-rule consensus tree summarized from the posterior trees. The node ages in the tree are the posterior medians, and the error bars at the nodes denote the 95% highest posterior density (HPD) intervals. The shade of each node circle represents the posterior probability of the corresponding clade. The color of the branch represents the mean relative evolutionary rate at that branch. The fossil sampling rate, ψ, varies along time in a piecewise constant manner, with four intervals divided at 145, 100, and 66 Ma. The solid line is the posterior mean, and the shade denotes the 95% HPD interval.

ILN model (0.98), while the sixth partition, which contains hind-limb characters, has the highest probability for the IGR model (0.83). This illustrates the importance of model averaging, because different character regions have distinct patterns of evolutionary rate variation due to varying natural selection and can thus be better fit by different distributions. The high evolutionary rates during early avian evolution largely correspond to extensive morphological modifications refining flight capability. As discussed in Zhang and Wang (2019; see also Yu et al. 2021), the high rate for axial skeleton reflects extensive morphological

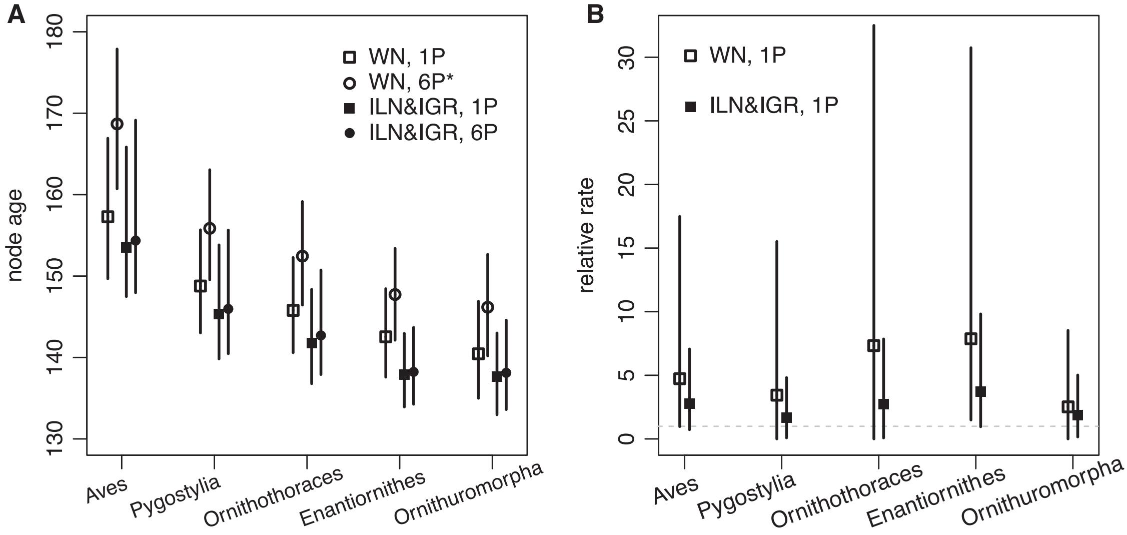

FIGURE 4. A, Five focal node ages (mean and 95% highest posterior density [HPD] interval) of early stem birds estimated under the white noise (WN) model for the unpartitioned data (WN, 1P), WN model for the partitioned data (WN, 6P*; the star denotes disallowing fossil ancestors in the fossilized birth-death [FBD] model), mixed independent lognormal (ILN) and independent gamma rates (IGR) models for the unpartitioned data (ILN&IGR, 1P), and mixed ILN and IGR models for the partitioned data (ILN&IGR, 6P). B, Relative evolutionary rates (mean and 95% HPD interval) at the five focal branches under the WN model for the unpartitioned data (WN, 1P) and mixed ILN and IGR models for the unpartitioned data (ILN&IGR, 1P).

TABLE 2. Posterior probability of the independent lognormal (ILN) clock model and acceptance rate of the reversible-jump Markov chain Monte Carlo (rjMCMC) proposal under each of the four fixed values of *w* and when *w* is auto-tuned.

P_{ILN}	*w*	Acceptance rate
0.57	1.0	0.1%
0.58	2.0	3.1%
0.60	3.0	2.7%
0.60	4.0	0.9%
0.58	Auto	3.0%

changes in the vertebral column, especially the forming of pygostyle, while the high rates for pectoral girdle and sternum correspond to the unique morphological changes related to flight apparatus and the deviation of flight styles transitioning from Pygostylia to Enantiornithes. There appears to be no dominant selective pressure in the skull and pelvis during early avian evolution.

Discussion

In this study, I developed a rjMCMC algorithm to average over the ILN and IGR clock models. The simulation study revealed the ability of the algorithm to infer the true model, while the estimates of divergence times and evolutionary rates in the Bayesian tip dating of Mesozoic birds were both improved by averaging over these two clock models.

The mixed ILN and IGR clocks also improved the MCMC sampling, as the partitioned analysis converged efficiently under the standard FBD model allowing fossil ancestors but failed under the WN model. In the WN model, the rate distribution is also gamma, but the variance of the branch rate is proportional to the branch length. Thus, a change of the time influences the rate and vice versa, but the MCMC proposals are not optimized for these correlated parameters. In the IGR or ILN model, the variance parameter is independent of the branch length, making the MCMC more efficient to update the times and rates. Thus, the overall convergence and mixing are improved.

Nevertheless, there is clearly some room for improvement in the rjMCMC algorithm, as it only achieved acceptance rate of about 3% (Table 2). Take the results of $P_{\mathrm{ILN}} = 0.6$ and

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press

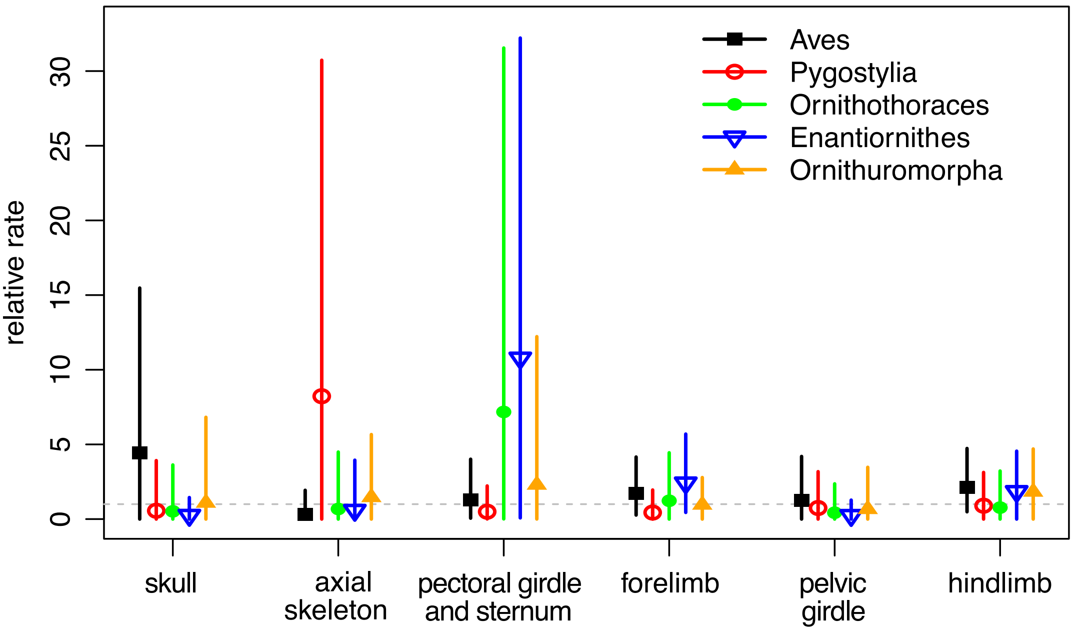

FIGURE 5. Relative evolutionary rates (mean and 95% highest posterior density [HPD] interval) for the six anatomical partitions at the five focal branches averaging over the independent lognormal (ILN) and independent gamma rates (IGR) models (ILN&IGR, 6P).

TABLE 3. Posterior probability of the independent lognormal (ILN) clock model for each of the six partitions and the variance parameters estimated (mean and 95% highest posterior density [HPD] interval).

Partition	P_{ILN}	σ_L	σ_G
Skull	0.69	6.0 (1.5, 11.4)	2.2 (1.0, 3.5)
Axial skeleton	0.51	4.6 (1.4, 8.6)	2.0 (1.0, 3.1)
Pectoral girdle and sternum	0.98	2.2 (0.8, 4.0)	1.0 (0.5, 1.6)
Forelimb	0.54	2.5 (0.9, 4.5)	1.1 (0.6, 1.7)
Pelvic girdle	0.49	3.4 (0.2, 7.5)	1.7 (0.5, 3.0)
Hind limb	0.17	2.7 (1.0, 4.8)	1.2 (0.7, 1.8)

$P_{IGR} = 0.4$, for example. One could imagine that if the chain moves to ILN with certainty and moves to IGR with probability P_{IGR}/P_{ILN}, such a move would be the most efficient with an acceptance rate of $P_{ILN} \times P_{IGR}/P_{ILN} + P_{IGR} \times 1 = 2 \times 0.4 = 0.8$. Note this is a theoretical upper limit, which is hard to reach in practice. Recall that the evolutionary rates are directly mapped between models, so the term that changed during rjMCMC is the joint probability distribution of the rates, which is the multiplication of either lognormal or gamma densities across branches. Apparently, matching these two multidimensional distributions would require a very accurate (and different) value of w in each iteration to avoid the move being rejected too often. The auto-tuning feature developed here has not accomplished this goal, and a smarter one is needed. Li and Drummond (2012). Took another direction by mapping the quantiles of two distributions, such that the probability distribution of the evolutionary rates is unchanged in the cross-model move, but the data likelihoods are different as the evolutionary rates themselves are changed. Because the likelihood would dominate the posterior, this strategy would probably result in a poorer acceptance rate. Also, the likelihood computation is usually much heavier than the prior calculation. Regardless, further studies are necessary to make more thorough evaluations and comparisons. A better algorithm to achieve good acceptance rate and efficiency might lie in between these two strategies, where the likelihood and evolutionary rate distribution need to be co-updated, which is left for further research.

A desirable feature of the rjMCMC algorithm appears to be including an autocorrelated relaxed clock model such as the ALN model. However, it is quite challenging to design efficient moves between the autocorrelated and uncorrelated clock models. The ALN model has rates of r_{i1} and r_{i2} at both ends of branch i, and r_{i2} follows a lognormal distribution with mean r_{i1} and variance $\sigma_A t_i$, where t_i is the time duration of branch i. One could map the branch rate r_i in the ILN model to $r_i' = (r_{i1} + r_{i2})/2$ in the ALN model, but this mapping sometimes reaches negative value of r_{i1} or r_{i2}, resulting in such moves being rejected. For the variance, one could use $\sigma_A = w\sigma_L$, where w can be chosen close to the mean of t_i values. Nevertheless, such an initial attempt tended to get stuck in different clock models. Further efforts are needed to achieve reasonable mixing while giving more consideration to the deviation of the autocorrelated and independent distributions. Alternatively, a mixed relaxed clock model has been introduced to balance the autocorrelated and uncorrelated models (Lartillot et al. 2016), which is a sensible approach in the context of model averaging.

This study draws attention to the importance of comparing alternative relaxed clock models in dating analyses, something frequently overlooked in empirical studies. It is a good practice to test the fit of different clock models, as the pattern of evolutionary rate variation is likely data dependent. Carefully selecting relaxed clock models could also improve parameter estimates and MCMC performance, as shown in this study. Although the rjMCMC algorithm introduced here is applied to the morphological data of Mesozoic birds, it is generally applicable in any node dating or total-evidence dating approaches for which a relaxed clock model is suitable. Thus, it provides a general framework for selecting and averaging relaxed clock models. Future research will have to show how the algorithm performs on molecular sequences or a combination of both morphological and molecular data.

Software Implementation

The rjMCMC algorithm has been implemented in the latest development branch of MrBayes (Ronquist et al. 2012b) available from GitHub (https://github.com/NBISweden/MrBayes) and will be included in the upcoming release version 3.2.8. The relevant commands of using this algorithm are listed here (with comments in square brackets). Complete commands are given in the Supplementary Material.

```
prset clockvarpr = mixed; [specify the mixed ILN and IGR clocks]
prset clockratepr = gamma(2, 100); [for the base evolutionary rate]
prset mixedvarpr = exp(0.5); [for the clock variance]
propset rj_clocks$ratio = 2.0; [initial value of w, or fixed if autotune = no]
mcmcp autotune = yes tunefreq = 5000; [auto-tuning (default) and batch length]
```

[Note:

1. Auto-tune is a global control, i.e., setting it to “no” disables auto-tuning for all proposals.
2. The relaxed clock model indicators (m_RCl) are recorded in .p files with 0 for IGR and 1 for ILN. The ‘sump’ command summarizes these values and prints the posterior probability of ILN (mean of m_RCl).
3. The “clockvarpr = igr” command in v. 3.2.7 and older is replaced by “clockvarpr = wn” in v. 3.2.8 to specify the WN model, while “clockvarpr = igr” and “clockvarpr = iln” in v. 3.2.8 specify the IGR and ILN models.]

Acknowledgments

I am funded by the Hundred Young Talents Program of Chinese Academy of Sciences and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB26000000). I thank Fredrik Ronquist and Nicolas Lartillot for helpful discussions on the rjMCMC algorithm and M. Wang, S. Ho, D. Silvestro, R. Borges, L. H. Liow, and three anonymous reviewers for their useful comments on the article.

Data Availability Statement

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.51c59zw5b.

Literature Cited

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29:2157–2167.

Baele, G., W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution 30:239–243.

dos Reis, M., P. C. J. Donoghue, and Z. Yang. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nature Reviews Genetics 17:71–80.

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

Gavryushkina, A., D. Welch, T. Stadler, and A. J. Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Computational Biology 10:e1003919.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Heath, T. A., and B. R. Moore. 2014. Bayesian inference of species divergence times. Pp. 277–318 *in* M.-H. Chen, L. Kuo, and P. O. Lewis, eds. Bayesian phylogenetics: methods, algorithms, and applications. CRC Press, Boca Raton, FL.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences USA 111:E2957–E2966.

Ho, S. Y. W. 2021. The molecular evolutionary clock, theory and practice. Springer, Cham, Switzerland.

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

Ho, S. Y. W., 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. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773–795.

Kishino, H., and M. Hasegawa. 1990. Converting distance to time: application to human evolution. Methods in Enzymology 183:550–570.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Molecular Biology and Evolution 18:352–361.

Lartillot, N., and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Systematic Biology 55:195–207.

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

Lee, M. S. Y. 2016. Multiple morphological clocks and total-evidence tip-dating in mammals. Biology Letters 12:20160033.

Lee, M. S. Y., A. Cau, D. Naish, and G. J. Dyke. 2014. Morphological clocks in paleontology, and a mid-Cretaceous origin of crown Aves. Systematic Biology 63:442–449.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669–2680.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Li, W., and A. Drummond. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution 29:751–761.

Rannala, B., and Z. Yang. 2007. Inferring speciation times under an episodic molecular clock. Systematic Biology 56:453–466.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology 61:973–999.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Stadler, T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 267:396–404.

Stadler, T. 2011. Simulating trees with a fixed number of extant species. Systematic Biology 60:676–684.

Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Systematic Biology 51:689–702.

Warnock, R., and A. Wright. 2021. Understanding the tripartite approach to Bayesian divergence time estimation. Elements of paleontology. Cambridge University Press, Cambridge.

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. 2014. Molecular evolution: a statistical approach. Oxford University Press, Oxford.

Yang, Z., and C. E. Rodriguez. 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. Proceedings of the National Academy of Sciences USA 110:19307–19312.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Molecular Biology and Evolution 17:1081–1090.

Yu, Y., C. Zhang, and X. Xu. 2021. Deep time diversity and the early radiations of birds. Proceedings of the National Academy of Sciences USA 118:e2019865118.

Zhang, C., and M. Wang. 2019. Bayesian tip dating reveals heterogeneous morphological clocks in Mesozoic birds. Royal Society Open Science 6:182062.

Zhang, C., T. Stadler, S. Klopfstein, T. Heath, and F. Ronquist. 2016. Total-evidence dating under the fossilized birth-death process. Systematic Biology 65:228–249.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 *in* V. Bryson and H. J. Vogel, eds. Evolving genes and proteins. Academic Press, New York.

https://doi.org/10.1017/pab.2021.42 Published online by Cambridge University Press