Bayesian Estimation of Divergence Times From Large Sequence Alignments

Stéphane Guindon

Bayesian Estimation of Divergence Times from Large Sequence Alignments

Stéphane Guindon*[,1,2]

[1]Department of Statistics, University of Auckland, Auckland, New Zealand
[2]Méthodes et Algorithmes pour la Bioinformatique, LIRMM, Centre National de la Recherche Scientifique, Montpellier, France

***Corresponding author:** E-mail: guindon@stat.auckland.ac.nz.

Associate editor: Jeffrey Thorne

Research article

Abstract

Bayesian estimation of divergence times from molecular sequences relies on sophisticated Markov chain Monte Carlo techniques, and Metropolis–Hastings (MH) samplers have been successfully used in that context. This approach involves heavy computational burdens that can hinder the analysis of large phylogenomic data sets. Reliable estimation of divergence times can also be extremely time consuming, if not impossible, for sequence alignments that convey weak or conflicting phylogenetic signals, emphasizing the need for more efficient sampling methods. This article describes a new approach that estimates the posterior density of substitution rates and node times. The prior distribution of rates accounts for their potential autocorrelation along lineages, whereas priors on node ages are modeled with uniform densities. Also, the likelihood function is approximated by a multivariate normal density. The combination of these components leads to convenient mathematical simplifications, allowing the posterior distribution of rates and times to be estimated using a Gibbs sampling algorithm. The analysis of four real-world data sets shows that this sampler outperforms the standard MH approach and demonstrates the suitability of this new method for analyzing large and/or difficult data sets.

Key words: phylogenetics, divergence time estimation, Bayesian estimation, Gibbs sampling, MCMC.

Introduction

Using biological sequences to date past evolutionary events was first proposed more than 40 years ago by the late Allan Wilson (1934–1991). In a pioneering study, Sarich and Wilson (1967) indirectly estimated the number of amino acid differences between homologous proteins in different species and used these measurements to date the divergence between humans and African apes. The conclusion that the earliest protohominids arose 5 Mya—as opposed to the then classical 25 Mya estimate obtained from fossil data—initiated a long series of controversies opposing molecular evolutionists to anthropologists.

A central assumption to Sarich and Wilson's work is that of a "molecular clock." This hypothesis, first proposed by Zuckerkandl and Pauling (1962), states that substitutions accumulate at constant pace over time and throughout lineages. Hence, given a set of homologous sequences and information on the age of the most recent common ancestor for a set of taxa, it is possible to estimate the rate at which the molecular clock ticks. Expected numbers of substitutions can then be translated into standard time units. Unfortunately, ample evidence suggests that the molecular clock is not unique (e.g., Gaut et al. 1992). Indeed, the substitution rates vary drastically across lineages for most sets of taxa and genetic markers. Hence, rather than a single clock, multiple clocks ticking at different rates seem to underlie molecular evolution. The sources of variability of substitution rates over time and lineages are multiple. Variations of the intensity of natural selection in a changing environment, fluctuation of population sizes, and/or generation times are the usual suspects.

Two options are then available to estimate dates of past evolutionary events using biological sequences. The first is to test the hypothesis of a single molecular clock. If this hypothesis cannot be rejected, then divergence times can be estimated using the standard approach mentioned above. Considerable efforts have therefore been devoted to test for the constancy of rates along phylogenetic trees (e.g., Wu and Li 1985; Li and Bousquet 1992; Robinson et al. 1998). The second and more recent option is to explicitly account for the variations of substitution rates along branches of the phylogeny in a statistical framework.

In such context, Sanderson (1997) proposed a nonparametric approach to estimate rates and times concomitantly. This method considers the branch lengths as the data from which rates and times are to be estimated. As there is an infinite number of combinations of rates and times leading to the same branch length estimates, a natural approach is to choose values of node times that minimize the variance of rates along the tree. Indeed, although substitution rates vary over time, it seems reasonable to assume that they are autocorrelated. Sanderson (2002) later reformulated this approach in a maximum likelihood framework, replacing the nonparametric rate smoothing technique by a term in the likelihood function that penalizes large changes of rates across ancestral and descendant lineages. The computer program "r8s" that implements this approach is very popular.

At the same period, Thorne et al. (1998) proposed a Bayesian approach to the same problem. This pioneering work inspired a large number of subsequent studies (e.g.,

Huelsenbeck et al. 2000; Aris-Brosou and Yang 2002; Yang and Rannala 2006; Drummond et al. 2006; Rannala and Yang 2007; Lepage et al. 2007), and the software "multidivtime" developed by J. Thorne is widely used nowadays. The Bayesian approach aims at estimating the joint posterior density of rates and times (plus other parameters considered as nuisance parameters) given an alignment of homologous sequences and, in most cases, the topology of a phylogenetic tree (but see Drummond et al. 2006). Such posterior densities are the product of the likelihood and the prior density of model parameters. Getting accurate time estimates essentially depends on the validity of the model that specifies such prior densities. In practice, the model is divided in two parts. The first describes the evolution of rates along the tree given some specified values for the node times. The second describes the distribution of node times, with no reference to the other model parameters. This latter aspect of the model is the least problematic. If sequences come from population-level data, a coalescent prior (Kingman 1982) seems to be the most reasonable choice. If they come from species-level data, a birth–death prior (see, e.g., Yang and Rannala 2006) can be used. Other priors, such as a Dirichlet distribution (Kishino et al. 2001), have also been proposed.

The part of the model dealing with the evolution of the rate of evolution along the phylogeny given the current node time estimates plays a central role. The first difficulty here is to clearly identify what feature is to be modeled. Some approaches focus on rate trajectories: they model the evolution of the substitution rate across successive time points. The compound Poisson process proposed by Huelsenbeck et al. (2000) is an example of this approach. Other methods focus on the substitution rate averaged over individual edges. The original model of Thorne et al. (1998), later modified by Kishino et al. (2001), belongs to this category. Note that some of the "average" models can be interpreted in terms of "trajectory" ones: Kitazoe et al. (2007) showed that in the "average" model of Kishino et al. (2001), the rate trajectory undergoes Brownian motion. Another important distinction between models concerns the autocorrelation of rates across lineages. Most models explicitly account for autocorrelation of rates. The models described by Drummond et al. (2006) and implemented in the software package "BEAST" (Drummond and Rambaut 2007) are a notable exception. These models are well suited to analyzing very fast-evolving organisms, such as viruses (Drummond et al. 2006), for which there is no strong evidence of rate autocorrelation. In most cases, however, rates of evolution show clear signs of autocorrelation. Indeed, the analysis of three real-world data sets by Lepage et al. (2007) showed that autocorrelation of rates is a strong feature of molecular evolution and generally needs to be accounted for.

A common characteristic of the Bayesian approaches aforementioned is the heavy computational burden involved. There are mainly two reasons to this. The first is the likelihood calculation. Thanks to Felsenstein's pruning algorithm (1981), the number of operations involved in likelihood calculation reduces to a polynomial function of the number of taxa rather than an exponential function. Despite this very significant advance plus other "tricks," such as site pattern aliasing, calculating the likelihood involves a large number of sum–product operations. Thorne et al. (1998) were the first to propose an alternative to this issue. They assumed that, given a fixed tree topology, the likelihood function can be approximated by a multivariate normal density. This approximation relies on a second-order Taylor series approximation of the likelihood function (see Material and Methods). The parameters of the normal density are the expected branch lengths and the corresponding covariance matrix, which are both obtained using standard numerical methods.

The second factor responsible for the heavy computational burden is more practical. Estimating model parameters in a Bayesian framework often relies on sophisticated Markov chain Monte Carlo (MCMC) methods. These techniques are used to build a Markov chain which stationary distribution is the posterior density of interest. Metropolis–Hasting (MH) sampling algorithms (Metropolis et al. 1953) are generally used to construct such Markov chain. This iterative approach consists of proposing a new state of the Markov chain given the current state and accept it with a probability proportional to its posterior density. In practice, the parameters of the MCMC are tuned such that the acceptance-to-rejection ratio is between 0.2 and 0.5 (Gamerman and Lopes 2006), which means that the majority of the calculations are used to evaluate least probable solutions. However, the MH algorithm can be replaced by a Gibbs sampler (Geman S and Geman D 1984) in particular situations. This approach is a special case of the MH algorithm for which new states are proposed with an acceptance-to-rejection ratio equal to 1, which makes this method potentially much faster than the standard MH technique.

The Gibbs algorithm is preferred over MH when full conditional densities of each model parameter are easy to sample from. Such distributions generally do not arise in phylogenetics, mostly because of the complex structure of the whole model (i.e., a tree combined to a stochastic process running along edges of this graph). A notable exception though was recently introduced by Lartillot (2006). Using data augmentation and appropriate priors, Lartillot was able to design a very efficient Gibbs sampling algorithm that clearly outperformed the standard MH approach in estimating phylogenies. The present study is very much inspired by Lartillot's ideas. A multivariate normal approximation to the likelihood is combined here to conjugate priors on rates and times. These three components define a Gibbs sampler that outperforms the traditional MH approach. I first introduce the method and then apply this new sampler to four real-world amino acid and nucleotide data sets. The divergence time estimates obtained with this new approach are virtually identical to those inferred with the MH algorithm, with the estimation process being significantly faster, making this new approach well suited to the analysis of large and difficult data sets.

Material and Methods

Notations are introduced, and the basics of Bayesian estimation of divergence times are described first. The approximation of the likelihood function using a multivariate normal is then justified, and the model of variations of rates along lineages and divergence times is introduced. The combination of this model to the likelihood function, giving rise to a Gibbs sampler, is explained next. The validation method and the data sets analyzed in this study are presented in the remaining sections.

Notations

Let D be the data, that is, an alignment of homologous nucleotide or amino acid sequences. $\Phi = \{\tau, Q, T, R, \theta, \nu\}$ is the phylogenetic model that describes how D arose. τ is the tree topology. In this study, τ is given a priori and considered as fixed. It is a rooted binary tree with $2n - 1$ nodes, n being the number of taxa. Q is the generator of the Markov model of substitution. The same Q matrix is used throughout the tree, which makes the substitution model homogeneous and stationary. Additional constraints on Q make the substitution process reversible (for a review on substitution models, see Bryant et al. 2005). $T = \{T_i\}$ is a vector of node times. By convention, the time is set to zero at the most recent node(s) in the phylogeny and all other node ages have negative values. $R = \{R_i\}$ is a vector of relative substitution rates along edges. R_i is the rate along the branch that has node i at its "distal" extremity, that is, i is the node that is the most distant from the root among the two nodes at the two ends of the edge of interest. The model introduced in this study considers R_i as the relative rate on edge i rather than the relative rate at a specific location in the tree. Relative rates are assumed to be constant along the corresponding edges and changes of rates occur at the nodes only. θ is the absolute substitution rate averaged over the branches of the phylogeny. The length of branch i, denoted as L_i, is therefore equal to $\theta R_i (T_i - T_{\mathrm{anc}(i)})$, where $\mathrm{anc}(i)$ denotes the direct ancestor of node i in the tree, that is, the first node encountered when going from node i to the root of the tree. The last parameter of the phylogenetic model, ν, describes the autocorrelation of substitution rates across adjacent edges.

Bayesian Estimation of Divergence Times and Substitution Rates

Estimating rates and times in a Bayesian framework relies on the joint posterior density of the model parameters:

$$p(T, R, \theta, \nu | D) = \frac{p(D|T, R, \theta, \nu) p(T, R, \theta, \nu)}{\int_T \int_R \int_\theta \int_\nu p(D|T, R, \theta, \nu) p(T, R, \theta, \nu) \mathrm{d}T\ \mathrm{d}R\ \mathrm{d}\theta\ \mathrm{d}\nu}, \quad (1)$$

which displays the two components divergence time estimation relies on (i.e., likelihood and prior on times, rates, and nuisance parameters). Once this density is estimated, it becomes possible to calculate various functions of the posterior densities of node times and substitution rates (e.g., posterior medians or means, credible intervals).

Estimating the joint posterior density defined in equation (1) is a difficult problem. Indeed, evaluating the integral in the denominator of equation (1) appears like a daunting task. MCMC methods provide computationally tractable solutions to overcome this limitation and the Gibbs sampler described in this study is one of them. Another limitation comes from the calculation of the likelihood. The next section describes an approximation that dramatically decreases the computational burden involved here.

Approximation of the Likelihood Function

Thorne et al. (1998) used a multivariate normal density to approximate the likelihood function. This approximation can be justified formally. Let D be the data and X the vector of parameters of interest. $\hat{X}$ denotes the maximum likelihood estimate of X. The second-order Taylor series approximation to the log likelihood around $\hat{X}$ is therefore as follows:

$$\log p(D|X) \simeq \log p(D|\hat{X}) + (X - \hat{X})^T (\partial \log p)_{\hat{X}} + \frac{1}{2}(X - \hat{X})^T (\partial^2 \log p)_{\hat{X}} (X - \hat{X}),$$

where $(\partial \log p)_{\hat{X}}$ is the gradient of $\log p(D|X)$ evaluated at $\hat{X}$ and $(\partial^2 \log p)_{\hat{X}}$ is the Hessian matrix. The gradient being equal to zero when evaluated at $\hat{X}$, we have

$$\log p(D|X) \simeq \log p(D|\hat{X}) + \frac{1}{2}(X - \hat{X})^T (\partial^2 \log p)_{\hat{X}} (X - \hat{X}).$$

The likelihood is therefore expressed as

$$p(D|X) \simeq p(D|\hat{X}) \exp\left(\frac{1}{2}(X - \hat{X})^T (\partial^2 \ln P)_{\hat{X}} (X - \hat{X})\right),$$

which can be rewritten as

$$p(D|X) \simeq p(D|\hat{X}) \exp\left(-\frac{1}{2}(X - \hat{X})^T \hat{\Sigma}^{-1} (X - \hat{X})\right) \propto \exp\left(-\frac{1}{2}(X - \hat{X})^T \hat{\Sigma}^{-1} (X - \hat{X})\right),$$

where $\hat{\Sigma} = -(\partial^2 \ln P)_{\hat{X}}^{-1}$. Hence, a normal density with mean vector $\hat{X}$ and covariance matrix $\hat{\Sigma}$ provides a second-order approximation to the likelihood function.

In the context of phylogenetic inference, for a fixed tree topology, X corresponds to the set of branch lengths and the parameters of the Markov model of character substitutions (e.g., transition/transversion ratio, gamma shape parameter). The maximum likelihood estimates and covariances between them are derived using standard numerical techniques. Once such precalculations are done, evaluating the likelihood comes at a low computational cost as calculating the density of a multivariate normal can be done efficiently.

Priors on Rates and Times

Defining priors on the model parameters is at the heart of all Bayesian analysis. In the context of molecular dating, we need to define priors on rates and times. In order to do so, the Bayesian methods proposed so far model the distribution of the random variables $R\,|\,T$ and T (the variables ν and θ do not show here for the sake of clarity of the argument). Typically, sampling from the posterior density of a given rate R_i relies on the following expression:

$$p(R_i|R_{-i},T,D) \propto p(R,T,D) \quad (2)$$

$$\propto p(D|R,T)p(R|T)p(T), \quad (3)$$

where R_{-i} is the set of all rates except R_i. The very same three components (i.e., $p(D|R,T)$, $p(R|T)$, $p(T)$) are also used to sample from the posterior density of a given node time T_i as $p(T_i|T_{-i},R,D) \propto p(R,T,D)$. Hence, when considering prior densities only, updating rates and times exclusively rely on the conditional density of all rates given every node times $(R\,|\,T)$ and the marginal density of node times (T). However, equation (3) is not the only solution to evaluating the posterior densities of interest. For instance, in order to sample from the posterior density of R_i, one can use the following expression:

$$p(R_i|R_{-i},T,D) \propto p(D|R,T)p(R_i|R_{-i},T). \quad (4)$$

Also, the posterior density of T_i can be expressed as

$$p(T_i|T_{-i},R,D) \propto p(D|R,T)p(T_i|T_{-i},R). \quad (5)$$

The present study precisely relies on equations (4) and (5) to evaluate the posterior densities of rates and times. To sum up, regarding priors on rates and times, instead of modeling the distribution of $R\,|\,T$ and T, we focus on the distributions of the random variables $R_i\,|\,R_{-i},T$, and $T_i\,|\,T_{-i},R$.

Borrowing on the ideas of Thorne et al. (1998), our model assumes that the prior distribution of the relative substitution rate on a given edge is a normal density centered on the relative rate of the ancestral lineage, with variance proportional to the time elapsed along this edge. Because substitution rates cannot be negative quantities, the normal distribution is truncated to nonnegative values. For example, taking the tree in figure 1 as reference, we have

$$R_3|R_1,T_1,T_3,\nu \sim \mathcal{N}^+(R_1,\nu(T_3-T_1)),$$

where ν is the autocorrelation of rates parameter. Its value is estimated from the data. The superscript "+" indicates that the normal distribution is truncated to nonnegative values. More generally, we write

$$R_i|R_{\mathrm{anc}(i)},T_i,T_{\mathrm{anc}(i)},\nu \sim \mathcal{N}^+(R_{\mathrm{anc}(i)},\nu(T_i-T_{\mathrm{anc}(i)})). \quad (6)$$

From this premise, it is possible to derive the distribution of the random variable $X_i \equiv R_i|R_{-i},T,\nu,\theta$. To do so, one notes that the conditional distribution of the rate on an internal branch does only depend on the rates on three edges connected to the branch of interest. For instance, the distribution of X_1 only depends on R_0, R_2, and R_3 (see fig. 1). Indeed, R_1 and R_4 are conditionally independent given

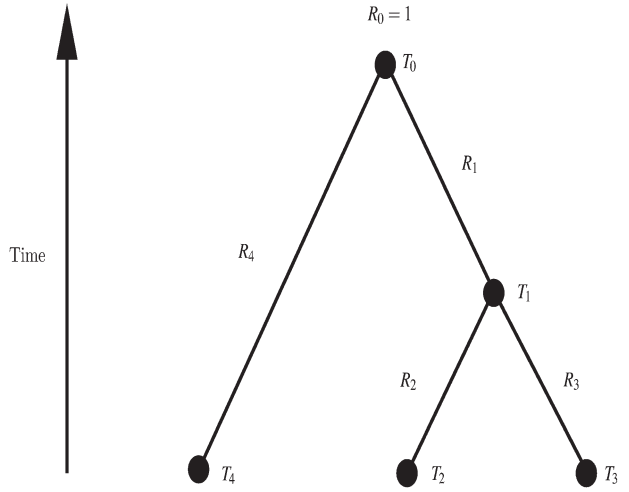

FIG. 1. Three-taxon rooted tree. R values are relative rates. R_0 is the relative rate on the edge above the root node. Its value is set to 1.0. T are time points. For this tree $T_4 = T_2 = T_3 > T_1 > T_0$.

R_0. Also, given R_2 (respectively R_3), knowing the rates on lineages below node 2 (respectively node 3) does not add any information about the distribution of X_1. Hence, we have

$$p(X_1) \propto p(R_1|R_0,T,\nu) \times p(R_2|R_1,T,\nu) \times p(R_3|R_1,T,\nu)$$

$$\propto \frac{\exp\left[-\frac{1}{2}\left(\frac{R_1-R_0}{\sigma_1}\right)^2\right]}{\Pr(R_1>0|R_0,\sigma_1)} \times \frac{\exp\left[-\frac{1}{2}\left(\frac{R_2-R_1}{\sigma_2}\right)^2\right]}{\Pr(R_2>0|R_1,\sigma_2)}$$

$$\times \frac{\exp\left[-\frac{1}{2}\left(\frac{R_3-R_1}{\sigma_3}\right)^2\right]}{\Pr(R_3>0|R_1,\sigma_3)},$$

where $\sigma_1^2 = \nu(T_1-T_0)$, $\sigma_2^2 = \nu(T_2-T_1)$, and $\sigma_3^2 = \nu(T_3-T_1)$. After a little algebra, the previous equation can be factorized as

$$p(X_1) \propto \frac{\exp\left[-\frac{1}{2}\left(\frac{R_1-\mu_1^*}{\sigma_1^*}\right)^2\right]}{\Pr(R_1^+)\Pr(R_2^+)\Pr(R_3^+)},$$

where $\Pr(R_1^+) = \Pr(R_1>0|R_0,\sigma_1)$, $\Pr(R_2^+) = \Pr(R_2>0|R_1,\sigma_2)$, and $\Pr(R_3^+) = \Pr(R_3>0|R_1,\sigma_3)$. The parameters σ_1^* and μ_1^* are defined as follows:

$$\sigma_1^{*2} = \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}+\frac{1}{\sigma_3^2}}$$

and

$$\mu_1^* = \sigma_1^{*2}\left(\frac{R_0}{\sigma_1^2}+\frac{R_2}{\sigma_2^2}+\frac{R_3}{\sigma_3^2}\right).$$

If R_1 is the rate on an external edge, we have

$$p(X_1) \propto \frac{\exp\left[-\frac{1}{2}\left(\frac{R_1-R_0}{\sigma_1}\right)^2\right]}{\Pr(R_1^+)},$$

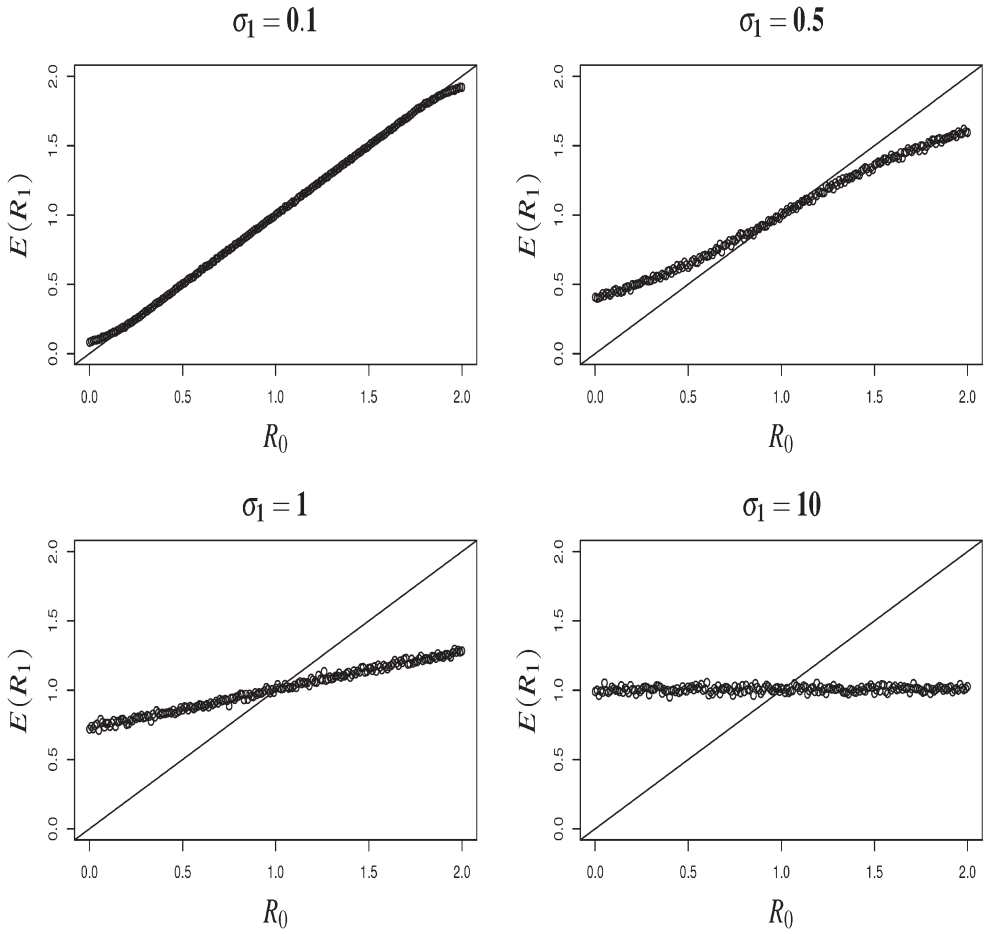

FIG. 2. Prior expectation of the descendant rate R_1 given the ancestral rate R_0 for various values of σ_1. R_1 and σ_1 are the rate and standard deviation of the relative rate on a given edge. R_0 is the ancestral rate. See figure 1 and text for precise definitions.

making the distribution of X_1 a normal density with mean R_0, standard deviation σ_1, and truncated to positive values.

Lepage et al. (2006) provide a relevant list of criteria that any model of evolution of the rate of evolution should satisfy. First, such model should be computational tractable: evaluating priors on rates should involve a "reasonable" number of operations. The model introduced here performs well regarding this criterion as evaluating normal densities is very fast. More importantly, a good model of rate evolution should have a well-defined stationary distribution, that is, when the process is run over a long period of time, rates should converge to a sensible distribution of values. The Ornstein–Uhlenbeck model of Aris-Brosou and Yang (2002) and the Cox Ingersol Ross (CIR) model introduced in phylogenetics by Lepage et al. (2006) both satisfy this constraint. Our model also fares well here, as demonstrated in figure 2. The abscissa gives the value of R_0, whereas the ordinates display the expected value of R_1 for various values of standard deviation of the rate on edge 1 (corresponding to varying values of the product $\nu(T_1 - T_0)$, see fig. 1). For small standard deviation values, $E(R_1) \simeq R_0$ for all values R_0 can take, matching the intuition that substitution rates generally do not undergo drastic changes over short periods of time. For larger standard deviation values, $E(R_1) > R_0$ if $R_0 < 1$ and $E(R_1) < R_0$ if $R_0 > 1$, corresponding to a balancing effect: slow ancestral rates tend to increase in descendant lineages and fast ancestral rates tend to decrease. In the limit, when the standard deviation reaches very large values, the stationary distribution of R_1 is uniform between 0.0 and 2.0. It does no longer depends on the value of R_0. Note that such appropriate behavior comes at a price: the maximum value a relative rate can take cannot be greater than 2.0. Hence, our model is probably not well suited to data sets for which very strong variations of rates across edges are likely. Fortunately, it is possible to detect problematic cases by visually inspecting the posterior distribution of rates and assess whether removing the left truncation would affect these.

Priors on node times are defined in a very straightforward manner. The distribution of the random variable $Y_i \equiv T_i | T_{-i}, R, \nu, \theta$ is chosen as uniform in the $[T_{\text{anc}}(i), \min(T_{\text{left}(i)}, T_{\text{right}(i)})]$ interval, where $\text{left}(i)$ and $\text{right}(i)$ are the first nodes encountered when going from node i toward the tips through the left and right of node i, respectively. For instance, Y_1 is uniformly distributed in the $[T_0, \min(T_2, T_3)]$ interval (see fig. 1). Note that just because the prior on a given node time is not a function of the relative rates does not mean that times and rates are

considered as independent from each other. Indeed, we have

$$p(T_i, T_{-i}|R) = \frac{p(T_i|T_{-i})p(T_{-i}, R)}{p(R)}$$

and $p(T_{-i}, R) = p(R|T_{-i})p(T_{-i}) \neq p(R)p(T_{-i})$, so that $p(T_i, T_{-i}|R) = p(T|R) \neq p(T)$.

Also, note that the prior on node ages does not depend on the absolute substitution rate (parameter θ). This feature of the model is convenient. Indeed, phylogenomic data sets often display data partitions, generally corresponding to distinct genes. Some partitions may, on average, evolve faster than others. A model in which the prior distribution of times is affected by the absolute rates would be difficult to specify. Indeed, whereas the absolute rates may vary across genes, the divergence times are the same across partitions. Hence, the independence of divergence times and absolute substitution rates is a requirement if one wants to analyze multigene data sets.

Full Conditional Distributions

The previous section focused on the distributions of $X_i \equiv R_i|R_{-i}, T, \theta, \nu$ and $Y_i \equiv T_i|T_{-i}, R, \theta, \nu$. It is now possible to define the full conditionals, that is, the distributions of the random variables $X_i' \equiv R_i|R_{-i}, T, \theta, \nu, D$ and $Y_i' \equiv T_i|T_{-i}, R, \theta, \nu, D$.

For our problem, the likelihood is approximated by a multivariate normal density with mean vector $\hat{L}$ and covariance matrix $\hat{\Sigma}$ (see eq. 2). Given a multivariate normal distribution of a random vector L, the conditional densities of $L_i|L_{-i}$ are also normally distributed. The mean $\hat{L}_i$ and variance $\hat{\varsigma}_i^2$ of the random variable $L_i|L_{-i}$ can be easily determined from $\hat{L}$ and $\hat{\Sigma}$ (see Gamerman and Lopes 2006, p. 22, for instance). Therefore, when considering L_i as the only random component of the vector L, the likelihood is

$$p(D|L_i, L_{-i}) \propto \exp\left[-\frac{1}{2}\left(\frac{L_i - \hat{L}_i}{\hat{\varsigma}_i}\right)^2\right],$$

which can also be expressed as a function of rates, times, and autocorrelation parameter:

$$p(D|R_i, R_{-i}, T, \theta, \nu) \propto \exp\left[-\frac{1}{2}\left(\frac{R_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)^2\right], \quad (7)$$

where $\hat{\mu}_i = \frac{\hat{L}_i}{\theta(T_i - T_{\text{anc}(i)})}$ and $\hat{\sigma}_i = \frac{\hat{\varsigma}_i}{\theta(T_i - T_{\text{anc}(i)})}$.

As for the prior, we have

$$p(R_i|R_{-i}, T, \theta, \nu) \propto \frac{\exp\left[-\frac{1}{2}\left(\frac{R_i - \mu_i^*}{\sigma_i^*}\right)^2\right]}{\Pr(R_i^+)\Pr(R_{\text{left}(i)}^+)\Pr(R_{\text{right}(i)}^+)}. \quad (8)$$

The parameters σ_i^* and μ_i^* are defined as follows:

$$\sigma_i^{*2} = \frac{1}{\frac{1}{\sigma_i^2} + \frac{1}{\sigma_{\text{left}(i)}^2} + \frac{1}{\sigma_{\text{right}(i)}^2}},$$

$$\mu_i^* = \sigma_i^{*2}\left(\frac{R_{\text{anc}(i)}}{\sigma_i^2} + \frac{R_{\text{left}(i)}}{\sigma_{\text{left}(i)}^2} + \frac{R_{\text{right}(i)}}{\sigma_{\text{right}(i)}^2}\right).$$

The posterior density of rate i can then be obtained by combining equations (7) and (8). We have

$$p(R_i|R_{-i}, T, \theta, \nu, D) \propto \frac{\exp\left[-\frac{1}{2}\left(\frac{R_i - \mu_i'}{\sigma_i'}\right)^2\right]}{\Pr(R_i^+)\Pr(R_{\text{left}(i)}^+)\Pr(R_{\text{right}(i)}^+)}, \quad (9)$$

where the expressions of the parameters μ_i' and σ_i' are given below as

$$\mu_i' = \frac{1}{\frac{1}{\hat{\sigma}_i^2} + \frac{1}{\sigma_i^{*2}}}\left(\frac{\hat{\mu}_i}{\hat{\sigma}_i^2} + \frac{\mu_i^*}{\sigma_i^{*2}}\right),$$

$$\sigma_i'^2 = \frac{1}{\frac{1}{\hat{\sigma}_i^2} + \frac{1}{\sigma_i^{*2}}}.$$

Ignoring the denominator in equation (9), the full conditional distribution of R_i is a normal density with parameters μ_i' and σ_i'. Sampling from this distribution is straightforward. However, for positive values of R_i, the actual distribution is distinct from a normal when $\Pr(R_{\text{left}(i)}^+)\Pr(R_{\text{right}(i)}^+) < 1$. This product of probabilities is generally very close to 1.0 in practice though. Hence, the full conditional distribution of R_i is indeed well approximated by a normal density. Therefore, sampling from the actual posterior density of R_i can be done efficiently by using a MH step with a proposal density determined by the normal density aforementioned. With the four real-world data sets analyzed in this study (see Results), the acceptance rate of such move is systematically above 97%. Such high acceptance rate indicates that new rate values proposed according to the relevant normal are almost always accepted, making this MH step virtually identical to a Gibbs sampling step.

The full conditional of any given node time, T_i, is proportional to the product of the likelihood when considering T_i as random, whereas T_{-i}, R, θ, and ν are fixed, by the prior density for T_i. The prior distribution of $Y_i \equiv T_i|T_{-i}, R, \theta, \nu$ is uniform (see Section Priors on Rates and Times). Hence, the full conditional is of the form:

$$p(T_i|T_{-i}, R, \theta, \nu, D) \propto p(D|T_i, T_{-i}, R, \theta),$$

that is, a multivariate truncated normal density, which is easy to sample from. The difficulty here lies in the fact that the likelihood is defined as a function of branch lengths, not node times. Or, sampling an internal node age is equivalent to sampling three branch lengths under some constraints. For instance, if the value of T_1 changes in the tree of figure 1, the lengths of the edges 1, 2, and 3 are also modified. Moreover, these three branch length modifications are not applied independently as the three lengths are all functions of T_1. The problem therefore needs to be defined in terms of sampling three edge lengths from truncated normal distributions under specific constraints. We have

$$L_1 = \theta R_1(T_1 - T_0),$$

$$L_2 = \theta R_2(T_2 - T_1),$$

$$L_3 = \theta R_3(T_3 - T_1),$$

and therefore

$$\begin{aligned} T_1 &= T_0 + \frac{L_1}{\theta R_1} \\ &= T_2 - \frac{L_2}{\theta R_2} \\ &= T_3 - \frac{L_3}{\theta R_3}. \end{aligned}$$

We then define two new random variables, Z_2 and Z_3

$$\begin{aligned} Z_2 &= \left(T_0 + \frac{L_1}{\theta R_1}\right) - \left(T_2 - \frac{L_2}{\theta R_2}\right), \\ Z_3 &= \left(T_0 + \frac{L_1}{\theta R_1}\right) - \left(T_3 - \frac{L_3}{\theta R_3}\right). \end{aligned}$$

Hence, jointly sampling L_1, L_2, and L_3, such that $\left(T_0 + \frac{L_1}{\theta R_1}\right) = \left(T_2 - \frac{L_2}{\theta R_2}\right) = \left(T_3 - \frac{L_3}{\theta R_3}\right)$, is equivalent to sampling L_1 given that $Z_2 = 0$ and $Z_3 = 0$. As Z_2 and Z_3 are linear combinations of the normally distributed variables L_1, L_2, and L_3, the joint distribution of L_1, Z_2, and Z_3 is multivariate normal too (with truncation). The conditional mean and variance for L_1, L_2, and L_3, as well as the corresponding covariances, are obtained from $\hat{L}$, and $\hat{\Sigma}$. It is then straightforward to express the mean, variances, and covariances for L_1, Z_2, and Z_3. Randomly sampling from the conditional distribution of $L_1 | Z_2 = 0, Z_3 = 0$, with the additional constraint that $L_1 \geqslant 0$, is done using an inversion method identical to the one used for sampling rates (see previous section).

There is no need to introduce new variables for the special case of the root node. Indeed, changing the time at that node modifies the length of only one branch (the edge on which the root node lies). Hence, no specific constraint applies here. In practice, the distribution of T_0 (see fig. 1) given T_{-0}, R, ν, and θ is chosen as uniform in the $[B, \min(T_1, T_4)]$ range, where B is set by the user. The use of this prior makes the posterior a truncated normal which bounds are functions of B and $\min(T_1, T_4)$. Here again, sampling from this distribution is straightforward.

Bayesian inference also requires samples from the posterior distributions of the nuisance parameters. In the context of this study, the absolute rate of substitution, θ, and the autocorrelation of rates parameter, ν, are considered as nuisance parameters. The full conditional for θ is

$$p(\theta | T, R, \nu, D) \propto p(D | T, R, \theta) p(\theta).$$

The prior distribution of θ is chosen as uniform in the $[1 \times 10^{-8}, 1.0]$ range. The posterior density of θ is therefore a truncated normal. However, here again, a particular constraint needs to be accounted for. Indeed, changing the value of θ modifies the values of every branch length, but the ratio of lengths between pairs of edges remains unchanged. In other words, we want to sample L_i $(i = 0, \ldots, 2n-2)$ under the constraint $L_i / R_i (T_i - T_{\text{anc}(i)}) = \theta$, with R and T being considered as fixed at that stage. We then define a new variable, U_i, such that

$$U_i = \frac{L_0}{R_0 (T_0 - T_{\text{anc}(0)})} - \frac{L_i}{R_i (T_i - T_{\text{anc}(i)})},$$

where branch 0 is any edge distinct from the edge on which lies the root node. For this particular branch, the difference $(T_i - T_{\text{anc}(i)})$ is replaced by the length of the path between the first descendant of the root node on the left-hand side and the first descendant of the root node on the right-hand side, measured in time units. The random variable $L_0 | U_{i=0,\ldots,2n-4} = 0$ then satisfies the constraint on relative edge lengths. As the U_i's are linear combinations of normally distributed variables, it is easy to determine the mean and variance of the variable $L_0 | U_{i=0,\ldots,2n-4} = 0$. A sample of θ from the corresponding full conditional is then given by $u / R_0 (T_0 - T_{\text{anc}(0)})$, where u is a realization of the random variable $L_0 | U_{i=0,\ldots,2n-4} = 0$.

As for the autocorrelation parameter ν, the posterior density is given below:

$$\begin{aligned} p(\nu | T, R, \theta, D) &\propto p(D | T, R, \theta) p(R | T, \nu) p(T) p(\nu) \\ &\propto p(R | T, \nu) p(\nu). \end{aligned}$$

The prior distribution on ν is chosen as a uniform distribution on $[1 \times 10^{-5}, 1.0]$. Such prior is not conjugate to the density of $R | T, \nu$. Hence, the full conditional is no longer a standard distribution that can easily be sampled from.

Gibbs Sampling Scheme

Gibbs sampling is a stochastic simulation approach that aims at estimating the joint density of model parameters from their full conditional densities (see, for instance, Gamerman and Lopes 2006, p. 142). The current study focuses on the joint posterior density of the model parameters, namely $p(T, R, \theta, \nu | D)$, or, using the extended notation, $p(T_0, \ldots, T_{2n-1}, R_0, \ldots, R_{2n-2}, \theta, \nu | D)$. This density is estimated by successive generations from the full conditional distributions. The algorithm can be described in the following way:

1. Initialize the iteration counter $j = 1$ and set the initial values of the parameters, that is, $T^{(0)} = (T_0^{(0)}, \ldots, T_{2n-1}^{(0)})$, $R^{(0)} = (R_0^{(0)}, \ldots, R_{2n-2}^{(0)})$, $\theta^{(0)}$, and $\nu^{(0)}$.
2. Obtain new values $T^{(j)}$, $R^{(j)}$, $\theta^{(j)}$, and $\nu^{(j)}$ from $T^{(j-1)}$, $R^{(j-1)}$, $\theta^{(j)}$, and $\nu^{(j)}$ by sampling successively from the

following full conditional densities:

$$\begin{aligned} R_0^{(j)} &\sim p\left(R_0 | R_{-0}^{(j-1)}, T^{(j-1)}, \theta^{(j-1)}, \nu^{(j-1)}, D\right), \\ &\vdots \\ T_0^{(j)} &\sim p\left(T_0 | R^{(j)}, T_{-0}^{(j-1)}, \theta^{(j-1)}, \nu^{(j-1)}, D\right), \\ &\vdots \\ \theta^{(j)} &\sim p\left(\theta | R^{(j)}, T^{(j)}, \nu^{(j-1)}, D\right), \\ \nu^{(j)} &\sim p\left(\nu | R^{(j)}, T^{(j)}, \theta^{(j)}, D\right). \end{aligned}$$

3. Increment counter from j to $j + 1$ and return to step 2 until convergence is reached.

The four full conditional densities involved in the Gibbs sampler have been defined previously. Updating the values of the relative rates, times, and absolute rates of substitution relies on sampling from the appropriate truncated normal distributions. As the full conditional density for the autocorrelation parameter does not correspond to a standard distribution, we rely on a MH step. Only one step of the MH algorithm is required here to ensure convergence of the MCMC to the stationary distribution (see Gilks et al. 1995, p. 84). Such Metropolis-within-Gibbs approach is a common alternative when the full conditional density of a model parameter does not correspond to a distribution that can be easily sampled from (for a similar example, see Lartillot 2006).

Implementation and Validation

Implementing stochastic techniques such as those described in this study can be tricky. Hence, validation procedures were designed in order to limit errors as much as possible. An MH sampler that estimates the parameters of the very same model as the one described above was therefore implemented in parallel. As the models are identical, both the Gibbs and the MH samplers should asymptotically return the same posterior distributions of parameters.

The MH algorithm relies on proposing new values of the model parameters and comparing the new posterior density to that of the current solution. The functions used to update the parameter values are essentially identical to those described in Thorne et al. (1998). Both the MH and the Gibbs sampler were implemented in the C language as part of the PhyML package (Guindon and Gascuel 2003). The sources are available from http://code.google.com/p/phyml.

Data

Four real-world alignments were considered in this study. The first consists of nucleotide sequences from Caviomorph rodents and Platyrrhine primates previously analyzed by Poux et al. (2006). Sixty-two homologous sequences from three nuclear genes and a total of 3,768 sites are considered here. Nine calibration points, including prior information on the time at the root node, are available. The second data set was recently studied by Rutschmann et al. (2007). It is made of 74 Myrtales sequences, 5,124 nucleotide long, spanning three plastid, and two nuclear genes. Seven calibration nodes, including the root node, are available. These vertices are chosen according to the assignment of fossils to calibration nodes described in Rutschmann et al. (2007). The third data set is the one analyzed by Douzery et al. (2004). It consists of a concatenation of 129 proteins from 36 eukaryotes and a total of 30,399 positions. Information on seven calibration points, including the root node, is available. The fourth data sets was put together and analyzed by Wahlberg (2006). It is made of 59 nucleotide sequences, 2,936 character long, consisting of one mitochondrial gene and two nuclear genes. Four calibration points were available for this data set. These four alignments and the corresponding tree topologies were retrieved from Treebase (Sanderson et al. 1994). The branch lengths of the phylogenies were estimated using PhyML under the HKY model (Hasegawa et al. 1985) for nucleotide sequences and the LG model (Le and Gascuel 2008) for amino acid sequences.

Table 1. Difference of Node Time Estimates Obtained with the Exact and the Multivariate Normal Approximate Likelihood Calculation. Each Posterior Mean Node Time Estimate was First Expressed as a Percentage of the Largest Average Node Time Estimate for Each Data Set. Δ Corresponds to the Differences of Rescaled Node Time Estimates Obtained Using the Gibbs and MH Methods Introduced in This Study. The Table Displays the Quantiles of Δ.

	Quantiles of Δ (%)				
Data sets	0	25	50	75	100
Poux et al. (2006)	−0.63	0.00	0.39	0.65	1.14
Rutschmann et al. (2007)	−4.57	−1.04	−0.17	0.18	1.85
Wahlberg (2006)	−2.24	−1.19	−0.58	0.04	0.43

Results

The validity of the multivariate normal approximation of the likelihood function was assessed first. Three of the four data sets included in this study were analyzed and the posterior distribution of node ages were estimated using two MH samplers. One sampler implements the exact likelihood function, whereas the other implements the approximation. For each analysis, the first 10^6 samples were discarded and the next 5×10^5 samples were considered. Model parameter values were collected every 10^3 samples. Ten repeats of this experiment were run in parallel, each starting from randomly chosen values for the model parameters. Posterior mean of node ages were first rescaled so as to correspond to percentages of the oldest time estimate (i.e., the root age). Such transformation makes the different data sets comparable. Table 1 gives the quantiles of the differences of scaled node time estimates obtained with the two samplers. The data set of Douzery et al. (2004) was not processed here due to heavy computational burdens. The median difference of node age estimates obtained with the exact and the approximate likelihood is between +0.39% and −0.58%. Also, the most important difference between node age estimates obtained with the two approaches is 4.57%. The multivariate normal approximation of the likelihood function is therefore extremely accurate. Moreover, the gain in terms of computing times is dramatic: on average, the approximate approach is between ≃500 and ≃1,500 times faster than the exact one (results not shown).

Table 2. Difference of Node Time Estimates Obtained with the MH and the Gibbs Samplers. 10^8 Samples were Collected in Total. Model Parameter Values were Recorded Every 10^4 Sample. See Caption of Table 1 for Details about the Calculation of Δ values.

	Quantiles of Δ (%)				
Data sets	0	25	50	75	100
Poux et al. (2006)	−0.07	0.01	0.04	0.08	0.15
Rutschmann et al. (2007)	−0.56	−0.03	0.31	0.95	1.79
Douzery et al. (2004)	−0.24	−0.07	0.02	0.05	0.10
Wahlberg (2006)	−0.15	0.06	0.10	0.19	0.31

Each data set was then analyzed using the Gibbs sampler introduced in this study and the equivalent MH sampler. For each analysis, the first 10^6 samples were discarded and the next 10^8 samples were considered. Model parameter values were collected every 10^4 samples. Thirty repeats of this experiment were run in parallel, each starting from randomly chosen values of the model parameters. The validity of the implementation of both the Gibbs and the MH samplers was first assessed by comparing the node time estimates returned by these two approaches. Table 2 gives the quantiles of the differences of scaled node time estimates obtained with the two samplers. For the first data set, the median difference is 0.04%. Roughly similar figures are obtained for the other three data sets and the most important difference between node age estimates obtained with the two approaches is only 1.79%. We also checked that both samplers converged to nearly identical average likelihoods (i.e., $p(D|T,R,\theta)$) and prior densities of relative rates (i.e., $p(R|T,\theta,\nu)$) (results not shown). Altogether, these results indicate that both approaches returned virtually identical model parameter estimates.

Specific node times estimates were then compared with those reported in other studies so as to further check the validity of our results. These studies used the computer program "multidivtime", which implements Thorne et al. (1998) and Kishino et al. (2001) Bayesian approaches to estimate divergence times. For the data of Poux et al. (2006) (fig. 3), our 95% credibility interval for the age of the Catarrhini/Platyrrhini divergence (node 1) is 40±6 Mya, which includes the estimate proposed by the authors (37 Mya). Our estimate of the earliest diversification of the extant platyrrhines (node 2) is 19±5 Mya, whereas the estimate of Poux et al. for the age of this node is 16.8 Mya. The earliest radiation of extant caviomorphs (node 3) is estimated to be within a 37±5 Mya interval according to us, which is very similar to the 36.7 Mya estimate proposed by Poux et al. As for the data set examined by Rutschmann et al. (2007), our estimate for the split of the Southeast Asian Crypteroniaceae from their West Gondwanan sister clade (node 1 in fig. 4) is 68±8 Mya, only slightly above the ∼79 Mya advanced by the authors. Turning to the data set of Douzery et al. (2004), our estimate of the diversification of eukaryotic kingdoms (node 1 in fig. 5) is 1,125±75 Mya, as opposed to 1,104 Mya. According to us, the divergence between animals and choanoflagellates (node 2) is 1,050±100 Mya, whereas the estimate of Douzery is 984 Mya. The most recent common ancestor of the subfamily Nymphalinae (node 1 in fig. 6) is estimated to have existed between 105 and 60 Mya according to Wahlberg (2006). Our estimate is 120±25 Mya, that is, slightly older than Wahlberg's proposed age even though the two proposed time intervals overlap. Wahlberg's estimated age of the Melitaeini clade (node 2) is in the range of 24–71 Mya, whereas our estimate is 70±15 Mya. Our analysis relied on a prior node age for the root of the tree that corresponds to the oldest prior estimates used by Wahlberg (2006), which could explain the slight differences between the two sets of estimates.

Table 3. Average Computing Times. Each Data Set was Analyzed 30 Times with Different Starting Points. The Table Displays the Average Time Duration Needed to Collect 10^8 Samples.

	Average time (s)	
Data sets	Gibbs	MH
Poux et al. (2006)	1,118	6,779
Rutschmann et al. (2007)	1,282	8,419
Douzery et al. (2004)	837	3,566
Wahlberg (2006)	1,078	6,430

The durations required to process each data set with both the Gibbs and the MH samplers were then compared. Table 3 lists the computing times averaged over the 30 repeats of the same experiment for each data set. The average times needed to collect the 10^8 samples using the Gibbs sampler is about four to six times less than the time required by the MH sampler. This difference of computing times is largely explained by the fact that the Gibbs sampler does not require updating the likelihood function when changing the values of the model parameters, as opposed to the MH algorithm. Note, however, that these results rely on a rather crude implementation of the MH sampler and reduced computing times for this method could probably be obtained.

Both the Gibbs and the MH techniques draw nonindependent samples from a target distribution (the joint posterior density of the model parameters). The effective sample size is the size of the sample that would be obtained if the draws were completely independent. It generally amounts to a fraction of the actual sample size (10^8 here). Effective sample sizes can be measured for every parameter of the model. Large effective sample size indicate good "mixing" properties. The effective sample sizes were calculated as described by Gelman in Gilks et al. (1995, p. 137). Table 4 displays the distribution of the ratio of effective sample sizes

Table 4. Quantiles of the Ratios of Effective Population Sizes for Node Time Estimates. Effective Population Sizes Were Calculated for Each Internal Node and Each Method (Gibbs and MH). The Ratio K of the Effective Population Sizes Obtained with the Gibbs Sampler Divided by the Size Obtained with the MH Sampler Were Then Calculated for Each Node. This Table Displays the Quantiles of K.

	Quantiles of K (%)				
Data sets	0	25	50	75	100
Poux et al. (2006)	0.66	1.39	2.13	2.90	5.18
Rutschmann et al. (2007)	1.13	2.76	4.28	8.57	23.91
Douzery et al. (2004)	0.79	2.17	4.69	9.10	20.10
Wahlberg (2006)	0.71	2.75	3.73	5.58	12.20

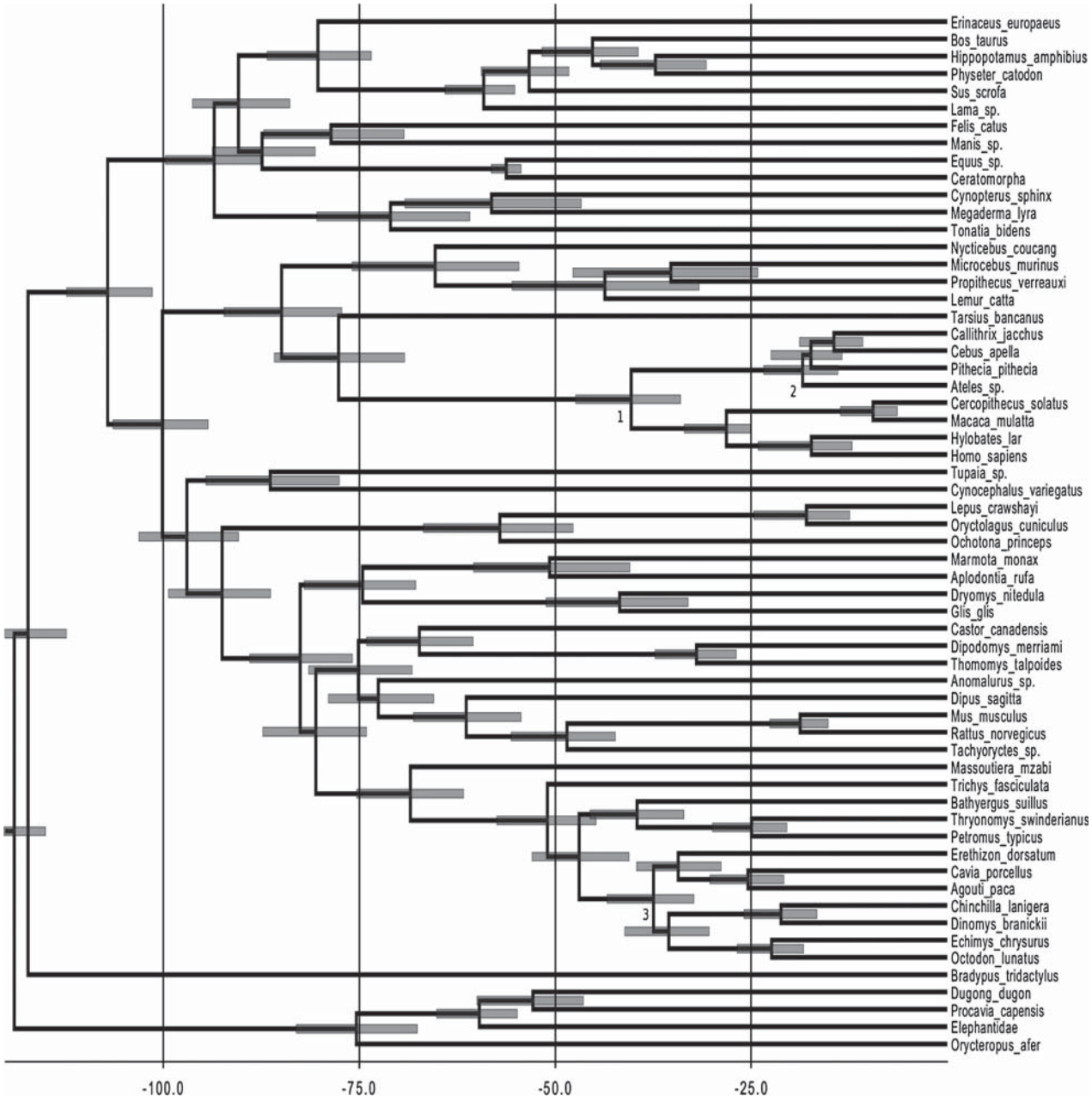

FIG. 3. Poux et al. tree. The node bars give the 95% credible intervals obtained with the method introduced in this study. Numbers below internal nodes denotes clades of particular interest.

measured for node time estimates obtained with the two approaches. The median effective sample size obtained with the Gibbs sampler is between 2.1 and 4.7 times higher with the Gibbs sampler compared with the MH one. The ratio attains 23.9 for one specific node and data set while it is seldom smaller than 1, confirming the clear advantage of the Gibbs sampler over MH.

Discussion

The present study describes a new approach to estimate divergence times from molecular data. This method relies on the approximation of the likelihood function by a multivariate normal density and conjugate priors on substitution rates. The combination of these two features leads to a Gibbs sampling algorithm that makes the estimation of the posterior density of rates and times faster than with the classical MH approach. This method is therefore well suited to analyze large data sets, which are now commonplace in phylogenomics.

An unusual feature of the approach presented here lies in the specification of priors on divergence times and rates. The estimation of model parameters relies exclusively on conditional densities of rates given times and times given rates. This contrasts with the standard approach that models the conditional distribution of rates given times and the

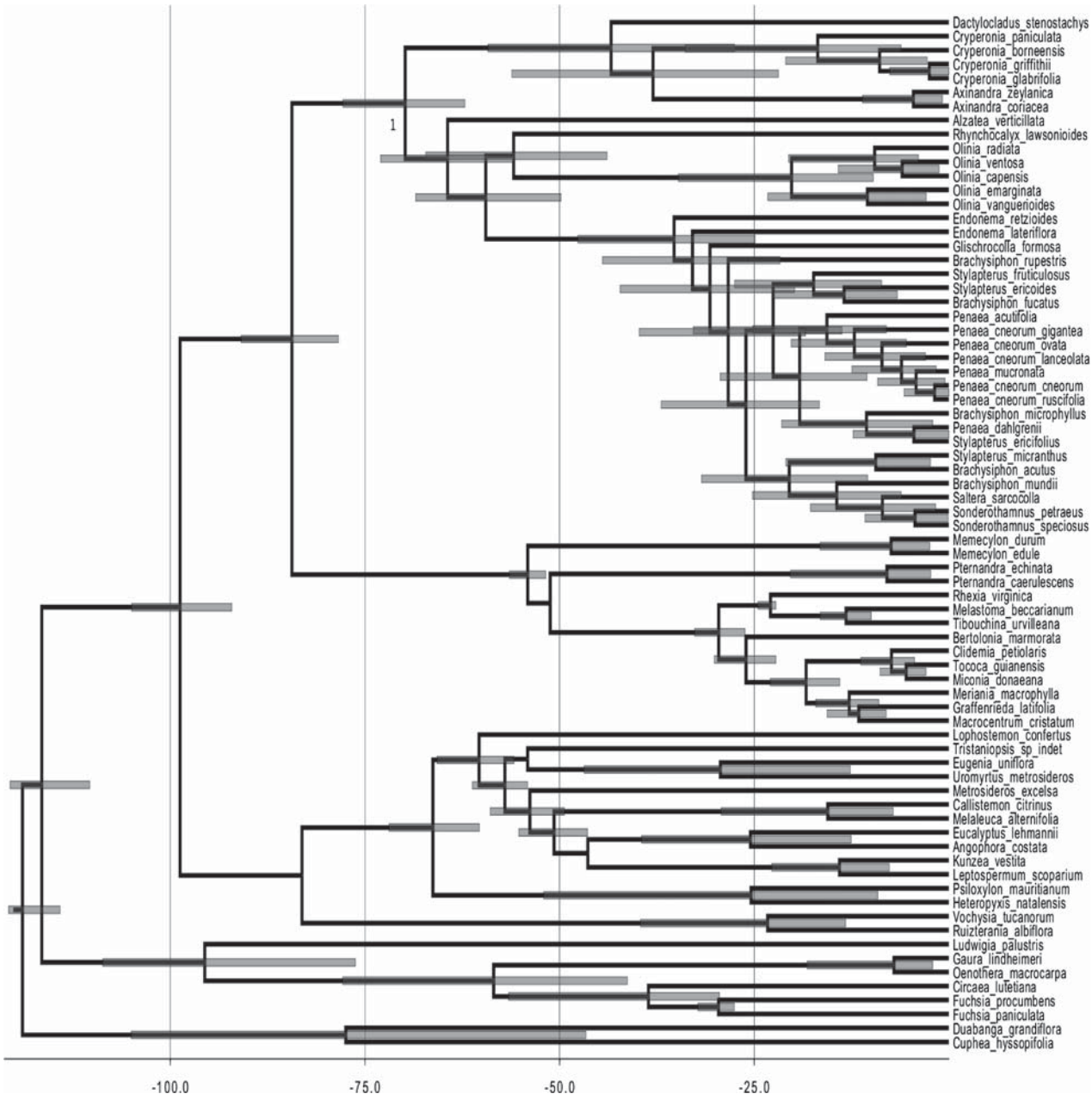

FIG. 4. Rutschmann et al. tree. See caption of figure 3.

marginal distribution of node times. To illustrate the potential benefits of this approach, I borrow an example given by Arnold et al. (2001). If one wishes to model the joint density of body weight and height, then modeling the distribution of weights given heights and the distribution of heights given weights is certainly easier than specifying a model that involves marginal distributions of weights or heights. Thus, in certain circumstances, models based on conditional distributions of parameters are relevant. However, no matter which variable is to be modeled, one must make sure that the joint posterior density of all the variables in the model actually exists. In the model presented here, the joint posterior density of the model parameters can be expressed as a product of the conditional densities of individual relative rates. These conditional densities being truncated normals, the joint posterior density of all model parameters can be evaluated for any value of the model parameters.

From a biological and statistical modeling perspective, it is important to make sure that rates converge to a sensible stationary distribution when the evolution process is run over a long period of time. The Ornstein–Uhlenbeck and CIR models introduced in phylogenetics by Aris-Brosou and Yang (2002) and Lepage et al. (2006), respectively, both have a normal density as stationary distribution. The stationary distribution for the model introduced in this study is uniform. When the product of the time interval on a given branch and the autocorrelation parameter tends to infinity, the distribution of the relative rate is indeed uniform between 0.0 and 2.0. Although a sensible stationary distribution is a boon to the proposed model, limiting the

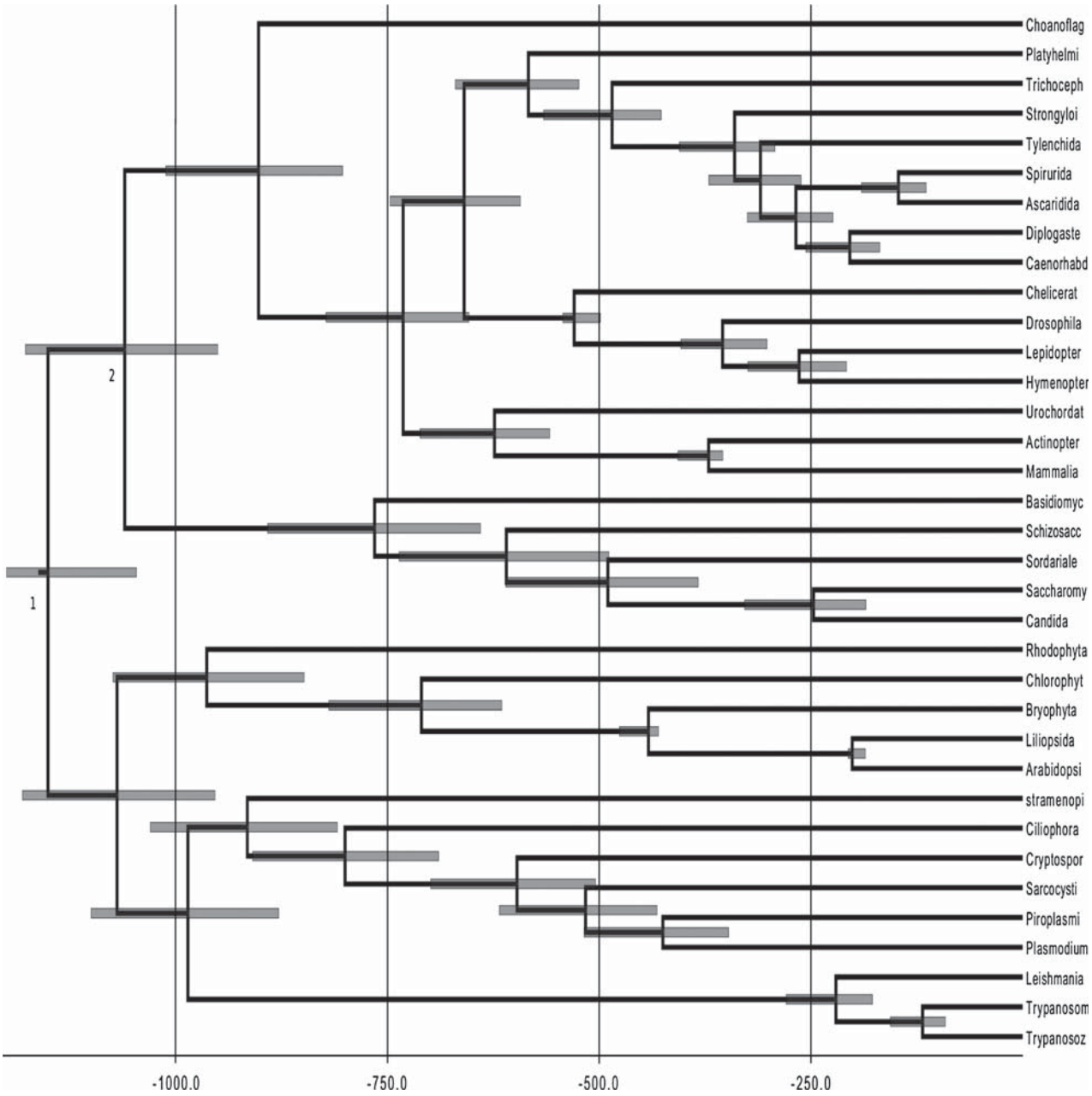

FIG. 5. Douzery et al. tree. See caption of figure 3.

maximum value a relative rate can take to 2.0 can be perceived as a drawback. Hence, our approach is not well suited to cases where sudden and large increases of substitution rates are suspected. Note that it is relatively straightforward to detect such situations by visually inspecting the posterior distribution of rates after fitting our model to the data. It then becomes possible to remove the very fast evolving taxa or to increase the taxon sampling intensity such that very fast lineages are broken into shorter time intervals showing moderate and gradual increases of the substitution rate.

From a biological perspective, other models of rate evolution are more realistic than the one introduced here. For instance, the compound Poisson process of Huelsenbeck et al. (2000) is likely to provide a more relevant description of the correlation of rates across lineages. Despite this, node age estimates obtained with the two methods will probably be fairly similar (see Lepage et al. 2007 for a comparison of models with respect to node age estimates). Therefore, the Gibbs sampling algorithm presented in this article could be used to "guide" the estimation process under a variety of more sophisticated models. More precisely, our algorithm could define an importance sampling function and save considerable amounts of time to the estimation under sophisticated models. The same argument is valid for priors on node ages. In the approach presented here, these priors are modeled using uniform distributions. The coalescent or a birth–death process could also be used in an importance sampling framework where the uniform densities would guide the estimation process under the more realistic priors. Flexible

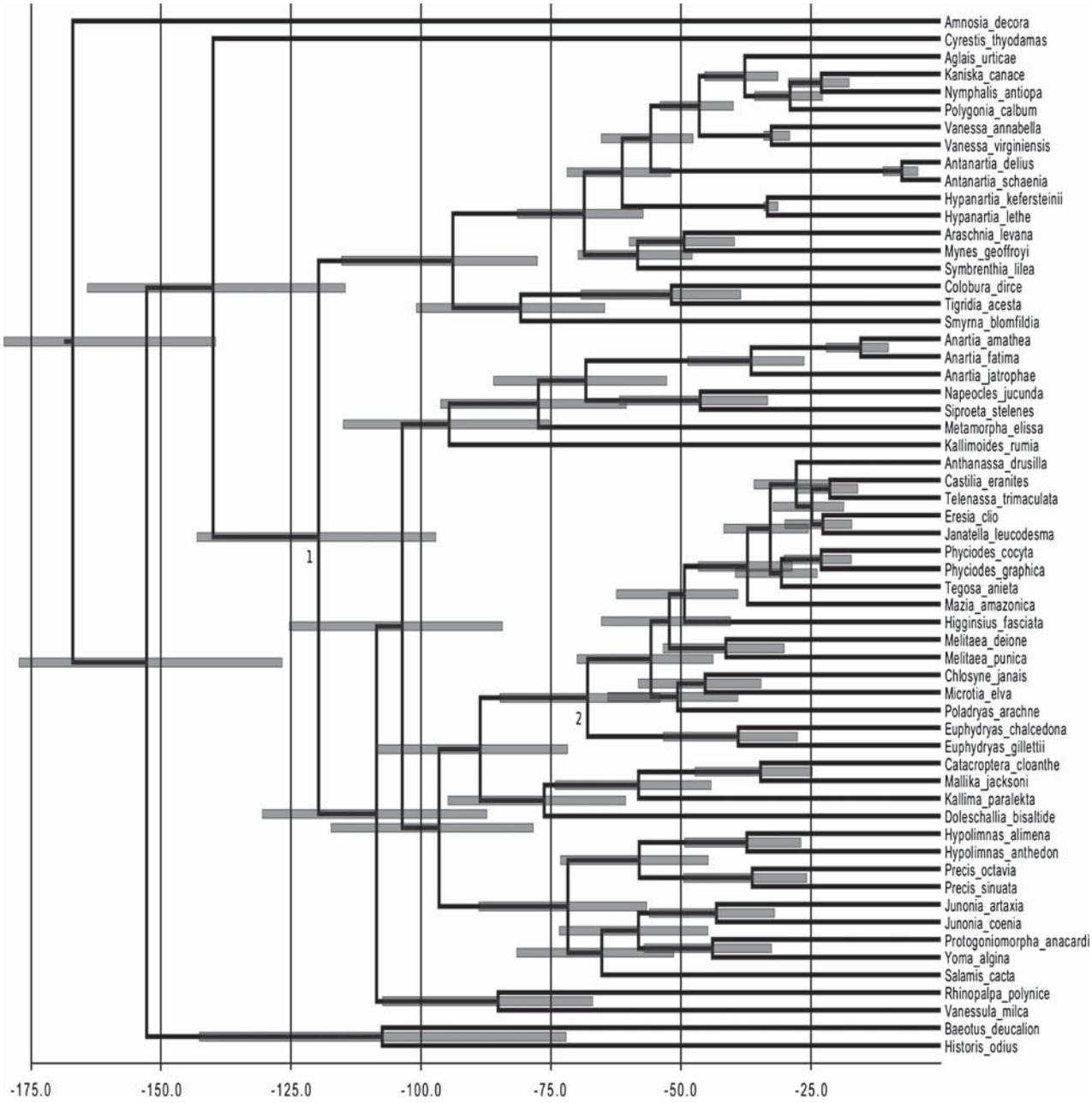

Fig. 6. Wahlberg tree. See caption of figure 3.

priors on calibration times, including soft bounds as proposed by Yang and Rannala (2006), could also be handled using the same importance sampling technique.

Another further development to the present study is inspired by the shape of the marginal posterior distributions of rates and times. Such distributions are generally roughly symmetrical and bell shaped. Hence, instead of approximating the likelihood function with a multivariate normal, it seems relevant to directly approximate the joint posterior density of rates and times using the same family of distributions. It is possible to apply numerical methods to estimate the first two moments of such distribution, just like it is done here with the likelihood function. This approach would be extremely fast as it would not involve any MCMC steps. Note that such method is actually very similar to maximizing a penalized likelihood function (Sanderson 2002), with the penalty term corresponding to the prior on rates and times. The proposed approach would therefore be valuable in providing confidence intervals on node time estimates for maximum likelihood–based methods.

Acknowledgments

This work is funded by the Royal Society of New Zealand through a Marsden grant to the author. Thanks to David Bryant, Mark Holmes, Ivan Kojadinovic, Renate Meyer, Russel Millar, Jeff Thorne, Peter Waddell, and an anonymous reviewer for their help with this study.

References

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

Arnold BC, Castillo E, Sarabia JM. 2001. Conditionally specified distributions: an introduction. *Stat Sci.* 16:249, 265.

Bryant D, Galtier N, Poursat M-A. 2005. Likelihood calculations in phylogenetics. In: Gascuel O, editor. Mathematics of evolution & phylogenetics. Oxford: Oxford University Press. p. 33–62.

Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A.* 101: 15386–15391.

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

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

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

Gamerman D, Lopes HF. 2006. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Boca Raton (FL): Chapman and Hall/CRC.

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

Geman S, Geman D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell.* 6:721–741.

Gilks WR, Richardson S, Spiegelhalter D. 1995. Markov chain Monte Carlo in practice. Boca Raton (FL): Chapman and Hall/CRC.

Guindon S, Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. *J Mol Evol.* 22:160–174.

Huelsenbeck J, Larget B, Swofford D. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879–1892.

Kingman JFC. 1982. The coalescent. Stochastic processes and their applications. Stochastic Processes and their Applications. 13: 235–248.

Kishino H, Thorne J, Bruno W. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol.* 18:352–361.

Kitazoe Y, Kishino H, Waddell PJ, Nakajima N, Okabayashi T, Watabe T, Okuhara Y. 2007. Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS ONE.* 2:e384.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J Comput Biol.* 13:1701–1722.

Le S, Gascuel O. 2008. An improved general amino-acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol.* 24: 2669–2680.

Lepage T, Lawi S, Tupper P, Bryant D. 2006. Continuous and tractable models for the variation of evolutionary rates. *Math Biosci.* 199:216–233.

Li W-H, Bousquet J. 1992. Relative-rate test for nucleotide substitutions between two lineages. *Mol Biol Evol.* 9:1185–1189.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys.* 21:1087–1092.

Poux C, Chevret P, Huchon D, de Jong WW, Douzery EJ. 2006. Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. *Syst Biol.* 55:228–244.

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol.* 56:453–466.

Robinson M, Gouy M, Gautier C, Mouchiroud D. 1998. Sensitivity of the relative-rate test to taxonomic sampling. *Mol Biol Evol.* 15:1091–1098.

Rutschmann F, Eriksson T, Salim KA, Conti E. 2007. Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points. *Syst Biol.* 56:591–608.

Sanderson M. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol.* 14:1218–1231.

Sanderson M. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol.* 19:101–109.

Sanderson M, Donoghue M, Piel W, Eriksson T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Bot.* 81:183.

Sarich V, Wilson A. 1967. Immunological time scale for hominid evolution. *Science* 158:1200–1203.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol.* 12:823–833.

Thorne J, Kishino H, Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15:1647–1657.

Wahlberg N. 2006. That awkward age for butterflies: insights from the age of the butterfly subfamily Nymphalinae (Lepidoptera: Nymphalidae). *Syst Biol.* 55:703–714.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A.* 82: 1741–1745.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in biochemistry. Amsterdam (The Netherlands): Elsevier. p. 189–225.