From Trajectories to Averages: An Improved Description of the Heterogeneity of Substitution Rates Along Lineages

Stéphane Guindon

Syst. Biol. 62(1):22–34, 2013

DOI:10.1093/sysbio/sys063
Advance Access publication on July 12, 2012

From Trajectories to Averages: An Improved Description of the Heterogeneity of Substitution Rates Along Lineages

Stéphane Guindon

Department of Statistics, University of Auckland, Auckland, 1010, New Zealand; and LIRMM, CNRS UMR 5506, Montpellier 34095, France;
**Correspondence to be sent to: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand;*
E-mail: s.guindon@auckland.ac.nz.

Received 10 July 2011; reviews returned 6 October 2011; accepted 5 July 2012
Associate Editor: Peter Foster

Abstract.—The accuracy and precision of species divergence date estimation from molecular data strongly depend on the models describing the variation of substitution rates along a phylogeny. These models generally assume that rates randomly fluctuate along branches from one node to the next. However, for mathematical convenience, the stochasticity of such a process is ignored when translating these rate trajectories into branch lengths. This study addresses this shortcoming. A new approach is described that explicitly considers the average substitution rates along branches as random quantities, resulting in a more realistic description of the variations of evolutionary rates along lineages. The proposed method provides more precise estimates of the rate autocorrelation parameter as well as divergence times. Also, simulation results indicate that ignoring the stochastic variation of rates along edges can lead to significant overestimation of specific node ages. Altogether, the new approach introduced in this study is a step forward to designing biologically relevant models of rate evolution that are well suited to data sets with dense taxon sampling which are likely to present rate autocorrelation. The computer programme PhyTime, part of the PhyML package and implementing the new approach, is available from http://code.google.com/p/phyml (last accessed 1 August 2012) . [Divergence date estimation; geometric Brownian process; MCMC; Bayesian estimation.]

Modelling the variations of nucleotide or amino acid substitution rates along a phylogeny or a genealogy is essential to deciphering the processes of molecular evolution. For instance, an accurate description of how substitution rates vary throughout a phylogeny is central to characterize the correlation between rates of evolution and life-history traits such as body mass index or longevity (Harvey and Pagel 1991; Lartillot and Poujol 2011). These models are also important to accurately estimate dates of divergence from molecular data. In this context, Zuckerkandl and Pauling (1962) first suggested that substitutions accumulate at constant pace over time and throughout lineages. The subsequent estimation of the date of divergence between humans and apes (Sarich and Wilson 1967) relied heavily of the so-called molecular clock hypothesis. However, with the accumulation of molecular data, it became clear that the molecular clock constraint did not always hold. It was then required to design more sophisticated models that allow substitution rates to vary during the course of evolution.

Sanderson first proposed a rate smoothing approach (Sanderson 1997, 2002) in which the rate variation along branches of the phylogeny are governed by a penalty term in the likelihood calculation. This approach was further extended by Britton et al. (2007) to handle large-phylogenomic data sets. However, Thorne and colleagues (Thorne et al., 1998; Kishino et al., 2001) were the first to propose a method of node age estimation that relied on an explicit model of rate evolution. The last decade has seen a number of studies that essentially rely on the framework set out by these authors (e.g., see Aris-Brosou and Yang, 2002; Drummond et al., 2006; Drummond and Suchard, 2010; Guindon, 2010; Huelsenbeck et al., 2000; Rannala and Yang, 2007; Yang and Rannala, 2006). Some of the models of rate evolution assume *a priori* that rates along branches are sampled independently from the same probabilistic distribution (e.g., Drummond et al., 2006). The variance of this distribution quantifies the deviation from the molecular clock. Although this type of model does not assume rates to be autocorrelated *a priori*, post-analysis processing of estimated rates can potentially reveal evidence of autocorrelation. Other models explicitly consider rates to be autocorrelated *a priori*. According to these models, fast evolving ancestors are more likely to give rise to fast evolving descendants than slow evolving ones. This hypothesis might not be relevant when taxon sampling is sparse because the time elapsed along branches is likely to be too great to generate autocorrelation across lineages. However, with dense sampling and therefore shorter time intervals between nodes in a phylogeny, accounting for autocorrelation becomes more important. For this reason, models with rate autocorrelation are well suited to analyse large phylogenomic data sets. This study focuses on this class of models.

The standard approach to model rates of evolution relies on two steps. In the first step, the rate at the end of a given branch (or a function of it such as the logarithm) is considered as a random variable, the distribution of which is a function of the rate (or the logarithm of it) at the beginning of the same branch. A common choice for this distribution is the normal density centred on the logarithm of the rate at the beginning of the branch. The logarithm of the rate trajectory then follows a Brownian process (Thorne et al., 1998; Kishino et al., 2001). Other stochastic processes have been proposed afterwards. Unlike the Brownian process, the Ornstein–Uhlenbeck

(OU) process has a stationary distribution, meaning that large rates of evolution are more likely to decrease rather than increase and *vice versa*. The OU model was introduced in phylogenetics by Aris-Brosou and Yang (2003). Lepage et al. (2006) used a Cox–Ingersoll–Ross (CIR) process instead. The CIR process is a generalization of the squared OU. As opposed to the OU process, the CIR process cannot take on negative values, which is of course relevant when modelling rates of evolution.

The second step consists in deriving average substitution rates along branches from the rates of evolution derived using one of the models mentioned previously. Combined with node ages, average rates are transformed into branch lengths which are then plugged in the calculation of the probability of observing the sequence alignment given the phylogenetic model (i.e., Felsenstein's likelihood). Average rates are generally derived using a deterministic function of the rates at both extremities of the corresponding branches. The most commonly used function here is the arithmetic average: the mean substitution rate along a branch is the arithmetic average of the rates observed at both its extremities. Therefore, although the rate trajectories are modelled using a stochastic process, the average rate along a branch is obtained through a deterministic calculation. This simplification amounts to ignoring the fact that the average rate along a branch given the rates at its extremities is itself a random variable.

Mathematical convenience is the main motivation behind such simplification. Indeed, the calculation of the probability of transition from one nucleotide or amino acid state to another along a given branch of specified length is relatively straightforward when this length is constant. If that length is a random variable, deriving these probabilities is more problematic. Lepage et al. (2006) explicitly tackled this issue. The authors of this study were able to derive an analytical expression for the probability of transition along a branch of a specified length given in calendar units, effectively integrating over all possible rate trajectories described by a CIR process. They also give the formula for calculating the likelihood of a 3-taxa unrooted tree. Because the number of terms involved in this calculation increases exponentially with the number of taxa, Monte Carlo techniques have to be used to evaluate likelihoods on larger phylogenies.

Another attempt to deal with the same issue was put forward by Huelsenbeck et al. (2000). According to the model proposed by these authors, the substitution rate along a branch (the rate trajectory) is constant until a "rate change event" occurs. The number of change events in a given time period is Poisson distributed. After a change event, the new rate is the rate before the event multiplied by a gamma-distributed scalar. Fitting this compound Poisson process to the data requires integrating over all the rate trajectories which, in principle, can be done using sophisticated Markov Chain Monte Carlo (MCMC) techniques. However, processing data sets of standard sizes under this model seems difficult. To the best of our knowledge, the software TreeTime (Himmelmann and Metzler, 2009) is the only one providing an implementation of this approach. This tool being relatively recent, the efficacy of its implementation has not been assessed yet.

Rannala and Yang (2007) have described another approach to the same problem. Assuming a standard Brownian process for the logarithm of rate trajectories, they were able to derive the joint prior distribution of the substitution rates at the midpoints of two sister branches (i.e., the two branches that are connected by the same ancestral node) given the rate at the midpoint of the ancestral branch (i.e., the branch directly "above" the two sister branches). The midpoint rate is then used here as a proxy to the average rate. The two authors rightly note that "[...] calculation of the average rate for each branch is approximate [in this algorithm]; ideally, the length of a branch should be calculated as an integral over the sample path of the geometric Brownian motion process, or otherwise calculation of the transition probability from one nucleotide to another along the branch has to take explicit account of the fluctuating rate."

This article essentially tackles the issue pointed out by Rannala and Yang in this last statement. Using a geometric Brownian process of rate variation across lineages, we propose a semianalytical solution to the integral of the transition probabilities over the average rate along a branch. This technique is very much inspired by Lepage et al. (2006). However, rather than presenting a fully analytical solution, we introduce an approximation and demonstrate its accuracy using simulations. Using parametric bootstrap, we show that rate autocorrelation and node age estimates inferred using the new approach are more precise than that estimated using the conventional model. We also show that ignoring the stochasticity of the average rates potentially leads to significant overestimation of internal node ages.

Materials and Methods

Preliminary

Before giving the details of our new approach, we first define the average rate along an edge mathematically. We then briefly describe how it is possible to integrate over these average rates when calculating probabilities of change from one nucleotide, amino acid, or codon state to another along a branch.

Let Z_{t,r_0,r_t} be the average rate along a branch of length t, in calendar units, with rate r_0 at one of its extremity and r_t at the other extremity. We have

$$Z_{t,r_0,r_t} = \frac{1}{t} \int_{s=0}^{s=t} R_{s|t,r_0,r_t} \mathrm{d}s, \tag{1}$$

where $R_{s|t,r_0,r_t}$ is the random variable that describes the rate trajectory. It corresponds to the rate at time s $(0 \le s \le t)$ along a branch of length t given that the rate at $s=0$ is r_0 and the rate at $s=t$ is r_t. For the sake of simplicity of notation, the random variables $R_{s|t,r_0,r_t}$ and Z_{t,r_0,r_t} will be denoted as R_s and Z, respectively, in what follows. Now, let X_s denote the character state (i.e., the nucleotide, amino acid, or codon) observed at time s along the same edge. The conditional probability of observing state x at time t given that y is observed at time 0 is given below:

$$\begin{aligned} &P(X_t=x|X_0=y,r_0,r_t,t) \\ &\quad = \int_{z=0}^{z=\infty} P(X_t=x|X_0=y,Z=z)P(Z=z|r_0,r_t,t)\mathrm{d}z \\ &\quad = \int_{z=0}^{z=\infty} [\exp(Qz)]_{x,y}P(Z=z|r_0,r_t,t)\mathrm{d}z, \qquad (2) \end{aligned}$$

where Q is the generator of the Markov process, commonly denoted as the rate matrix. Equation (2) corresponds to $[\mathbf{E}(\exp(Qz))]_{x,y}$, that is, the moment-generating function of Z (restricted to the pair x,y). Analytical solutions to the integral in this equation are not available in the general case. However, for specific probabilistic distributions of Z, it is possible to get a closed-form formula. In particular, if Z is distributed as a gamma density with shape $\alpha=\mathbf{E}^2(Z)/\mathbf{V}(Z)$ and scale $\beta=\mathbf{V}(Z)/\mathbf{E}(Z)$, we have

$$P(X_t=x|X_0=y,r_0,r_t,t)=[(I-\beta Q)^{-\alpha}]_{x,y}, \qquad (3)$$

which can be computed efficiently by calculating eigenvalues and eigenvectors of the rate matrix Q (using $f(Q)=Uf(\Lambda)U^{-1}$, where U and Λ are the matrices of eigenvectors and eigenvalues of the rate matrix Q, respectively). In other words, the length of a branch is here replaced by 2 parameters, α and β, that describe the probabilistic distribution of the number of substitutions per site. Note that the same technique was proposed by others (Huelsenbeck et al., 2008; Suchard et al., 2002, 2003) in the context of Bayesian estimation of phylogenies where the prior distribution of the average branch rate is exponentially or gamma distributed.

When estimating species divergence dates in a Bayesian framework, the goal is to reconstruct the joint posterior density of the vectors of node ages ($\mathbf{T}$) and node rates ($\mathbf{R}$). Using Bayes theorem, we have

$$P(\mathbf{T},\mathbf{R}|D)\propto\int_{\mathbf{Z}} P(D|\mathbf{Z},\mathbf{T})P(\mathbf{Z}|\mathbf{R},\mathbf{T})P(\mathbf{R}|\mathbf{T})P(\mathbf{T})\mathrm{d}\mathbf{Z}, \qquad (4)$$

where $\mathbf{Z}$ is the vector of average rates along edges and D corresponds to the sequence alignment. $P(D|\mathbf{Z},\mathbf{T})$ is Felsenstein's likelihood (Felsenstein, 1981), $P(\mathbf{Z}|\mathbf{R},\mathbf{T})$ is the probability density of average rates along edges given node rates and times, $P(\mathbf{R}|\mathbf{T})$ is the probability density of node rates along the tree, and $P(\mathbf{T})$ is the prior on node heights. The next section describes a model of rate trajectory for which the distribution of the average substitution rate along a branch is well approximated by a gamma density, making the integral $\int_{\mathbf{Z}}P(D|\mathbf{Z},\mathbf{T})P(\mathbf{Z}|\mathbf{R},\mathbf{T})\mathrm{d}\mathbf{Z}$ calculable analytically using the standard sum-product algorithm in which each term is given by Equation (3). The model itself is first introduced. The mean and variance of the distribution of the average rate are given next.

From Rate Trajectories to Averages

The standard Brownian process is a continuous-time stochastic process that has been widely used in physics or economics to describe the variation of random quantities with time. According to this model, the variable of interest has normally distributed independent increments, that is, the differences between the value of the variable at the end and the beginning of each of two nonoverlapping time intervals are two independent random variables with normal distributions. Logarithm of substitution rates can take on negative or positive values and a standard Brownian process is therefore suitable in this respect. Also, the logarithm function is a widely used transformation to normality. Hence, although rate trajectories might not be well described by a Brownian process, their logarithms might fit that model better. It is therefore not surprising that the geometric Brownian process, which is a Brownian process of the log-transformed variable of interest, has already been well studied by others (Kishino et al., 2001; Rannala and Yang, 2007; Thorne et al., 1998). The section below describes the derivation of the mean and variance of the average substitution rate along a branch given that the rate trajectory is governed by a geometric Brownian process which values at the beginning and the end of the branch of interest are fixed. Such a process is generally referred to as a bridged geometric Brownian process. The "bridged" term being justified here by the fact that the rates at the beginning and the end of a branch are "tied down" to specific values (i.e., the node rates).

The average substitution rate along a branch is given by the following integral:

$$Z=\frac{1}{t}\int_{s=0}^{s=t} \mathrm{e}^{V_s}\mathrm{d}s, \qquad (5)$$

where $V_s=\log_{\mathrm{e}}(R_s)$ is a normally distributed random variable. The expected value of Z is therefore given by

$$\mathbf{E}(Z)=\frac{1}{t}\int_{s=0}^{s=t} \mathbf{E}(\mathrm{e}^{V_s})\mathrm{d}s. \qquad (6)$$

This last formula illustrates a convenient mathematical property of the geometric Brownian process. Indeed, the expected value of the average rate is a function of the moment-generating function of a normally distributed random variable. The previous formula therefore simplifies to

$$\mathbf{E}(Z)=\frac{1}{t}\int\limits_{s=0}^{s=t}\exp\left(\mathbf{E}(V_s)+\frac{1}{2}\mathbf{V}(V_s)\right)\mathrm{d}s. \tag{7}$$

Having V_s normally distributed not only makes the moment-generating function of $\mathbf{E}(Z)$ a simple function of $\mathbf{E}(V_s)$ and $\mathbf{V}(V_s)$, we will also see below that closed-form formulas for these two quantities are available that make the integral calculable. Because V_s defines a Brownian bridge starting at value $v_0=\log_{\mathrm{e}}(r_0)$ and stopping at $v_t=\log_{\mathrm{e}}(r_t)$, we have

$$\mathbf{E}(V_s)=v_0+\frac{(v_t-v_0)}{t}s \tag{8}$$

and

$$\mathbf{V}(V_s)=\frac{\nu(t-s)s}{t}, \tag{9}$$

where ν is the rate autocorrelation parameter. The derivation of these last 2 equations is given in the Appendix. We therefore have

$$\mathbf{E}(Z)=\frac{1}{t}\exp(v_0)\int\limits_{s=0}^{s=t}\exp\left(\frac{(v_t-v_0)}{t}s+\frac{1}{2}\frac{\nu(t-s)}{t}s\right)\mathrm{d}s. \tag{10}$$

It is worth noting that when $\nu=0$, that is, rates do not fluctuate at all, then $r_0=r_t$ and therefore, from Equation (10), $\mathbf{E}(Z)=1/t\ \exp(v_0)t=\exp(v_0)=r_0$, as expected. Also, Equation (10) shows that as ν increases, $\mathbf{E}(Z)$ increases too. This observation also makes sense because, for large values of ν, rate trajectories are likely to reach high peaks which will dominate in the calculation of the average. This property of the proposed model is interesting because it suggests that part of the information conveyed by the data on rate autocorrelation comes from the total amount of substitutions accumulated, a quantity that can be relatively well estimated from sequence alignments.

In order to compute the previous integral efficiently, we use a Taylor series approximation of the exponential function. We therefore have

$$\mathbf{E}(Z)\simeq\frac{1}{t}\exp(v_0)\int\limits_{s=0}^{s=t}\sum_{k=0}^{N>0}\frac{1}{k!}\left(\frac{(v_t-v_0)}{t}s+\frac{1}{2}\frac{\nu(t-s)}{t}s\right)^k\mathrm{d}s, \tag{11}$$

the analytical expression of which can be derived for various values of N using Maple for instance.

As for the variance of Z, we use a similar approach. Let U_s define a Brownian bridge with $U_0=0$ and $U_t=0$. We have

$$\begin{aligned}\mathbf{V}(Z)&=\frac{1}{t^2}\mathbf{V}\left(\int\limits_{s=0}^{s=t}\exp(V_s)\mathrm{d}s\right)\\&=\frac{1}{t^2}\mathbf{V}\left(\int\limits_{s=0}^{s=t}\exp\left[v_0+\frac{(v_t-v_0)}{t}s+U_s\right]\right)\mathrm{d}s\\&=\frac{1}{t^2}\exp(2v_0)\mathbf{V}\left(\int\limits_{s=0}^{s=t}\exp\left[\frac{(v_t-v_0)}{t}s+U_s\right]\mathrm{d}s\right)\\&=\frac{1}{t^2}\exp(2v_0)\int\limits_{a=0}^{a=t}\int\limits_{b=0}^{b=t}\mathrm{Cov}\Big(\exp\left[\frac{(v_t-v_0)}{t}a+U_a\right],\\&\qquad\exp\left[\frac{(v_t-v_0)}{t}b+U_b\right]\Big)\mathrm{d}a\ \mathrm{d}b\\&=\frac{1}{t^2}\exp(2v_0)\int\limits_{a=0}^{a=t}\int\limits_{b=0}^{b=t}\exp\left[\frac{(v_t-v_0)}{t}(a+b)\right]\\&\qquad\mathrm{Cov}\Big(\exp(U_a),\exp(U_b)\Big)\mathrm{d}a\ \mathrm{d}b.\end{aligned} \tag{12}$$

We now focus on the expression of $\mathrm{Cov}\Big(\exp(U_a),\exp(U_b)\Big)$. Because U_a and U_b are normally distributed random variables, we have

$$\begin{aligned}\mathbf{E}(\exp(U_a))&=\exp\left[\mathbf{E}(U_a)+\frac{1}{2}\mathbf{V}(U_a)\right]\\&=\exp\left[\frac{\nu a(t-a)}{2t}\right].\end{aligned} \tag{13}$$

Therefore,

$$\begin{aligned}\mathbf{E}(\exp(U_a))\mathbf{E}(\exp(U_b))&=\exp\left[\frac{\nu a(t-a)}{2t}\right]\exp\left[\frac{\nu b(t-b)}{2t}\right]\\&=\exp\left[\frac{\nu}{2t}\Big(a(t-a)+b(t-b)\Big)\right].\end{aligned} \tag{14}$$

Also, we have

$$\begin{aligned}&\mathbf{E}(\exp(U_a)\exp(U_b))\\&=\exp\left[\mathbf{E}(U_a+U_b)+\frac{1}{2}\mathbf{V}(U_a+U_b)\right]\\&=\exp\Big[\mathbf{E}(U_a)+\mathbf{E}(U_b)+\frac{1}{2}(\mathbf{V}(U_a)\\&\qquad+\mathbf{V}(U_b)+2\mathrm{Cov}(U_a,U_b))\Big]\\&=\exp\left[\frac{1}{2}(\mathbf{V}(U_a)+\mathbf{V}(U_b))+\mathrm{Cov}(U_a,U_b)\right].\end{aligned} \tag{15}$$

We now focus on the covariance term $\mathrm{Cov}(U_a, U_b)$ in Equation (15). Let W_s define the standard Brownian process. We have

$$\begin{aligned}\mathrm{Cov}(U_a,U_b) &= \mathbf{E}(U_a U_b) - \mathbf{E}(U_a)\mathbf{E}(U_b)\\ &= \mathbf{E}(U_a U_b)\\ &= \mathbf{E}\left(\left(W_a - \frac{a}{t}W_t\right)\left(W_b - \frac{b}{t}W_t\right)\right)\\ &= \mathbf{E}(W_a W_b) - \frac{b}{t}\mathbf{E}(W_a W_t) - \frac{a}{t}\mathbf{E}(W_b W_t)\\ &\quad + \frac{ab}{t^2}\mathbf{E}(W_t^2). \end{aligned} \tag{16}$$

If $a \le b$, we have

$$\begin{aligned}\mathrm{Cov}(U_a,U_b) &= \nu a - \frac{b}{t}\nu a - \frac{a}{t}\nu b + \frac{ab}{t^2}\nu t\\ &= \frac{\nu a(t-b)}{t}. \end{aligned} \tag{17}$$

Otherwise,

$$\mathrm{Cov}(U_a,U_b) = \frac{\nu b(t-a)}{t}. \tag{18}$$

Therefore, if $a \le b$, we have

$$\begin{aligned}&\mathbf{E}(\exp(U_a)\exp(U_b))\\ &= \exp\left[\frac{\nu a(t-a)}{2t} + \frac{\nu b(t-b)}{2t} + \frac{\nu a(t-b)}{t}\right]\\ &= \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b) + 2a(t-b)\Big)\right]. \end{aligned} \tag{19}$$

Otherwise,

$$\begin{aligned}&\mathbf{E}(\exp(U_a)\exp(U_b))\\ &= \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b) + 2b(t-a)\Big)\right]. \end{aligned} \tag{20}$$

We now have an expression for each term required to derive $\mathrm{Cov}(\exp(U_a), \exp(U_b))$. If $a \le b$, then

$$\begin{aligned}&\mathrm{Cov}(\exp(U_a),\exp(U_b))\\ &= \mathbf{E}(\exp(U_a + U_b)) - \mathbf{E}(\exp(U_a))\mathbf{E}(\exp(U_b))\\ &= \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b) + 2a(t-b)\Big)\right] -\\ &\quad \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b)\Big)\right]. \end{aligned} \tag{21}$$

Otherwise,

$$\begin{aligned}&\mathrm{Cov}(\exp(U_a),\exp(U_b))\\ &= \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b) + 2b(t-a)\Big)\right]\\ &\quad - \exp\left[\frac{\nu}{2t}\Big(a(t-a) + b(t-b)\Big)\right]. \end{aligned} \tag{22}$$

Going back to Equation (12), we break down the integral in two parts so as to consider the two cases, that is, $a \le b$ and $a > b$, separately:

$$\begin{aligned}\mathbf{V}(Z) &= \frac{1}{t^2}\exp(2v_0)\int_{b=0}^{b=t}\int_{a=0}^{a=b} \exp\left[\frac{(v_t - v_0)}{t}(a+b)\right]\\ &\quad \mathrm{Cov}\Big(\exp(U_a), \exp(U_b)\Big)\mathrm{d}a\,\mathrm{d}b\\ &\quad + \frac{1}{t^2}\exp(2v_0)\int_{a=0}^{a=t}\int_{b=0}^{b=a} \exp\left[\frac{(v_t - v_0)}{t}(a+b)\right]\\ &\quad \mathrm{Cov}\Big(\exp(U_a), \exp(U_b)\Big)\mathrm{d}a\,\mathrm{d}b, \end{aligned} \tag{23}$$

and replace the covariance terms by their expressions as given in Equations (21) and (22). After a few mathematical simplifications, we obtain the following expression for the variance:

$$\begin{aligned}\mathbf{V}(Z) &= \frac{2}{t^2}\exp(2v_0)\int_{a=0}^{a=t}\int_{b=0}^{b=a} \exp\left[\frac{(v_t - v_0)}{t}(a+b)\right.\\ &\quad \left. + \frac{\nu}{2t}\Big(a(t-a) + b(t-b) + 2b(t-a)\Big)\right]\mathrm{d}a\,\mathrm{d}b\\ &\quad - \frac{2}{t^2}\exp(2v_0)\int_{a=0}^{a=t}\int_{b=0}^{b=a} \exp\left[\frac{(v_t - v_0)}{t}(a+b)\right.\\ &\quad \left. + \frac{\nu}{2t}\Big(a(t-a) + b(t-b)\Big)\right]\mathrm{d}a\,\mathrm{d}b. \end{aligned} \tag{24}$$

Note that Equation (24) satisfies $\mathbf{V}(Z) = 0$ if $\nu = 0$, as expected. Here again, the exponential terms can be approximated as a Taylor series, which makes the double integral calculable analytically.

In practice, a 10th order approximation was used for the series expansion of the mean and variance of Z. The accuracy of this approximation was assessed by simulating the process governing rate trajectories. We generated rate trajectories according to a bridged geometric Brownian process and approximated the probability density function of Z for various values of r_0, r_t, ν, and t. For each combination of parameters, 10 000 rate trajectories were simulated. The average rate was recorded for each of them. These simulations were performed using the R software (R Development Core Team, 2011). Figure 1 shows the quantile–quantile plots for the distributions of the simulated values of Z and the corresponding gamma densities with mean and variance set using the 10th order approximations of Equations (11) and (24). The corresponding distributions themselves are also displayed in the embedded graphics. These graphics show that the gamma distribution provides a good fit to the actual distribution of Z. They also demonstrate that the 10th order approximation for the mean and variance of the gamma distribution is satisfactory. Therefore, the mean and variance of Z can be well approximated using Equations (11) and (24), respectively, and it is

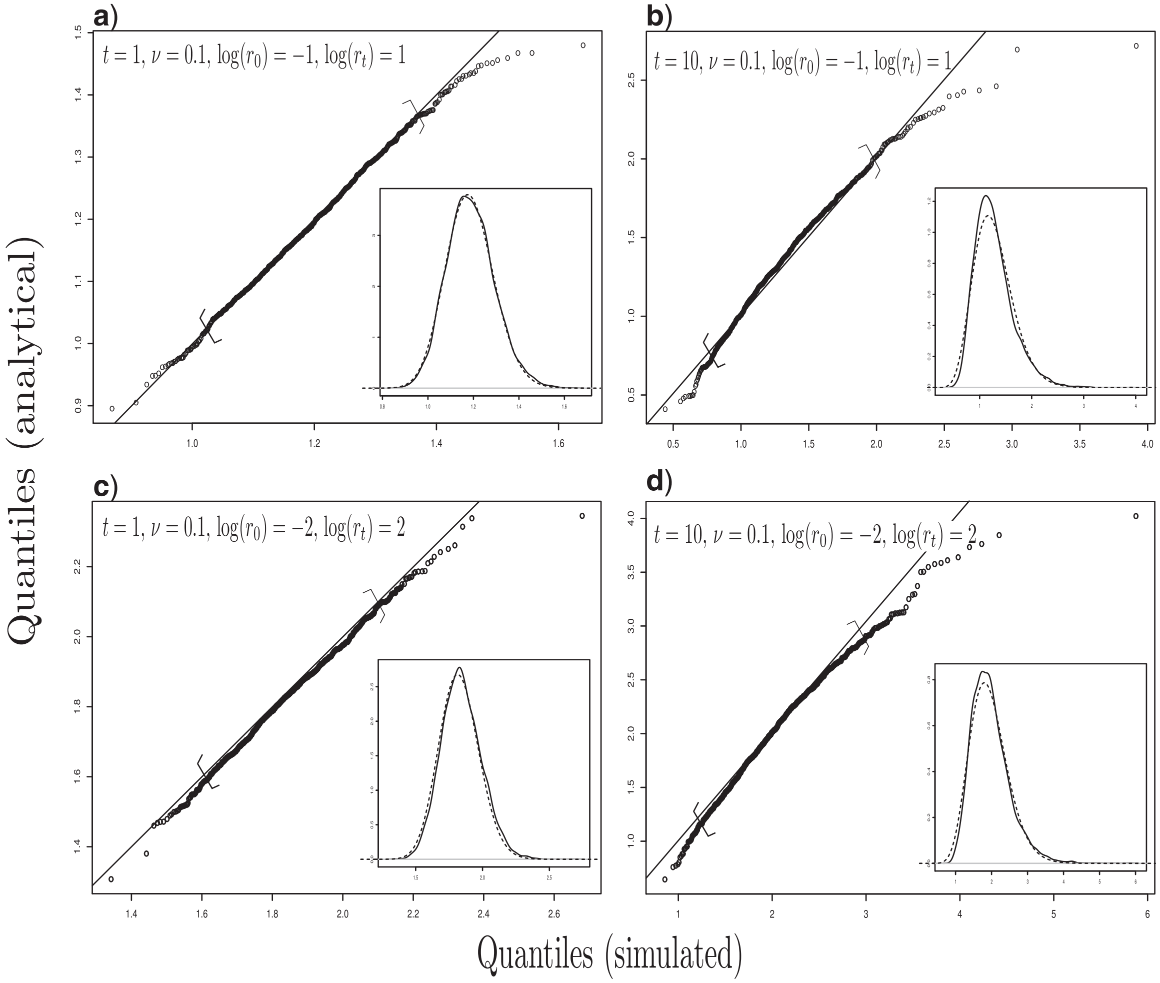

FIGURE 1. Quantile-quantile plots of the simulated vs. gamma-approximated distribution of the average substitution rate. t is the branch length, r_0 and r_t are the rates at its 2 extremities. ν is the scaling factor governing the variance of the Brownian process (see text). The brackets indicate the 5% and 95% quantiles. The solid line in the embedded graphics corresponds to the estimated density from the simulated process. The dashed line is the density obtained using the analytical approximation.

straightforward to calculate the probability density of this random variable for any value it takes.

Joint Prior of Node Ages

The previous section focuses on the distribution of substitution rate on edges given the rest of the parameters of the phylogenetic model, including node heights. This section deals with the prior distribution of node heights. We here follow the steps of Rannala and Yang (1996) and Yang and Rannala (1997, 2005) (but see also Nee, 2001; Stadler, 2009, 2010). These authors give the joint prior density of divergence times under a birth–death process (Kendall 1948) generalized to account for species sampling. The model considered in this study corresponds to the pure-birth or Yule process. It is a special case of the one proposed by Yang and Rannala (2005) as it assumes complete sampling and also considers that species never go extinct (i.e., $\mu = 0$ and $\rho = 1$). However, Yang and Rannala's results apply to the case where only a single sampling event took place. Our model accounts for the case where groups of taxa were sampled on several occasions over a certain period of time, as is often the case with viruses for instance (e.g., see Shankarappa et al. 1999).

In the case where data are not sampled through time and for the Yule model, the conditional density of $n-2$ internal node heights conditioned on the height of the root node is given by the order statistic of $n-2$ independent variables distributed as a right-truncated exponential with parameter λ, where λ is the birth rate (Yang and Rannala, 2006). The truncation results from the fact that internal node heights are constrained to lie between the present (time 0) and the age of the root node. More formally, we have

$$f(\mathbf{T}_{-r}|T_r, \lambda, \tau) \propto \prod_{i \neq r}^{n-1} \frac{\lambda e^{-\lambda |T_i|}}{(1 - e^{-\lambda |T_r|})}, \tag{25}$$

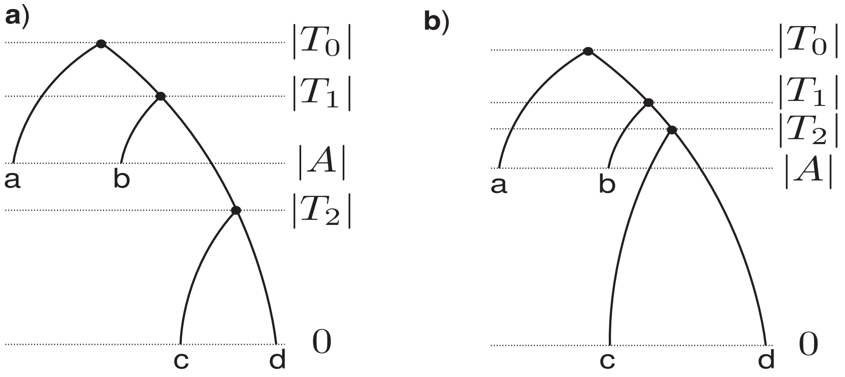

FIGURE 2. Two rooted 4-taxon trees with taxa observed at times 0 and $|A|$.

where $\mathbf{T}_{-r}$ corresponds to the vector of all internal node heights (denoted as $\mathbf{T}$) after having removed T_r, the age of the root node, from it. τ is the tree topology. The proportionality factor is a function of τ. Because the topology is considered as fixed in our study, we can safely ignore it.

In order to explain how Equation (25) needs to be modified to account for multiple sampling events, we consider an example with four taxa and two time points. In Figure 2, sequences were collected at time $|A|$ and 0. Given this constraint, $|T_0| \geq |T_1| \geq |A|$ and $|T_1| \geq |T_2| \geq 0$. Hence, although $|T_1|$ has to be greater than $|A|$, $|T_2|$ can be smaller (Fig. 2a) or larger (Fig. 2b) than $|A|$. Under a pure-birth model, the time required to go from x to $x+1$ lineages is exponentially distributed with parameter $x\lambda$. For the tree on the left-hand side, the joint density of T_0, T_1, and T_2 and the tree topology τ is therefore given by the following formula:

$$\begin{aligned} f_{\mathrm{L}}(T_0, T_1, T_2|\lambda, A, \tau) = \; & \frac{1}{2}(2\lambda \exp(-2\lambda|T_0 - T_1|)) \times \\ & \exp(-3\lambda|T_1 - A|) \times \\ & \lambda \exp(-\lambda|A - T_2|) \times \\ & \exp(-2\lambda|T_2|). \end{aligned} \tag{26}$$

For the tree on the right-hand side, we have

$$\begin{aligned} f_{\mathrm{R}}(T_0, T_1, T_2|\lambda, A, \tau) = \; & \frac{1}{2}(2\lambda \exp(-2\lambda|T_0 - T_1|)) \times \\ & \frac{1}{3}(3\lambda \exp(-3\lambda|T_1 - T_2|)) \times \\ & \exp(-4\lambda|A - T_2|) \times \\ & \exp(-2\lambda|A|). \end{aligned} \tag{27}$$

Note that $f_{\mathrm{L}}(T_0, T_1, T_2|\lambda, A, \tau) = f_{\mathrm{R}}(T_0, T_1, T_2|\lambda, A, \tau)$. The marginal density of T_0 is obtained by integrating over all possible values for T_1 and T_2 for the two trees. We have

$$\begin{aligned} f(T_0|\lambda, A, \tau) = & \int_{|T_1|=|A|}^{|T_0|} \int_{|T_2|=0}^{|A|} f_{\mathrm{L}}(T_0, T_1, T_2|\lambda, A, \tau) \mathrm{d}|T_1| \mathrm{d}|T_2| \\ & + \int_{|T_1|=|A|}^{|T_0|} \int_{|T_2|=|A|}^{|T_1|} f_{\mathrm{R}}(T_0, T_1, T_2|\lambda, A, \tau) \mathrm{d}|T_1| \mathrm{d}|T_2|. \end{aligned} \tag{28}$$

After expanding and factorizing the integrals, we end up with the following formula for the conditional density of T_1 and T_2 given T_0:

$$\begin{aligned} f(T_1, T_2|T_0, A, \lambda) & \propto \frac{f(T_1, T_2, T_0|A, \lambda, \tau)}{f(T_0|A, \lambda, \tau)} \\ & = \frac{\lambda \mathrm{e}^{-\lambda|T_1|}}{(\mathrm{e}^{-\lambda|A|} - \mathrm{e}^{-\lambda|T_0|})} \times \frac{\lambda \mathrm{e}^{-\lambda|T_2|}}{(1 - \mathrm{e}^{-\lambda|T_0|})}. \end{aligned} \tag{29}$$

Let $|A_i|$ denote the lower bound for $|T_i|$, the previous formula then generalizes to any tree with n serially sampled taxa. We have

$$f(\mathbf{T}_{-r}|T_r, \lambda, \mathbf{A}) \propto \prod_{i \neq r}^{n-1} \frac{\lambda \mathrm{e}^{-\lambda|T_i|}}{(\mathrm{e}^{-\lambda|A_i|} - \mathrm{e}^{-\lambda|T_r|})}, \tag{30}$$

where $\mathbf{A}$ is a vector giving the times at which sampling events occurred. This last formula indicates that the conditional density of the $n-2$ internal node heights knowing the height of the root node is given by the order statistic of $n-2$ independent variables distributed as a right- and left-truncated exponential distribution with parameter λ. The left truncation stems from the lower bounds imposed to certain nodes heights when sequences were collected during successive events (e.g., the lower bound for $|T_1|$ in Fig. 2 is $|A|$).

Our treatment of calibration nodes is less sophisticated than that provided by Yang and Rannala (2005). The marginal distribution of the calibration node heights is governed by the Yule process, excluding values that fall below (above) the lower (upper) bounds for these nodes.

Let **L** and **U** be the vectors defining the lower and upper bounds for the absolute value of each internal node age, respectively. Let $I_i = \min(U_i, |T_r|)$ and $J_i = \max(L_i, |A_i|)$, we have

$$f(\mathbf{T}_{-r}|T_r, \lambda, \mathbf{A}, \mathbf{U}, \mathbf{V}) \propto \prod_{i \neq r}^{n-1} \frac{\lambda e^{-\lambda |T_i|}}{(e^{-\lambda J_i} - e^{-\lambda I_i})}. \qquad (31)$$

Altogether, the joint density of node heights used in this study is given by

$$f(\mathbf{T}|\lambda, \mathbf{A}, \mathbf{U}, \mathbf{V}) = f(\mathbf{T}_{-r}|T_r, \lambda, \mathbf{A}, \mathbf{U}, \mathbf{V}) f(T_r|\lambda, \mathbf{A}, \mathbf{U}, \mathbf{V}), \qquad (32)$$

where $f(\mathbf{T}_{-r}|T_r, \lambda, \mathbf{A}, \mathbf{U}, \mathbf{V})$ is given by Equation (31) and $f(T_r|\lambda, \mathbf{A}, \mathbf{U}, \mathbf{V})$, the marginal distribution of the age of the root node, is uniform in $[I_r, J_r]$.

Data Sets

Three real sequence alignments were considered in this study. The first consists of HIV-1 sequences collected in a single infected individual at ten successive time points (Shankarappa et al., 1999). The 87 sequences from the *env* gene, 561 nucleotide long, were first processed with the software PhyML (Guindon et al., 2010) in order to estimate the maximum-likelihood phylogeny under the GTR + Γ_4 model (Tavaré, 1986; Yang, 1994), using an SPR search for the best tree topology. The tree topology was then considered as fixed in the estimation of divergence dates. The calibration nodes here correspond to the tips of the tree for which the time of collection (in months) was recorded over a period of 74 months. The second data set consists of nucleotide sequences from Caviomorph rodents and Platyrrhine primates previously analysed by Poux et al. (2006). Sixty-two homologous sequences from three nuclear genes and a total of 3768 sites are considered here. Nine calibration points, including prior information on the time at the root node, are available. The third data set was assembled and analysed by Wahlberg (2006). It is made of 59 nucleotide sequences, 2936 character long, consisting of one mitochondrial gene and two nuclear genes collected from the butterfly subfamily Nymphalinae. Four calibration points were available for this data set. The last two alignments and the corresponding tree topologies were retrieved from Treebase (Sanderson et al., 1994).

Results

We use a parametric bootstrap approach to assess the accuracy and precision of the estimated autocorrelation parameter and the node ages when rates evolve along a tree using one of two distinct models. The first is the standard deterministic model where the average rate along a branch is the arithmetic average of the rates at its two extremities, that is, average rates along edges, and therefore edge lengths, are not considered as random variables here. The second is the stochastic model where the average rate is a gamma-distributed random variable, with the shape and scale parameter of this distribution derived analytically (see "Materials and Methods"). For each real sequence data set (see Materials and Methods) and each model of rate evolution, a MCMC algorithm is used to sample model parameters from their prior distribution. Once the effective sample sizes for the autocorrelation parameter, the age of the root node and the overall substitution rate have all reached 100, model parameters are recorded and nucleotide sequences are generated along the corresponding tree using the HKY model of substitution (Hasegawa et al., 1985) with nucleotide frequencies set to the values estimated from the actual data set and transition/transversion ratio sampled from its prior distribution. For each of the three real-world data sets considered in this study, 500 sequence alignments are simulated this way. For each of these 500 alignments, the same MCMC algorithm is used to sample from the posterior distribution of model parameters under the HKY model for both models of rate evolution. Here again, the sampler is run as long as the effective sample sizes for any of the autocorrelation parameter, the age of the root node or the overall substitution rate do not exceed 100. Additionally, we have used the R library `coda` (Plummer et al., 2010) to run convergence diagnostics on a subset of the MCMCs. None of these tests indicated a lack of convergence.

Accuracy

Figure 3 (top row) shows scatterplots of the true values of the autocorrelation parameter (a), the age of the root node (b), and the age of the other internal nodes in the tree (c) for data simulated under the stochastic model of rate evolution, against the corresponding values estimated under the deterministic model. Sequences were simulated along the HIV-1 phylogeny (see Materials and Methods section). Here, the estimated values correspond to the posterior medians of the parameters of interest. Although the accuracy of the rate autocorrelation estimation is satisfactory for small values of this parameter (corresponding to slight deviations from the molecular clock), larger values are more difficult to infer when branch lengths are derived using the deterministic approach. The same observation applies to the age of the root node: young root ages are recovered accurately, while the accuracy drops sharply for older root ages. Also note that root age tends to be underestimated, for example, for actual ages close to 120 time units, all the estimates are younger than 115 time units. The accuracy for the other internal nodes is very satisfactory and using the deterministic approximation does not seem to impact on the accuracy of the estimates here.

Figure 3 (middle row) shows the accuracy of parameter estimates for Poux et al. data set. Here again, the accuracy decreases with increasing values of the rate autocorrelation parameter. This decrease in accuracy also comes with an increase in bias, with large values

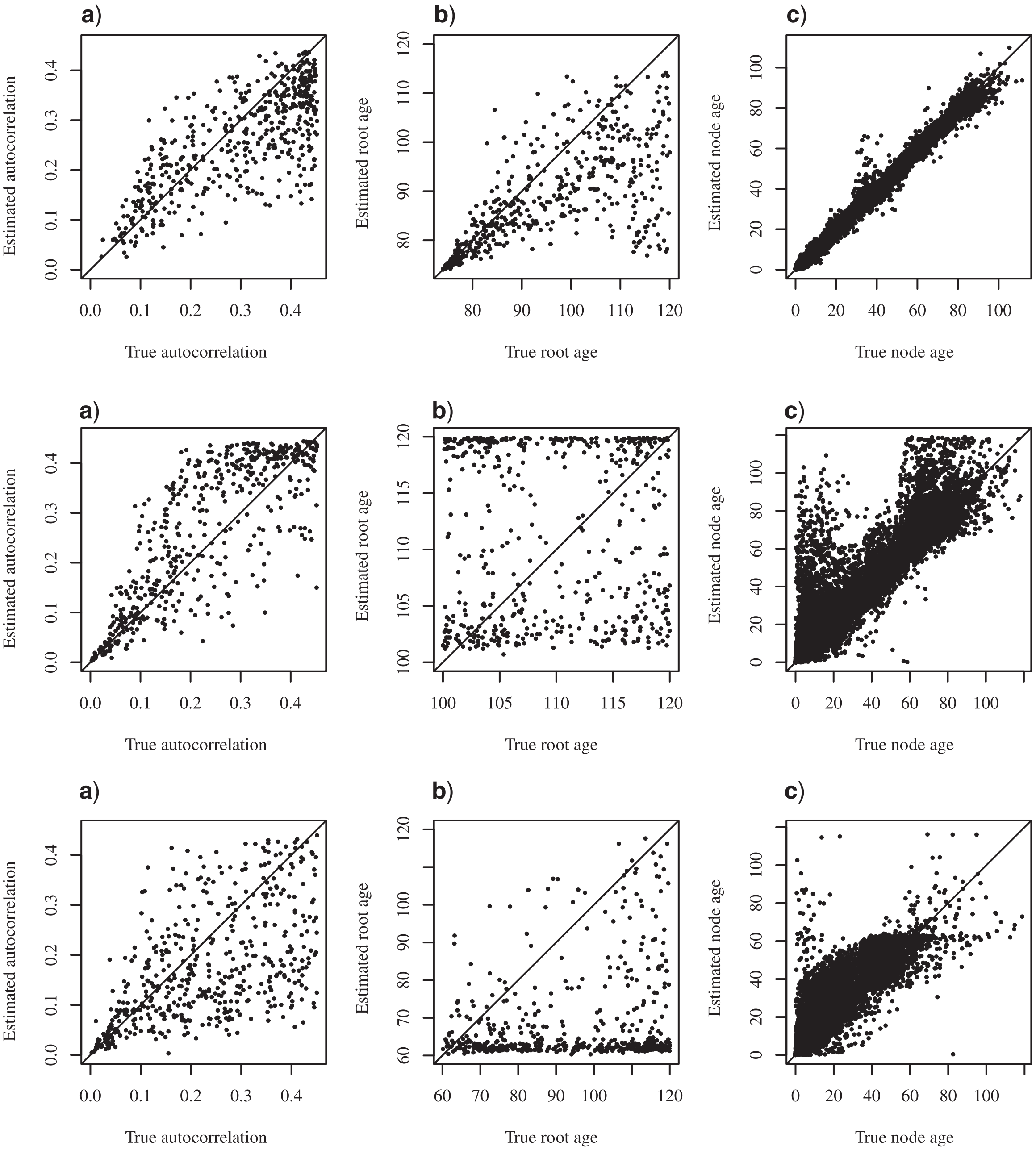

FIGURE 3. Accuracy of the deterministic approach for estimating the rate autocorrelation parameter (a), the age of the root node (b), and other internal node ages (c) (top row: HIV-1 data set; middle row: Poux et al. data set; bottom row: Wahlberg data set). Sequences were simulated under the stochastic model. Parameters were estimated under the deterministic model. The true parameter values are on the x-axis. The medians of the posterior densities for each parameter are on the y-axis. The solid line corresponds to the first diagonal.

of the parameter tending to be overestimated. As for the root age, the estimates obtained by applying the deterministic approach to data generated using the stochastic method show very poor accuracy, with no obvious correlation between the estimated and the true values of this parameter. Ignoring the stochasticity of average rates along edges here amounts to a total inability of the model to extract phylogenetic signal from the data. Regarding the other internal nodes, the correlation between estimated and true values is much stronger. However, some node ages are clearly overestimated. For instance, in our simulations, some nodes that were 60 time units old were estimated to be 120 time units old. Also note that the uneven distribution of values along both axes is explained by the constraints on calibration nodes, setting boundaries of internal node heights in both simulated and estimated trees.

Figure 3 (bottom row) shows the accuracy of parameter estimates for Wahlberg data set. A pattern similar to that observed for the other two data sets is obtained for the rate autocorrelation parameter. Also, the age of the root node is almost systematically underestimated. The inaccuracy is such that the correlation between the true and estimated values of this parameter is virtually nonexistent. The other internal node estimates show better correlation with the corresponding true values but, here again, some node ages are dramatically overestimated.

Precision

Figure 4 (top row) gives the 95% credibility intervals for the autocorrelation parameter (a), the age of the root node (b), and the age of the other internal nodes in the tree (c) for the HIV-1 data set. The intervals defined by the solid lines were obtained from data simulated under the deterministic model, with the same deterministic approach used to estimate the parameters. The dashed lines were derived from data simulated and parameter estimated using the stochastic approach. The width of the credibility intervals obtained for the autocorrelation parameters are clearly smaller for the stochastic model, indicating more precise estimates of this parameter under the stochastic approach compared with the deterministic one (one-sided Welch two-sample *t*-test, P-value $<2.2\times10^{-16}$). Also note that the precision of the estimates for this parameter does not seem to be affected by its actual value: the difference between the 2.5% and 97.5% quantiles remains approximately constant as one moves along the x-axis. This observation indicates that it is possible to quantify the deviation from the molecular clock constraint in a precise manner, even in cases where the variation of rates across lineages is strong. As for the node age estimates, including the root node (Fig. 4b, c, top row), the precision of the estimates are virtually identical with both the deterministic and the stochastic approaches. For internal node heights, however, the intervals obtained with the stochastic approach are statistically smaller than those obtained with the deterministic ones (means: 3.82 vs. 4.04 time units, P-value $<2.2\times10^{-16}$). The difference between the two approaches regarding the precision of these node age estimates, while being statistically significant, is therefore very slight if not negligible and mostly reflect the very large sample sizes used here.

Figure 4 (middle and bottom rows) shows the credibility intervals obtained for Poux et al. and Wahlberg data sets, respectively. Here again, the estimates of the rate autocorrelation parameter are much more precise under the stochastic than the deterministic approach (P-value $<2.2\times10^{-16}$ in both cases). Contrary to our observation for the HIV-1 data set, the estimation of the root age is more precise under the stochastic than the deterministic approach (Poux et al. data, means: 12.37 vs. 14.87 time units, P-value $<2.2\times10^{-16}$; Wahlberg data, means: 31.28 vs. 35.58 times units, P-value $=3.5\times10^{-09}$). However, the precision of the other internal node age estimates are very similar with both approaches, even though the stochastic approach returns statistically smaller intervals on average in both cases (Poux et al. data, means: 5.89 vs. 7.03, P-value $<2.2\times10^{-16}$; Wahlberg data, means: 4.89 vs. 5.55 time units, P-value $<2.2\times10^{-16}$)

Discussion

This article focuses on how rate trajectories describing the variation of substitution rates along a phylogeny are translated into average evolutionary rates along edges. Current approaches approximate average rates using a deterministic function of the rates at the nodes of the tree, ignoring the stochasticity of the process. Assuming rate trajectories are governed by a geometric Brownian process, the average rate along any given branch can be well approximated by a gamma distribution. We were able to derive the mean and variance of this distribution given the node rates using a semi-analytical approach. Moreover, it is relatively straightforward to calculate matrices of character change probability along a branch whose length is considered as a gamma-distributed random variable. The combination of these two properties makes the geometric Brownian model of rate evolution particularly attractive from a practical perspective because the computational load required by the stochastic approach is virtually identical to that of the deterministic one.

Our results show that ignoring the stochasticity of average rates along branches impacts negatively on the accuracy of the estimated node ages and rate autocorrelation parameter. Most importantly, the large overestimation of some internal node ages observed with two of the three data sets considered in this study is concerning. Interestingly however, in their review article comparing species divergence dates estimated using molecular versus palaeontological data, Benton and Ayala (2003) note that the molecular age estimates for several important divergence dates (including that of the origins of metazoans) are about twice as old as the oldest fossils. This observation is well in line with our results. Further investigations are required to confirm that ignoring the stochasticity of average rates could partly be responsible for the discrepancy between molecular and palaeontological date estimates. Given the equivalent computational costs compared with the standard approach, we believe that future studies relying on molecular data to estimate species divergence dates should account for the stochasticity of average substitution rates anyway. Note however that both the deterministic and stochastic approach largely agree on all the node ages estimated from the three actual data sets considered in this study (results not shown). These estimates are also very similar to those presented by Poux et al. (2006) and Wahlberg (2006) because the method used to infer species divergence dates in these studies relies on the geometric Brownian model combined to the deterministic approximation of branch lengths. Therefore, although the deterministic approach might

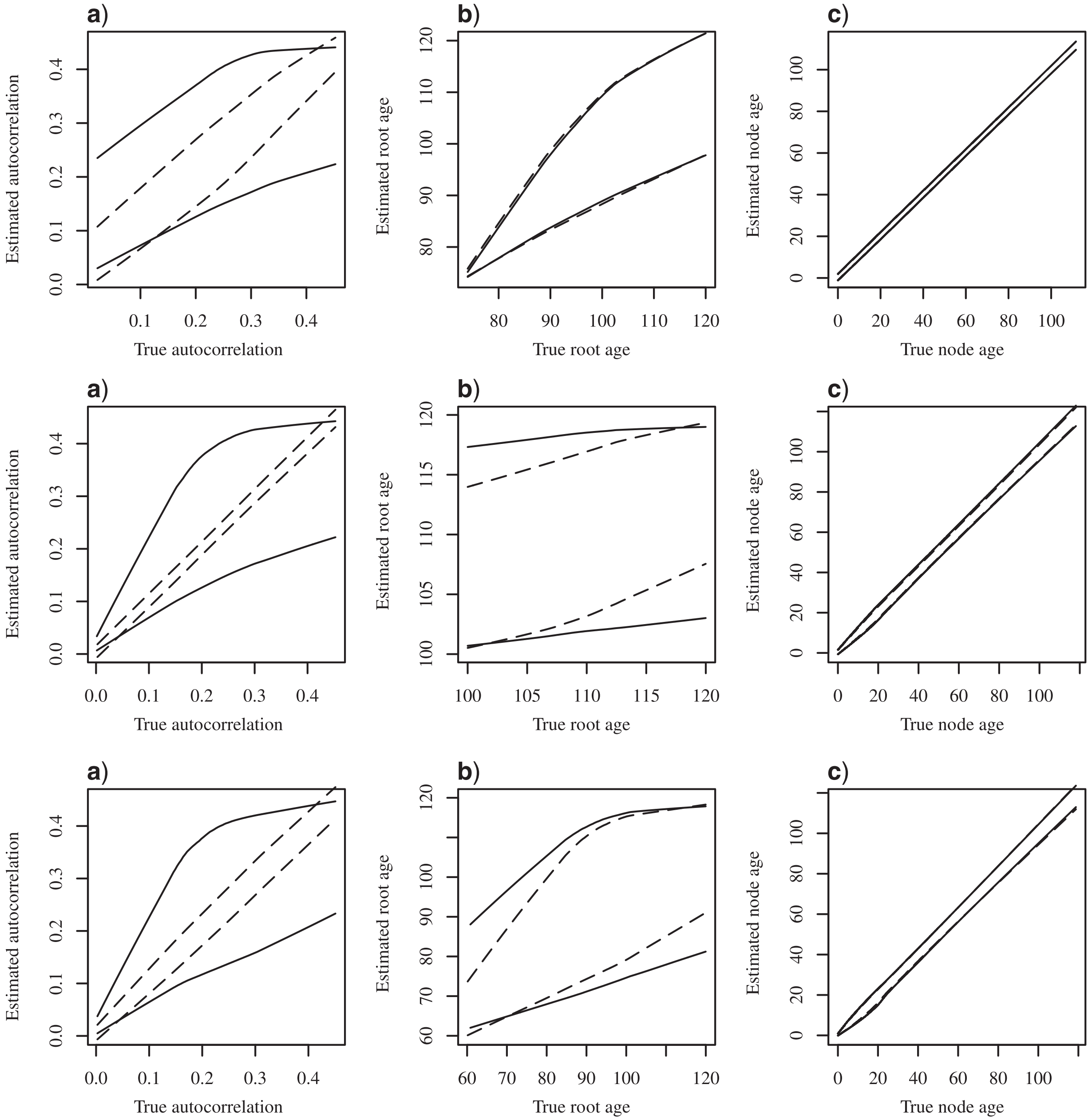

FIGURE 4. Precision of the estimated rate autocorrelation parameter (a), the age of the root node (b), and the age of other internal nodes (c) using the stochastic (dashed lines) and the deterministic (solid lines) approaches (top row: HIV-1 data set; middle row: Poux et al. data set; bottom row: Wahlberg data set). Sequences simulated under the stochastic approach along the original phylogenies were processed using the same stochastic model (dashed lines). Also, sequences simulated under the deterministic approach were processed using the same deterministic approach (solid lines). For each of the 500 simulated data sets and for each method (deterministic and stochastic), the 2.5% and 97.5% quantiles of the posterior distribution of each parameter were calculated. The solid and dashed lines were obtained by plotting a locally weighted polynomial regression (function `lowess` in `R`) for each of the 2 quantiles and methods.

overestimate node ages in some instances, particularly in data sets displaying large deviations from the molecular clock constrain, it is likely that both methods will return very similar node age estimates in most cases.

Our simulations also show that the estimates of the root age and, most notably, the rate autocorrelation parameter are more precise under the stochastic than the deterministic approach. Augmenting the number of parameters in a model generally decreases the precision of the estimates. The increased precision observed here with the most parameter-rich model is best explained by the ability of the stochastic approach to extract relevant signal from the data. Indeed, although the deterministic model estimates rate autocorrelation from node rates only, the stochastic approach also uses information from the average rates along edges.

Increasing the precision of the estimated autocorrelation of substitution rates has important consequences from the biological point of view. For instance, lineage-specific changes in the natural selection patterns or reduction in effective population sizes can explain deviations from the molecular clock constraint. The technique proposed in this study therefore offers the opportunity to revisit the tests of the molecular clock hypothesis using an improved statistical approach. In particular, given an estimate of the autocorrelation parameter derived from the Bayesian approach described in this article, it is possible to detect sudden changes of the substitution rate in specific lineages. Correlating those changes to known phenotypical or environmental variations is relevant from a biological perspective, and this study provides an adequate statistical framework to perform such analysis.

The proposed approach allows rates to vary along the phylogeny and also across sites. The mean of the branch length distribution may indeed be shifted to larger or smaller values using a gamma-distributed multiplicative factor for instance (Yang, 1994). Other models of rate variation across the elements of a data partition, such as different loci for instance, can also be envisaged and implemented using the very same techniques as that used with the more standard models of rate evolution. Also, integrating over geometric Brownian trajectories amounts to modelling heterotachy (Lopez et al., 2002), that is, site-specific patterns of variation of rates along lineages. This feature makes our approach (along with that of Lepage et al., 2006) distinct from the large majority of models describing the heterogeneity of rates along lineages. For instance, the exponential and log-normal models implemented in BEAST, the geometric Brownian model implemented in the softwares Multidivtime or MCMCtree (Yang, 2007), and the OU model combined to the deterministic derivation of branch lengths put forward by Aris-Brosou and Yang (2002), are all homotachous approaches in the sense that the ratio of the average rate of substitution on two distinct edges are constrained to be the same throughout the alignment. Note, however, that our approach should not be considered as "fully heterotachous." Indeed, although the average rates along edges is a random quantity and therefore can vary along sites and lineages according to an heterotachous process, the rates at individual nodes only vary across sites in a homotachous fashion. Such an approach therefore provides an interesting balance between model flexibility and over-fitting the data, even though validating that claim warrants further investigation.

Acknowledgements

I thank Eric Frichot for his help with linear algebra; David Welch and Tanja Stadler for discussions on the prior density of node heights for serially sampled sequences; David Bryant for pointing out interesting features of the proposed approach; and Chieh-Hsi Wu for her comments and corrections. I also thank the three anonymous reviewers and Prof. Peter Foster.

Funding

This work was supported by the Royal Society of New Zealand through a Marsden [to S.G.].

Appendix

Let $V_{s|t,r_0,r_1,\nu}$ be the logarithm of the substitution rate at time s along a branch of length t. v_0 and v_t are the logarithm of rates at the start and the end of this branch. ν is the autocorrelation of rate parameter. For the sake of clarity of notations, $V_{s|t,r_0,r_1,\nu}$ will be denoted as V_s in what follows. We have

$$V_s = v_0 + \frac{(v_t - v_0)}{t} s + U_s,$$

and

$$U_s = W_s - \frac{s}{t} W_t,$$

where W_s defines the standard Brownian process. Therefore, W_s is normally distributed with mean 0 and variance νs. U_s defines a Brownian bridge, the value of which is 0 at $s=0$ and $s=t$. V_s is thus a Brownian bridge too with $V_0 = r_0$ and $V_t = v_t$. Moreover,

$$\begin{aligned}
\mathbf{E}(V_s) &= v_0 + \frac{(v_t - v_0)}{t} s + \mathbf{E}(U_s) \\
&= v_0 + \frac{(v_t - v_0)}{t} s + \mathbf{E}(W_s) - \frac{s}{t} \mathbf{E}(W_t) \\
&= v_0 + \frac{(v_t - v_0)}{t} s.
\end{aligned}$$

As for the variance, we have

$$\begin{aligned}
\mathbf{V}(V_s) &= \mathbf{V}(U_s) \\
&= \mathbf{V}\left(W_s - \frac{s}{t} W_t\right) \\
&= \mathbf{V}(W_s) + \frac{s^2}{t^2} \mathbf{V}(W_t) - 2\frac{s}{t} \mathrm{Cov}(W_s, W_t) \\
&= \nu s + \frac{s^2}{t^2} \nu t - 2\frac{s}{t} \left(\mathbf{E}(W_s W_t) - \mathbf{E}(W_s)\mathbf{E}(W_t)\right) \\
&= \nu s + \frac{s^2}{t^2} \nu t - 2\frac{s}{t} \mathbf{E}(W_s W_t) \\
&= \nu s + \frac{s^2}{t^2} \nu t - 2\frac{s}{t} \nu s \\
&= \frac{\nu s(t-s)}{t}.
\end{aligned}$$

Hence, V_s is a normally distributed random variable with mean $\mu_s = v_0 + (v_t - v_0)s/t$ and variance $\sigma_s^2 = \nu(t-s)s/t$.

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.

Aris-Brosou S., Yang Z. 2003. Bayesian models of episodic evolution support a late pre-Cambrian explosive diversification of the Metazoa. Mol. Biol. Evol. 20:1947–1954.

Benton M., Ayala F. 2003. Dating the tree of life. Science 300:1698–1700.

Britton T., Anderson C., Jacquet D., Lundqvist S., Bremer K. 2007. Estimating divergence times in large phylogenetic trees. Syst. Biol. 56:741–752.

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

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

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

Guindon S. 2010. Bayesian estimation of divergence times from large sequence alignments. Mol. Biol. Evol. 27:1768–1781.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Harvey P., Pagel M. 1991. The comparative method in evolutionary biology. Oxford: Oxford Univerity Press.

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.

Himmelmann L., Metzler D. 2009. TreeTime: an extensible C++ software package for Bayesian phylogeny reconstruction with time-calibration. Bioinformatics 25:2440–2441.

Huelsenbeck J., Ané C., Larget B., Ronquist F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst. Biol. 57:406–419.

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

Kendall D. 1948. On the generalized birth-and-death process. Ann. Math. Stat. 19:1–15.

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.

Lartillot N., Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28:729–744.

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

Lopez P., Casane D., Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

Nee S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Plummer M., Best N., Cowles C., Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Plummer M., Best N., Cowles K., Vines K. 2010. coda: output analysis and diagnostics for MCMC. R package Version 0.13-5.

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

R Development Core Team. 2011. R: a language and environment for statistical computing. Austria: R Foundation for Statistical Computing Vienna. ISBN 3-900051-07-0.

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

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

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. Botany 81:183.

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

Shankarappa R., Margolick J., Gange S., Rodrigo A., Upchurch D., Farzadegan H., Gupta P., Rinaldo C., Learn G., He X., Huang X.-L., Mullins J. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73:10489–10502.

Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.

Stadler T. 2010. Sampling-through-time in birth-death trees. J. Theor. Biol. 267:396–404.

Suchard M., Weiss R., Dorman K., Sinsheimer J. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. 51:715–728.

Suchard M., Weiss R., Dorman K., Sinsheimer J. 2003. Inferring spatial phylogenetic variation along nucleotide sequences. J. Am. Stat. Assoc. 98:427–437.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

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.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

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

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.

Yang Z., Wong W., Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22: 1107–1118.

Zuckerkandl E., Pauling L. 1962. Molecular disease, evolution, and genic heterogeneity. In: Kasha M. and Pullman B., editors. Horizons in Biochemistry. New York: Academic Press, p. 189–225.