Continuous and tractable models for the variation of evolutionary rates

Thomas Lepage [a], Stephan Lawi [b], Paul Tupper [a], David Bryant [a,c,*]

[a] *Department of Mathematics and Statistics, McGill University, Montréal, Canada*
[b] *Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII, Canada*
[c] *Department of Mathematics, Auckland University, Auckland, New Zealand*

Abstract

We propose a continuous model for variation in the evolutionary rate across sites and over the phylogenetic tree. We derive exact transition probabilities of substitutions under this model. Changes in rate are modelled using the CIR process, a diffusion widely used in financial applications. The model directly extends the standard gamma distributed rates across site model, with one additional parameter governing changes in rate down the tree. The parameters of the model can be estimated directly from two well-known statistics: the index of dispersion and the gamma shape parameter of the rates across sites model. The CIR model can be readily incorporated into probabilistic models for sequence evolution. We provide here an exact formula for the likelihood of a three-taxon tree. The likelihoods of larger trees can be evaluated using Monte–Carlo methods.
© 2005 Elsevier Inc. All rights reserved.

* Corresponding author. Address: McGill Centre for Bioinformatics, 3775 University, Montréal, Québec, Canada H3A 2B4. Tel.: +1 514 398 4578; fax: +1 514 398 3387.
E-mail address: bryant@mcb.mcgill.ca (D. Bryant).

1. Introduction

Understanding evolutionary rates and how they vary is one of the central concerns of molecular evolution. It has been clearly shown that inadequate models of rate variation, between lineages and between loci, can dramatically affect the accuracy of phylogenetic inference [1–3]. The dependency of molecular dating on evolutionary rate models is even more critical: we will only obtain precise divergence time estimates from molecular data once we can model the rate at which sequences evolve [4,5].

Modelling the evolutionary rate is made difficult by the number and variety of factors influencing it. The base rate of mutation can vary because of changes in the accuracy of replication machinery [6], DNA repair mechanisms [7], and metabolic rate [8]. At the cellular level, selective pressures can lead to variation of rate between loci and over time, as evidenced by differential rates of the three codon positions [9,10], the slower evolutionary rate of highly expressed genes [11], and the effect of tertiary structure on patterns of sequence conservation [12].

Selection also affects the evolutionary rate at the level of populations. For the most part, the only mutations that affect phylogenetics are those that are fixed in the population. Hence evolutionary rate is a combination of mutation rate and fixation rate. Fluctuations in population size, generation times, and environmental pressures affect fixation rates and thereby influence evolutionary rate [13–15].

Because of this complexity, the strategies employed for modelling evolutionary rate have tended to be statistical in nature. As with all statistical inference, there is an iterative sequence of model formulation, model assessment, and model improvement. The aim is to construct a model that accurately explains the observed variation but is as simple, and tractable, as possible.

Our goal in this paper is to derive a continuous model for rate evolution that avoids many of the problems of existing approaches. We base our model on the CIR process, a continuous Markov process that is widely used in finance to model interest rate fluctuations [16]. As we shall see, the model fits well into existing protocols for phylogenetic inference.

The CIR process has a stationary distribution given by a gamma distribution and yet, unlike the rates-across-sites (RAS) model of Uzzell and Corbin [17], the rate is allowed to vary along lineages. The CIR model adds only one parameter to the RAS model, and this parameter can be estimated directly from the index of dispersion or the autocorrelation (see below). Furthermore, we can derive exact transition probabilities when we incorporate CIR based rate variation into the standard models for sequence evolution.

The outline of the paper is as follows:

- In the following section we summarise the key characteristics of models for rate evolution, and show how existing models are classified with respect to these characteristics.
- In Section 3 we present the CIR model for rate evolution and discuss its basic properties.
- In Section 4 we derive transition probabilities for standard substitution models where the rate is described as a Markov process.
- In Section 5 we focus on the case where the rate is modelled by a CIR process.
- In Section 6 we extend this one step further to derive an expression for likelihood of a three-taxon tree using a substitution model with rate determined by the CIR process. We note that three-taxon trees are often used to study differences in evolutionary rate.

We conclude with a discussion on how these results can be applied to phylogenetic analyses for more than three taxa.

2. Properties of models for rate variation

In this section we examine several important characteristics that can be used to distinguish, and choose between, different models for rate variation. We discuss how the existing models fit into this scheme and summarise the differences between them in Table 1.

The rate of evolution for a given site at time $t \geqslant 0$ is denoted by R_t. For each $t > 0$, R_t is a non-negative random variable, and different models of rate evolution give different distributions for the rates R_t, $t \geqslant 0$.

Here and throughout the paper we will restrict out attention to *Markov processes*. That is, for any $t_1 \leqslant t_2 \leqslant t_3$, we assume that R_{t_3} conditioned on R_{t_2} is independent of R_{t_1}. In other words, the future depends on the past only through the present.

Table 1
Models for the substitution rate, classified according to the properties of Section 2

	I Rate class	II Ergodicity	III Closed form for the transition probability	IV Closed form for the autocovariance	Ref.
Models from population genetics					
Fluctuating mutation rates	CTMC with continuous state space	No	Yes	None	[14]
Fluctuating neutral space	CTMC with continuous state space	Yes	None	None	[14]
Compound Poisson process	CTMC with finite state space	Yes	None	None	[14,38]
Episodic evolution	CTMC with finite state space	Yes	None	None	[26]
Models from phylogenetics					
Covarion	CTMC with finite state space	Yes	Yes	None	[18,27,19,20]
HLS	CTMC with continuous state space	No	None	None	[21]
Log-normal	Diffusion	No	None	Constant autocovariance	[23,5]
Ornstein–Uhlenbeck	Diffusion	No	None	Exponentially decreasing	[4,22]
CIR process	Diffusion	Yes	Yes	Exponentially decreasing	[16]

CTMC stands for 'continuous-time Markov chain'.

2.1. Property I: Continuous or discontinuous sample paths

The first characteristic is whether sample paths of the process are continuous or discontinuous with respect to time. Typically, models with discontinuous paths have rates R_t that are constant except for discrete points in time at which there is a jump in the value (Fig. 1-1(a)). If the number of possible values for the rate is finite, then the rate can easily be described as a continuous-time Markov chain with a finite dimensional infinitesimal rate matrix. For example, in the covarion process [18] the basic rates are 'off' ($R_t = 0$) or 'on' ($R_t = 1$) and transitions occur between them at exponentially distributed random time intervals. Galtier [19,20] generalizes this process to one with more than two possible states. In other models, the range of possible values for the rate is continuous, as in the model of Huelsenbeck [21], where a rate change event consists of multiplying the previous rate by a gamma random variable. The rate change events are still discrete and exponentially distributed.

There are also models that describe the rate as a continuous function with time, and the most important class of Markov processes with continuous paths are *diffusions* (Fig. 1-2(a)). Examples include the CIR process presented here, the Ornstein–Uhlenbeck model of Aris-Brosou and Yang [4,22], and the log-normal model of Kishino et al. [5,23].

Finally, it is also possible for R_t to make jumps in value at a discrete set of times while also changing continuously in between these points.

2.2. Property II: Long term behaviour and ergodicity

The second property we consider is the distribution of R_t as t goes to infinity, that is, the distribution of the rate of evolution in the long term. Surprisingly, many models of rate evolution are very badly behaved in the limit.
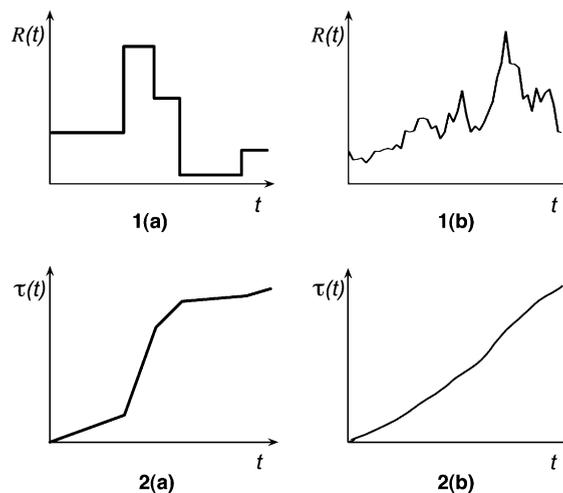
Fig. 1. A representation of the two classes of rate process with respect to the classification of property I. On top are examples of the rate history. Below are the corresponding integrated rates $\tau(t) = \int_{s=0}^{t} R_s \, ds$. The figures 1(a), 2(a) refer to a continuous-time Markov chain with discrete rate change events, and in figures 1(b), 2(b), $R(t)$ is modelled as a diffusion process, with continuous paths.

One problematic class of processes that have already been applied to rates in phylogenetics is the martingales. We say that a Markov process is a *martingale* if, for all $s, t \geqslant 0$ we have $E[M_{t+s}|M_t] = M_t$ [24]. An example of a Markov martingale is Brownian motion. As a result of this fairly innocuous looking condition, a martingale M_t has the property that either $E[|M_t|]$ is unbounded in time or M_t converges to a random constant [25]. Either possibility is undesirable from a modelling point of view. This may not be a problem if we only look at the process over a finite time, but neither is it particularly desirable. The processes of Kishino et al. [5,23] and Huelsenbeck et al. [21] all have the property that either R_t or $\log(R_t)$ is a martingale.

At a purely theoretical level, we observe that an ever-increasing variance will result for almost any signal that is *only* driven by its initial value and a stochastic force, with no directional bias. The position of a particle subjected to a random force produced by collisions with other particles is a classical example of such a case. In our context, the effects on the evolutionary rate are not independent of the actual rate: whatever the theoretical framework we consider, a high evolutionary rate is not as likely to increase (or to stay at high values) as to go back to smaller values. The theory of episodic evolution [26] fits particularly well with this idea. Periods of drastic adaptation with high evolutionary rates are naturally followed by periods where a population is adapted and its genome evolves much more slowly. Even according to the neutral theory, as argued by Takahata [14], the overall dynamic of the rate should behave as a random function that takes high values whenever bottlenecks occur and goes back to small values afterwards.

The concept of ergodicity naturally arises from this observation. We say that a Markov process is *ergodic* if for any initial rate R_0 the distribution of R_t converges to a unique distribution as t goes to infinity. The limiting distribution is known as the invariant or stationary distribution. Examples of ergodic processes include the OU process, the CIR model and (usually) the discrete space covarion and covarion-type models [27,19,20].

One possible way for a process to not be ergodic is if for some initial rate R_0 the distribution of R_t does not converge for large t. This must be the case if R_t is a martingale and does not converge to a constant, as is the case with Brownian motion. Another possibility is that R_t converges to different stationary distributions for different values of R_0.

2.3. Property III: Tractability

A highly desirable feature of any model is its tractability, both mathematical (does there exist a closed formula?) and computational (can we compute probabilities efficiently?). Nowadays, Monte Carlo methods make it possible to use arbitrarily complex models: however, explicit analytical formulae permit more efficient sampling [28].

There are several probability distribution functions that are important to have when working with rate processes. The most basic is the distribution of the rate R_t given the rate at time $t = 0$. This we have for the models [4,22,23,5] and for the CIR model, but not for the models of [21].

In phylogenetics we incorporate the model for evolutionary rate into the substitution model for sequence evolution at a site. These interact to give a joint process (R_t, X_t) for both the rate R_t at time t and the nucleotide or amino acid state X_t at time t. To evaluate the likelihood we require an expression for the joint conditional transition probabilities of X_t and R_t. If both random variables are discrete, then their joint transition probability is a probability mass function, which we denote

$Pr[X_t = j, R_t = s | X_0 = i, R_0 = r].$

If R_t is continuous, then its marginal distribution is best described by a probability density function $f_{R_t|R_0=r}(s)$, defined as

$Pr[R_t \in \boldsymbol{ds} | R_0 = r] = f_{R_t|R_0=r}(s)\mathrm{d}s,$

where $\mathrm{d}s$ is infinitesimal and $\boldsymbol{ds} = [s, s + \mathrm{d}s]$. The joint distribution of X_t and R_t will then take the general form

$$
\begin{aligned}
Pr[&X_t = j, R_t \in \boldsymbol{ds} | X_0 = i, R_0 = r] \\
&= Pr[X_t = j | X_0 = i, R_t \in \boldsymbol{ds}, R_0 = r]Pr[R_t \in \boldsymbol{ds} | R_0 = r] \\
&= Pr[X_t = j | X_0 = i, R_t = s, R_0 = r]f_{R_t|R_0=r}(s)\mathrm{d}s,
\end{aligned}
$$

where the last step is valid when R_t is a diffusion process. Even though it is sometimes possible to perform Monte Carlo computations to estimate this probability without a formula (as in [21]), having a formula will speed up the computations significantly without having to resort to approximations, as in [23,5].

2.4. Property IV: Autocovariance and dispersion

There is general agreement [29–31] on the relevance of autocorrelation in the modelling of evolutionary rate. Broadly speaking, if the various causes that explain rate variation (generation time, population size, environmental fitness) vary with time, it should be reflected in rate variations. The extent to which the rate varies can be studied using the *index of dispersion* [32–34]. Let $N(t)$ be the number of substitutions or mutations of a sequence over time t. The index of dispersion $I(t)$ is defined as

$$
I(t) = \frac{\text{Var}[N(t)]}{\text{E}[N(t)]}. \tag{1}
$$

This statistic can be estimated by comparing the number of substitutions that have accumulated in different lineages [33,35]. These estimates are consistently found to be greater than one, hinting that the substitution process is *overdispersed* and deviates from the Poisson process [29,36]. As Zheng showed [37], the infinitesimal matrices currently in use (with constant rate) are not likely by themselves to explain a large increase in the index of dispersion.

Many models of molecular evolution have been suggested to explain the observed amount overdispersion [14,38,39]. They model the impact of environmental change, fluctuation selective pressure and multiple simultaneous mutations. In phylogenetics, which generally operates at a longer time scale than population biology, the emphasis is on autocorrelation of rates [40,31]. In effect, the many population level processes influencing evolutionary rate are collapsed into one random process.

The first attempt of estimation of divergence times with an overdispersed clock was performed by Cutler [40], but instead of inferring the mean number of substitutions along each branch using his own method, he rather used Sanderson's non-parametric method [31]. Here, we propose a new model for the evolutionary rate, the CIR model, that can be applied to phylogenetic analysis, and at the same time incorporates overdispersion of the substitution process.

The index of dispersion resulting from a particular model of rate variation is a function of the *autocovariance* of that model. The autocovariance for a process R_t is defined by

$$\rho(t) = \text{Cov}(R_0, R_t). \tag{2}$$

For many processes we can derive an explicit formula for the autocovariance. If we assume that the substitutions occur according to a Poisson process with rate governed by our rate process (that is, the substitutions follow a *doubly stochastic* or *Cox* process, see Section 4) and the rate process has autocovariance function $\rho(t)$ then

$$I(t) = 1 + \frac{2 \int_0^t \left(1 - \frac{s}{t}\right) \rho(s)\, \mathrm{d}s}{\mathrm{E}(R_t)}, \tag{3}$$

as stated by a theorem in [41], and the stationary index of dispersion [26] is then

$$I(\infty) = \lim_{t \to \infty} I(t) = 1 + \frac{2 \int_0^\infty \rho(s)\, \mathrm{d}s}{\mu}, \tag{4}$$

provided that μ, the stationary mean of the process $R(t)$, and the limit, exist. Note that if there is *any* stochastic variation in rate then the index of dispersion will be greater than one [26].

Some rate models in phylogenetics [23,22] don't model the rate explicitly, but instead assign a (fixed) rate to each branch, so that the expected number of substitutions on a particular branch is equal to its length times its assigned (constant) rate.

A close look at the log-normal model from Thorne et al. [5], which differs from their previous version [23] in that the rate is explicitly modelled, shows that the rate has constant autocovariance, since this rate process is close to a transform of the Brownian motion, and Brownian motion has a constant autocovariance function. Put into Eq. (4), we see that the index of dispersion diverges. This problematic result illustrates the necessity of a balance between the presence of autocorrelation on one side, and the decrease of autocorrelation on a large time scale.

2.5. Property V: Heterotachy or homotachy

There are two general ways that models for evolutionary rate can be incorporated into phylogenetics. On one hand, we can introduce a distinct rate process for each site or locus. In this paper, our focus is principally on the modelling of *heterotachy*, which describes changes in the rate that are site-specific [2]. The transition probabilities that we derive in Section 4 can be applied primarily in a heterotachous context. Therefore, an independent substitution process as well as an independent rate process occur on each site.

The alternative to having an independent rate process for each site is to model dependencies between sites, or, in the extreme case, a single rate process that applies simultaneously to all sites. This extreme case can be modelled by trees for which the paths from the root to the leaves have different lengths, and capture what Langley and Fitch [34] called the *lineage effect*, i.e. the part of rate variation that is common to all sites, for example the consequence of a variable generation time. This kind of rate variation explains the extent to which the evolution of the sequences has violated the molecular clock.

3. A continuous diffusion model for the evolutionary rate

A Markov process with continuous paths and satisfying some additional smoothness conditions on its transition probabilities [42] is called a *diffusion*. There are many ways of specifying a diffusion process: perhaps the most intuitive one is by giving the probability distribution function (PDF) of R_t given $R_0 = r_0$, for arbitrary r_0. We denote this PDF by $f_R[R_t|r_0]$. For example, Brownian motion with parameter σ^2 is defined by the condition that $f_R[R_t|r_0]$ is a normal density with mean r_0 and variance $\sigma^2 t$.

A mathematically convenient representation of a diffusion is by means of a *stochastic differential equation* (SDE). In the same way that a dynamical system can be defined as the solution of a differential equation, a diffusion process R_t can be defined as the solution of an equation taking the general form (see [24, p. 61])

$$dR_t = \alpha(t, R_t)\, dt + \beta(t, R_t)\, dB_t. \tag{5}$$

Here, $\alpha(t, R_t)$ represents the deterministic effect on R_t, $\beta(t, R_t)$ the stochastic part, and dB_t is an infinitesimal 'random' increment. Brownian motion corresponds to the case when $\alpha(t, R_t) = 0$ for all t, $\beta(t, R_t)$ is constant and the SDE becomes

$$dR_t = \sigma\, dB(t).$$

Note that if $\beta(t, R_t) = 0$ for all t and R_t then (5) becomes a deterministic ordinary differential equation.

Going from an SDE such as (5) to a PDF for the diffusion involves solving a variable-coefficient second-order partial differential equation (PDE). For general functions α and β this PDE has no analytic solution. There are very few diffusions known that have closed form equations for their pdfs, and even fewer of these are ergodic. The simplest ergodic diffusions with closed-form expressions for the PDF are the Ornstein–Uhlenbeck and the CIR (Cox-Ingersoll-Ross) [16] processes.

The Ornstein–Uhlenbeck (OU) process is described by the SDE

$$dR_t = -bR_t\, dt + \sigma\, dB_t.$$

The PDF for R_t given $R_0 = r_0$ is the normal density with mean $r_0 e^{-bt}$ and variance $\sigma^2(1 - e^{-2\theta t})$. Its stationary distribution is normal with mean 0 and variance σ^2. The OU process was used by Aris-Brosou and Yang [22] to model evolutionary rates. However, the OU process can take on negative values, and it is not clear how it can be used directly without any transformation, such as a reflected OU or a squared OU. Aris-Brosou and Yang also proposed another model, the EXP (for exponential) model, defined as the following: the rate assigned to a branch is drawn from an exponential distribution with mean equal to the rate of its ancestral branch. Hence their EXP model was a martingale. They observed that the OU model seemed to provide a better fit to their data than the EXP model. Even though the reason of this better fit is still to be investigated, it seems reasonable to suggest that the ergodic property of the OU model could be a important factor. They also mentioned that the σ^2 parameter of the OU model was hard to infer, perhaps because the OU model has an insufficient number of free parameters.

The use of the CIR model solves the problem, since it is a generalization of the squared OU process, that has two independent parameters instead of three for the CIR. Fixing the mean parameter of the CIR process to one, we are left with two parameters that can be used

to independently estimate the variance and the autocorrelation. In the mathematical literature it is often called the *squared Bessel process* (see [43]).

The CIR process satisfies the SDE

$$dR_t = b(a - R_t) dt + \sigma\sqrt{R_t} dB_t, \tag{6}$$

and the PDF $f_R(R_t|r_0)$ for R_t given $R_0 = r_0$ is a non-central χ^2 distribution with degree of freedom $4ab/\sigma^2$ and parameter of non-centrality $\frac{4br_0 e^{-bt}}{\sigma^2(1-e^{-bt})}$. Its mean and variance are equal to

$$E[R_t] = r_0 e^{-bt} + a(1 - e^{-bt}) \tag{7}$$

$$Var[R_t] = r_0 \frac{\sigma^2}{b}(e^{-bt} - e^{-2bt}) + \frac{a\sigma^2}{2b}(1 - e^{-bt})^2. \tag{8}$$

As t goes to infinity, the parameter of non-centrality of f_R goes to zero, and the stationary distribution of R_t is a gamma distribution with shape parameter $2ab/\sigma^2$ and scale parameter $\sigma^2/2b$, or equivalently, the stationary rate equals $\frac{\sigma^2}{4b}$ times a χ^2 random variable with degree of freedom $\frac{4ab}{\sigma^2}$. Hence the mean of the stationary distribution is a and the variance is $\frac{a\sigma^2}{2b}$ [16].

Unlike an OU process, if r_0, a, and b are all positive a CIR process is always non-negative. The square of an OU process is a special case of the CIR process. Furthermore, by multiplying R_t by a constant in Eq. (6), we see that multiplying a CIR process by a positive constant gives another CIR process.

The covariance of the stationary CIR process can be exactly computed as

$$\rho(t) = \text{Cov}(R_0, R_t) = \frac{a\sigma^2}{2b} e^{-bt}. \tag{9}$$

From this, (3) leads to a closed formula for the index of dispersion:

$$I_{\text{CIR}}(t) = 1 + \frac{\sigma^2}{b^3 t}(bt - 1 + e^{-bt}).$$

Thus

$$I_{\text{CIR}}(\infty) = \lim_{t\to\infty} I_{\text{CIR}}(t) = 1 + \frac{\sigma^2}{b^2}. \tag{10}$$

From (6) we see that the CIR process possesses three parameters a, b, and σ^2. These parameters can be interpreted as the stationary mean a, the stationary variance $2a\sigma^2/b$, and the intensity of the force that drives the process to its stationary distribution, b. The parameter b determines how fast the process autocovariance goes to 0 as t increases.

The three parameters of the CIR process can be quickly estimated from standard statistics in molecular evolution. The parameter a is a scale parameter. It determines the expected rate at any time given no other information. Throughout the paper, we will assume that $a = 1$, so that the model has an expected rate equal to one. This parallels the constraint that the gamma distribution has an expected rate equal to one in the Rate-Across-Site (RAS) model [1].

The CIR process has a stationary distribution given by a gamma distribution. To make the stationary distribution coincide with the gamma distribution of a RAS model with parameter Γ we choose σ and b such that

$$\Gamma = \frac{\sigma^2}{b}. \tag{11}$$

The stationary index of dispersion, $I_{\text{CIR}}(t)$, can be estimated empirically [34,29] for a given locus. Because the difference between $I_{\text{CIR}}(t)$ and $I_{\text{CIR}}(\infty)$ is of the order of $1/t$, we can then use (10) and (11) to obtain the estimates

$$\hat{b} = \frac{\hat{\Gamma}}{\hat{I}_{\text{CIR}}(\infty) - 1},$$

$$\hat{\sigma}^2 = \frac{\hat{\Gamma}^2}{\hat{I}_{\text{CIR}}(\infty) - 1}.$$

4. Substitution models with a rate process

The standard model for the substitution process at a particular locus is a continuous-time Markov chain. This kind of process is defined by a square matrix Q called the *infinitesimal rate matrix*. Suppose, to begin, that there is a constant evolutionary rate r_0. As above, we let X_t denote the state (e.g. amino acid) at time t. The transition probabilities are then given by

$$\Pr[X_t = j | X_0 = i] = [\mathrm{e}^{Qr_0 t}]_{ij}. \tag{12}$$

We suppose that the process has a unique stationary distribution π, where π_j is the stationary probability of state j and

$$\pi_j = \lim_{t \to \infty} \Pr[X_t = j | X_0 = i]$$

for all i and j. We assume that Q has been normalised so that in the stationary distribution the expected number of substitutions over time t equals $r_0 t$. Note that the transition probabilities (12) depend only on the product $r_0 t$, so will be the same if we double the rate and halve the time, for example.

Suppose now that the rate is not constant, but instead varies according to some fixed function r_s, $s \geqslant 0$. Eq. (12) then becomes

$$\Pr[X_t = j | X_0 = i, r] = [\mathrm{e}^{Q\tau_r}]_{ij}, \tag{13}$$

where

$$\tau_r = \int_{s=0}^{s=t} r_s \, \mathrm{d}s$$

is the area under the curve r_s.

In the models we will consider, the fixed function $r = (r_t)_{t \geqslant 0}$ is replaced by a random process $R = (R_t)_{t \geqslant 0}$ that is dependent only on the starting rate r_0. The integral

$$\tau_R = \int_{s=0}^{s=t} R_s \, \mathrm{d}s \tag{14}$$

is also random in this case; let g_R denote its PDF. The transition probabilities can be determined from the expected value of (13) with τ_r replaced by the random variable τ_R. By the law of total expectation, this simplifies to

$$\Pr[X_t = j | X_0 = i] = \int_\tau [\mathrm{e}^{Q\tau}]_{ij} g_R(\tau) \, \mathrm{d}\tau. \tag{15}$$

Let $M(\eta) = \mathrm{E}_\tau[\mathrm{e}^{\eta\tau_R}]$ denote the *moment generating function* (MGF) for the random variable τ_R. Then (15) can be rewritten

$$\Pr[X_t = j | X_0 = i] = [M(Q)]_{ij}$$

where the function M is interpreted as a matrix function [44]. We assume that Q can be diagonalised as $Q = V\Lambda V^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix formed from the eigenvalues of Q. The matrix function $M(Q)$ can then be evaluated as $M(Q) = VM(\Lambda)V^{-1}$, where

$$M(\Lambda) = \mathrm{diag}(M(\lambda_1), \ldots, M(\lambda_n)).$$

See [45] for a more details on matrix functions. The problem of determining pattern probabilities therefore boils down to the problem of determining the moment generating function of the integrated rate, τ_R (Eq. (14)). Tuffley and Steel use this approach to derive distance estimates for the covarion process [27].

For applications in phylogenetics, we need the MGF of τ_R conditioned on just the starting rate, or both the starting and finishing rate. MGF of τ_R conditioned on a starting rate of r_0 is

$$M_{r_0}(\eta) = \mathrm{E}[\exp(\eta\tau_R) | R_0 = r_0] = \mathrm{E}\left[\exp\left(\eta \int_{s=0}^t R_s \, \mathrm{d}s\right) | R_0 = r_0\right]. \tag{16}$$

As before, we let $f_R(R_t | r_0)$ denote the PDF of R_t conditioned on $R_0 = r_0$. Let $\delta(x)$ denote the Dirac delta distribution, so that $\int \delta(x) f(x) \, \mathrm{d}x = f(0)$ for all smooth functions f. MGF of τ_R conditioned on both the starting and finishing rates is

$$\begin{aligned} M_{r_0, r_t}(\eta) &= \mathrm{E}[\exp(\eta\tau_R) | R_0 = r_0, R_t = r_t] \\ &= \frac{1}{f_R(r_t | r_0)} \mathrm{E}[\exp(\eta\tau_R) \delta(R_t - r_t) | R_0 = r_0] \\ &= \frac{1}{f_R(r_t | r_0)} \mathrm{E}\left[\exp\left(\eta \int_{s=0}^t R_s \, \mathrm{d}s\right) \delta(R_t - r_t) | R_0 = r_0\right]. \end{aligned} \tag{17}$$

Eqs. (16) and (17) hold irrespective of whether R is discrete or continuous, a diffusion, jump process, or a continuous time Markov chain.

We note in passing that analytic formulae for $M_{r_0}(\eta)$ and $M_{r_0, r_t}(\eta)$ exist in the case that R is a continuous time Markov chain, for example in the covarion-type model of Galtier [19]. Suppose that the evolutionary rate switches between rate values $g_1, g_2, \ldots, g_k$ following a continuous time Markov chain with infinitesimal rate matrix G. Let D be the $k \times k$ diagonal matrix with entries $g_1, g_2, \ldots, g_k$. A careful reworking of the proof of Theorem 1 in [46] gives MGF of τ_R conditioned on both the starting and finishing rate. MGF for τ_R conditioned on $r_0 = g_i$ is then

$$M_{g_i} = \sum_{j=1}^{k} (\mathrm{e}^{(G+\eta D)t})_{ij}$$

while MGF of τ_R conditioned on $r_0 = g_i$ and $r_t = g_j$ is

$$M_{g_i,g_j} = \frac{(\mathrm{e}^{(G+\eta D)t})_{ij}}{(\mathrm{e}^{Gt})_{ij}}.$$

This provides an independent derivation of the formula in [20] for transition probabilities under a covarion-type model.

5. Moment generating functions and transition probabilities for the CIR model

In this section we derive expressions for the (joint) transition probabilities

$$\Pr[X_t = j | X_0 = i, R_0 = r_0]. \tag{18}$$

and

$$\Pr[X_t = j, R_t \in \boldsymbol{ds} | X_0 = i, R_0 = r_0] \tag{19}$$

As we have seen, to evaluate these probabilities we need to determine the moment generating functions (MGFs) defined in Eqs. (16) and (17).

We use the Feynman-Kac formula [47,24] to derive analytic formulae for $M_{r_0}(\eta)$ and $M_{r_0,r_t}(\eta)$ under the CIR model. Let $g(\cdot)$ be a real-valued function. Define the function $v = v(t,x)$ by

$$v(t,x) = \mathrm{E}\left[\exp\left(\eta \int_0^t R(s)\,\mathrm{d}s\right) g(R_t) \middle| R_0 = x\right] \tag{20}$$

The Feynman-Kac formula [47] asserts that $v(t, x)$ solves the following partial differential equation (PDE)

$$\frac{\partial}{\partial t} v(t,x) = b(1-x)\frac{\partial}{\partial x}v(t,x) + \frac{1}{2}\sigma^2 x(t,x)\frac{\partial^2}{\partial x^2}v(t,x) + \eta x v \tag{21}$$

for $t > 0$, $x \in \mathbb{R}$, and with boundary condition

$$v(0,x) = g(x) \quad \text{for all } x \in \mathbb{R}. \tag{22}$$

We apply the methods in [48,49] to solve these PDEs with the different boundary conditions.

First consider the case when we condition only on the initial rate, Eq. (16). To make (20) equal to (16) we set $g(x) = 1$ for all x. The boundary condition (22) in this case becomes

$$v(0,x) = 1 \quad \text{for all } x \in \mathbb{R}.$$

With this boundary condition, the PDE (21) has solution

$$v(t,x) = \Psi(\eta,t)\mathrm{e}^{-x\Xi(\eta,t)},$$

where

$$\Psi(\eta, t) = \left(\frac{\overline{b} e^{bt/2}}{\overline{b} \cosh(\overline{b}t/2) + b \sin h(\overline{b}t/2)} \right)^{\frac{2b}{\sigma^2}}, \tag{23}$$

$$\Xi(\eta, t) = \left(\frac{2\eta \sinh(\overline{b}t/2)}{\overline{b} \cosh(\overline{b}t/2) + b \sinh(\overline{b}t/2)} \right), \tag{24}$$

$$\overline{b} = \sqrt{b^2 - 2\eta\sigma^2}. \tag{25}$$

We therefore have

$$M_{r_0}(\eta) = \Psi(\eta, t) e^{-r_0 \Xi(\eta, t)}. \tag{26}$$

The case when both the starting and finished rates are specified is more complicated. From (17) MGF M_{r_0, r_t} can be written

$$M_{r_0, r_t} = \frac{1}{f_R(r_t | r_0)} v(t, x),$$

where, in this case, $v(t, x)$ is given by (20) with $g(x) = \delta(x - r_t)$. The boundary condition (22) therefore becomes

$$v(0, x) = \delta(x - r_t).$$

With this new boundary condition, the PDE (21) has solution

$$v(t, x) = c \exp\left[-\frac{bt}{\sigma^2}(\overline{b} - b) + \frac{b - \overline{b}}{\sigma^2}x - \frac{b + \overline{b}}{\sigma^2}r_t - c(r_t + x)e^{-\overline{b}t} \right]$$

$$\times \left(\frac{r_t}{x e^{-\overline{b}t}} \right)^{\frac{b}{\sigma^2} - 1/2} I_{\frac{2b}{\sigma^2} - 1}\left(2c\sqrt{x r_t e^{-\overline{b}t}} \right), \tag{27}$$

where

$$c = \frac{2\overline{b}}{\sigma^2(1 - e^{-\overline{b}t})},$$

$$\overline{b} = \sqrt{b^2 - 2\eta\sigma^2}$$

and $I_\nu(x)$ is the modified Bessel function of the first kind with parameter v [50]. Hence MGF conditioned on initial and final rates is given by

$$M_{r_0, r_t}(\eta) = c \exp\left[-\frac{bt}{\sigma^2}(\overline{b} - b) + \frac{b - \overline{b}}{\sigma^2}r_0 - \frac{b + \overline{b}}{\sigma^2}r_t - c(r_t + r_0)e^{-\overline{b}t} \right]$$

$$\times \left(\frac{r_t}{r_0 e^{-\overline{b}t}} \right)^{\frac{b}{\sigma^2} - 1/2} I_{\frac{2b}{\sigma^2} - 1}\left(2c\sqrt{r_0 r_t e^{-\overline{b}t}} \right) \frac{1}{f_R(r_t | R_0 = r_0)},$$

where c and $\overline{b}$ are defined above and, from Section 3, $f_R(r_t | R_0 = r_0)$ is the PDF for a non-central χ^2 distribution with degree of freedom $4ab/\sigma^2$ and parameter of non-centrality $\frac{4br_0 e^{-bt}}{\sigma^2(1 - e^{-bt})}$.

Bringing everything together, we have our main result.

Theorem 1. *Let X_t be a substitution process with infinitesimal matrix Q, and diffusion process R_t with probability density function $f_{R_t}(s)$. Define P as the joint PDF of X_t and R_t, so that*

$$P[X_t = j, R_t \in \boldsymbol{ds}]$$
$$= \Pr[X_t = j, R_t \in \boldsymbol{ds} | X_0 = i, R_0 = r_0]$$
$$= \Pr[X_t = j | X_0 = i, R_t = s, R_0 = r_0] f_{R_t}(s)\, ds,$$

where ds is infinitesimal and $\boldsymbol{ds} = [\,s, s + ds\,]$. Suppose that Q is an infinitesimal matrix that can be written as $Q = V \Lambda V^{-1}$ where Λ is a diagonal matrix containing the eigenvalues $\lambda_1, \ldots, \lambda_n$ of Q. Then

$$P[X_t = j, R_t \in \boldsymbol{ds}] = M(Q)f_{R_t}(s)ds = VM(\Lambda)V^{-1}f_{R_t}(s)ds$$

where $M(\Lambda)$ is the diagonal matrix with, for all i,

$$M(\Lambda)_{ii} = M_{r_0,s}(\lambda_i),$$

and $M_{r_0,s}(\lambda_i)$ given by Eq. (17).

6. Three-taxon phylogenies

The simplest phylogeny for which we can distinguish between constant and variable evolutionary rates is a tree with three taxa. For this reason, there are many methods for testing, and estimating, rate variation that are based on three taxon analyses [29]. Here we show that the likelihood for a three-taxon tree, under the CIR model of rate variation, can be computed exactly. The problem for general phylogenies is more complex since we have to integrate out rates for the internal nodes. Here, we consider a heterotachous model, so that each site has its own rate history. Because the sites (and the rate at each site) evolve independently from each other, the likelihood of a sequence will be the product of all site-specific likelihoods. Therefore, we only require the likelihood computation for one site.

We recall that the stationary distribution of the CIR is a gamma distribution $\Gamma(v, \omega)$, where $\omega = v = 2b/\sigma^2$, i.e.

$$f_{R_0}(r) = \frac{\omega^v}{\Gamma(v)} r^{v-1} e^{-\omega r}. \tag{28}$$

Therefore the stationary mean and variance are *1* and $\sigma^2/2b$.

In order to get the transition probabilities, we will use MGF of τ_R unconditioned on the final rate, given by Eq. (26). The transition probability matrix of the substitution process, given initial rates, can be obtained by Eqs. (26) and (1). Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of Q. Using eigenvalue decomposition, we can find vectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}$ and $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ such that

$$\Pr[X_t = i | X_0 = j, R_0 = r_0] = \sum_{k=1}^{n} {\mathbf{u}_j^{(k)}}^T \mathbf{v}_i^{(k)} M_{r_0}(\lambda_k, t), \tag{29}$$

where we changed slightly our notation and explicitly wrote the dependency of M_{r_0} on *t*.
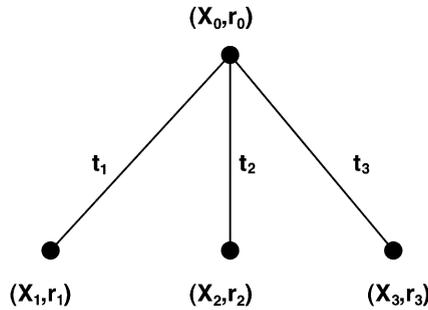
Fig. 2. A three-taxon rooted star tree, with branch lengths and one character state and rate value associated to each leaf.

Now consider the three-taxon tree with branches of lengths t_1, t_2, t_3 leading to leaves labelled with states x_1, x_2, x_3 (Fig. 2). If we condition on a rate r_0 and state x_0 at the root then the probability of observing x_1, x_2, x_3 at the leaves is given by

$$
\begin{aligned}
L(x_1, x_2, x_3 | x_0, r_0) &= P[X_{t_1} = x_1 | x_0, r_0] P[X_{t_2} = x_2 | x_0, r_0] P[X_{t_3} = x_3 | x_0, r_0] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} B_{ijk} M_{r_0}(\lambda_i, t_1) M_{r_0}(\lambda_j, t_2) M_{r_0}(\lambda_k, t_3)
\end{aligned}
\tag{30}
$$

where

$$
B_{ijk} = \mathbf{u}_{x_0}^{(i)} \mathbf{v}_{x_1}^{(i)} \mathbf{u}_{x_0}^{(j)} \mathbf{v}_{x_2}^{(j)} \mathbf{u}_{x_0}^{(k)} \mathbf{v}_{x_3}^{(k)}.
$$

The rate at the root is assumed to have the stationary distribution f_{R_0} given by (28). The likelihood integrated with respect to r_0 is then

$$
L(x_1, x_2, x_3 | x_0) = \int_{r_0} L(x_1, x_2, x_3 | x_0, r_0) f_{R_0}(r_0) \, dr_0
$$

which by (30) equals

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} B_{ijk} \int_{r_0} M_{r_0}(\lambda_i, t_1) M_{r_0}(\lambda_j, t_2) M_{r_0}(\lambda_k, t_3) f_{R_0}(r_0) \, dr_0.
\tag{31}
$$

We now use the formula (26) for MGFs derived above.

$$
\begin{aligned}
&M_{r_0}(\lambda_i, t_1) M_{r_0}(\lambda_j, t_2) M_{r_0}(\lambda_k, t_3) f_{R_0}(r_0) \\
&\quad = \Psi(\lambda_i, t_1) \mathrm{e}^{-r_0 \Xi(\lambda_i, t_1)} \Psi(\lambda_j, t_2) \mathrm{e}^{-r_0 \Xi(\lambda_j, t_2)} \Psi(\lambda_k, t_3) \mathrm{e}^{-r_0 \Xi(\lambda_k, t_3)} \frac{\omega^v}{\Gamma(v)} r_0^{v-1} \mathrm{e}^{-\omega r_0} \\
&\quad = \Psi(\lambda_i, t_1) \Psi(\lambda_j, t_2) \Psi(\lambda_k, t_3) \frac{\omega^v}{\Gamma(v)} r_0^{v-1} \mathrm{e}^{-r_0(\omega + \Xi(\lambda_i, t_1) + \Xi(\lambda_j, t_2) + \Xi(\lambda_k, t_3))}
\end{aligned}
\tag{32}
$$

Using integration by parts, or simply using the fact that the gamma PDF integrates to 1, we get

$$\int_{r_0} M_{r_0}(\lambda_i, t_1) M_{r_0}(\lambda_j, t_2) M_{r_0}(\lambda_k, t_3) f_{R_0}(r_0)\, \mathrm{d}r_0$$

$$= \Psi(\lambda_i, t_1)\Psi(\lambda_j, t_2)\Psi(\lambda_k, t_3)\left(\frac{\omega}{\omega + \Xi(\lambda_i, t_1) + \Xi(\lambda_j, t_3) + \Xi(\lambda_k, t_3)}\right)^{v}.$$

Finally, we can substitute this back into (31) to obtain

$$L(x_1, x_2, x_3 | x_0) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} B_{ijk}\Psi(\lambda_i, t_1)\Psi(\lambda_j, t_2)\Psi(\lambda_k, t_3)\left(\frac{\omega}{\omega + \Xi(\lambda_i, t_1) + \Xi(\lambda_j, t_3) + \Xi(\lambda_k, t_3)}\right)^{v}.$$

The formula extends immediately to phylogenies with n leaves attached to the root, though the number of terms in the summation increases exponentially. Our approach has been to use Monte–Carlo techniques to evaluate likelihoods on complete phylogenies.

7. Discussion

7.1. Summary

We have shown that, given a few natural criteria for our model selection, the CIR is the simplest continuous model that is at the same time ergodic, has a non-zero autocovariance function and that can account for an arbitrarily large index of dispersion. Moreover, we provided simple ways to estimate its parameters with the help of two observable statistics, namely the RAS gamma parameter and empirical index of dispersion. Another very interesting practical aspect of the CIR process is that it can be easily, and without approximations, implemented in the MCMC framework.

7.2. Future applications and extensions

The most straightforward application of our model would be to test the presence of heterotachy in specific data sets, meaning that we have to be able to apply our model to tree with more than 3 taxa. For such trees, we need to integrate efficiently over the possible rate histories. We have implemented (together with Nicolas Lartillot) an MCMC algorithm to perform Bayesian phylogenetic analysis under this model. The problems faced when the huge number of variables (one rate value for every node and every site) in the integration were a source of considerable computational difficulties, and convergence is rather slow. The implementation and application will be discussed in a forthcoming paper.

A possible future extension of our model could involve jump models, in which the rate path is discontinuous as in the continuous-time Markov chain, but also varies as diffusion between these discontinuities. However, the use of such a model implies the use of more parameters, and it may well be the case that the relative weakness of the rate of evolution signal cannot allow the use of more than two parameters, because of identifiability problems.

References

[1] Z. Yang, Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, Mol. Biol. Evol. 10 (1993) 1396.

[2] P. Lopez, D. Casane, H. Philippe, Heterotachy, an important process of protein evolution, Mol. Biol. Evol. 19 (2002) 1.

[3] E. Susko, Y. Inagaki, C. Field, M.E. Holder, A.J. Roger, Testing for differences in rates-across-sites distributions in phylogenetic subtrees, Mol. Biol. Evol. 19 (2002) 1514.

[4] S. Aris-Brosou, Z. Yang, Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny, Syst. Biol. 51 (2002) 703.

[5] H. Kishino, J.L. Thorne, W.J. Bruno, Performance of a divergence time estimation method under a probabilistic model of rate evolution, Mol. Biol. Evol. 18 (2001) 352.

[6] R.J. Britten, Rates of DNA sequence evolution differ between taxonomic groups, Science 231 (1986) 1393.

[7] K.H. Wolfe, P.M. Sharp, W.-H. Li, Mutation rates differ among regions of the mammalian genome, Nature 337 (1989) 283.

[8] A.P. Martin, S.R. Palumbi, Body size, metabolic rate, generation time, and the molecular clock, Proc. Natl. Sci. USA 90 (1993) 4087.

[9] N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences, Mol. Biol. Evol. 11 (1994) 725.

[10] S.V. Muse, B.S. Gaut, A likelihood method for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome, Mol. Biol. Evol. 11 (1994) 715.

[11] C. Pàl, B. Papp, L.D. Hurst, Highly expressed genes in yeast evolve slowly, Genetics 158 (1998) 927.

[12] D.M. Robinson, D.T. Jones, H. Kishino, N. Goldman, J.L. Thorne, Protein evolution with dependence among codons due to tertiary structure, Mol. Biol. Evol. 20 (1998) 1692.

[13] T. Ohta, H. Tachida, Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution, Genetics 126 (1990) 210.

[14] N. Takahata, On the overdispersed molecular clock, Genetics 116 (1987) 169.

[15] C.D. Laird, B.L. Mc Conaughy, B.J. Mc Carthy, Rate of fixation of nucleotide substitutions in evolution, Nature 224 (1969) 149.

[16] J.C. Cox, J.E. Ingersoll, S.A. Ross, A theory of the term structure of interest rates, Econometrica 53 (1985) 385.

[17] T. Uzzell, K.W. Corbin, Fitting discrete probability distributions to evolutionary events, Science 172 (1971) 1089.

[18] W.M. Fitch, E. Markowitz, An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution, Biochem. Genet. 4 (1970) 579.

[19] N. Galtier, Maximum-likelihood phylogenetic analysis under a covarion-like model, Mol. Biol. Evol. 18 (2001) 866.

[20] N. Galtier, Markov-modulated markov chains and the covarion process of molecular evolution, J. Comput. Biol. 11 (2004) 727.

[21] J.P. Huelsenbeck, B. Larget, D. Swofford, A compound Poisson process for relaxing the molecular clock, Genetics 154 (2000) 1879.

[22] S. Aris-Brosou, Z. Yang, Bayesian models of episodic evolution support a late precambrian explosive diversification of the metazoa, Mol. Biol. Evol. 20 (2003) 1947.

[23] J.L. Thorne, H. Kishino, I.S. Painter, Estimating the rate of evolution of the rate of molecular evolution, Mol. Biol. Evol. 15 (1998) 1647.

[24] B. Øksendal, Stochastic Differential Equations: an Introduction with Applications, Springer, 1998.

[25] D. Williams, Probabilities with Martingales, Cambridge Mathematical Textbooks, 1991.

[26] J.H. Gillespie, The Causes of Molecular Evolution, Oxford University, 1991.

[27] C. Tuffley, M.A. Steel, Modeling the covarion hypothesis of nucleotide substitution, Math. Biosci. 147 (1998) 63.

[28] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer, 2001.

[29] J.H. Gillespie, Lineage effects and the index of dispersion of molecular evolution, Mol. Biol. Evol. 6 (1989) 636.

[30] L. Chao, D.E. Carr, The molecular clock and the relationship between population size and generation time, Evolution 47 (1993) 688.

[31] M.J. Sanderson, A nonparametric approach to estimating divergence times in the absence of rate constancy, Mol. Biol. Evol. 14 (1997) 1218.

[32] T. Ohta, M. Kimura, On the constancy of the evolutionary rate in cistrons, J. Mol. Evol. 1 (1971) 18.

[33] M. Kimura, The Neutral Theory of Molecular Evolution, Cambridge University, Cambridge, 1983.

[34] C.H. Langley, W.M. Fitch, The constancy of evolution: a statistical analysis of the α and β haemoglobins, cytochrome c, and fibrinopeptide A, in: Genetic Structure of Populations, Univ. of Hawaii, Honolulu, 1973.

[35] M. Bulmer, Estimating the variablility of substitution rates, Genetics 123 (1989) 615.

[36] T. Ohta, Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory, J. Mol. Evol. 40 (1995) 56.

[37] Q. Zheng, On the dispersion index of a Markovian molecular clock, Math. Biosci. 172 (2001) 115.

[38] D.J. Cutler, Understanding the overdispersed molecular clock, Genetics 154 (2000) 1403.

[39] Y. Iwasa, Overdispersed molecular evolution in constant environments, J. Theoret. Biol. 164 (1993) 373.

[40] D.J. Cutler, Estimating divergence times in the presence of an overdispersed molecular clock, Mol. Biol. Evol. 17 (11) (2000) 1647.

[41] D.R. Cox, V. Isham, Point Processes, Chapman and Hall, NewYork, 1980.

[42] S. Karlin, H.M. Taylor, A Second Course in Stochastic Processes, Academic Press, New York, 1981.

[43] D. Revuz, M. Yor, Continuous Martingales and Brownian Motion, Springer, 2001.

[44] R. Horn, C. Johnson, Matrix Analysis, Cambridge University, 1985.

[45] G.H. Golub, C.F. Van Loan, Matrix Computations, John Hopkins University, 1996.

[46] J.N. Darroch, K.W. Morris, Passage-time generating functions for continuous-time finite Markov chains, J. Appl. Prob. 5 (1968) 414.

[47] M. Kac, On Some connections between probability theory and differential and integral equations, in: Proceedings of the Second Berkeley Symposium on Probability and Statistics, University of California, Berkeley, 1951.

[48] S.E. Shreve, Stochastic Calculus for Finance, Springer Finance, 2004.

[49] C. Albanese, S. Lawi, Laplace Transforms for Integrals of Markov Processes, Markov Proc. Rel. Fields 11 (2005) 667.

[50] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1965.