J. Math. Biol. (2008) 56:391–412
DOI 10.1007/s00285-007-0120-8

Mathematical Biology

Counting labeled transitions in continuous-time Markov models of evolution

Vladimir N. Minin · Marc A. Suchard

Received: 5 February 2007 / Published online: 14 September 2007

Abstract Counting processes that keep track of labeled changes to discrete evolutionary traits play critical roles in evolutionary hypothesis testing. If we assume that trait evolution can be described by a continuous-time Markov chain, then it suffices to study the process that counts labeled transitions of the chain. For a binary trait, we demonstrate that it is possible to obtain closed-form analytic solutions for the probability mass and probability generating functions of this evolutionary counting process. In the general, multi-state case we show how to compute moments of the counting process using an eigen decomposition of the infinitesimal generator, provided the latter is a diagonalizable matrix. We conclude with two examples that demonstrate the utility of our results.

V.N.M. was supported by a Dissertation Year Fellowship from the UCLA Graduate Division. M.A.S. is an Alfred P. Sloan Research Fellow.

V. N. Minin · M. A. Suchard
Department of Biomathematics, David Geffen School of Medicine at UCLA,
Los Angeles, CA, USA

Present Address:
V. N. Minin
Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA
e-mail: vminin@stat.washington.edu

M. A. Suchard
Department of Biostatistics, UCLA School of Public Health,
Los Angeles, CA, USA

M. A. Suchard (✉)
Department of Human Genetics, David Geffen School of Medicine at UCLA,
695 Charles E. Young Dr., South, Los Angeles, CA 90095-7088, USA
e-mail: msuchard@ucla.edu

Keywords Counting processes · Continuous-time Markov chains · Evolution · Phylogenetics

Mathematics Subject Classification (2000) 60J27 · 92D15 · 92D20

1 Introduction

Continuous-time Markov chains (CTMCs) have become standard modeling tools in evolutionary biology. Coupled with a phylogenetic tree that defines the evolutionary relationship among species, the Markov chain describes how a genetically inherited trait changes state over the tree. Such probabilistic models of evolution induce a likelihood of the trait values, observed at the tips of a tree. Estimation of model parameters then proceeds using either maximum likelihood or Bayesian frameworks [7]. Naturally, advances in evolutionary model parameter estimation go hand in hand with the development of new statistically rigorous hypothesis testing procedures. These developments often involve revisiting heuristics proposed by evolutionary biologists and reformulating the heuristic test statistics in terms of stochastic processes induced by Markov models of evolution [8,19,30]. One important, but insufficiently developed example is the counting process that keeps track of labeled changes in state experienced by a trait over the course of its evolutionary history.

Counting processes associated with CTMCs receive substantial attention in mathematical modeling of ion channel gating behavior [2,5]. One commonly views these processes as a particular case of a Markovian arrival process, a well studied object from queuing theory [16,17]. However, mathematical evolutionary biologists have not treated counting processes systematically. Analytic results are available only for limited evolutionary models or for specific counting problems [12,28,34]. Consequently, evolutionary biologists routinely estimate properties of evolutionary counting processes via computationally costly and often inaccurate simulation algorithms. Foremost among these algorithms is stochastic mapping [19]. In this paper, we show that it is possible to recover certain useful properties of evolutionary counting processes analytically. Although we capitalize on results derived in ion channel modeling and the engineering literature, our objectives require additional developments since we are interested in neither the stationary properties of the counting process nor in its long term behavior.

We start with a CTMC model of binary trait evolution. This simple model plays an important role in evolutionary developmental biology, where trait states often represent the presence or absence of an evolutionary interesting morphological feature [21,22]. In this two-state case, we obtain closed-form solutions for both the probability mass function and the probability generating function of the evolutionary counting process. We then proceed to examine models for traits with more than two discrete states; these models are ubiquitous in molecular sequence studies. In the general multistate case, closed-form expressions of the probability mass and probability generating functions are less practical. Therefore, we focus our attention on recovering factorial moments of the evolutionary counting processes. We show that when the generator

of the underlying CTMC is diagonalizable, the moments are recoverable analytically, provided the eigensystem of the CTMC generator is known, a common situation in evolutionary applications.

To demonstrate the utility of our results, we provide two examples. In the first example, we consider the evolution of a binary trait along a rooted phylogenetic tree with three tips (leaves). We show how conditioning on different observed data patterns at the tips of the tree effects the distribution of the total number of changes experienced by the trait. This distribution makes feasible a formal statistical test of "independent origins" hypotheses often confronted for morphological features [21]. Next, we turn to the analysis of DNA sequence evolution. We simulate an alignment of three sequences and compute the mean number of mutations for each site in the alignment. This simple site-specific summary of sequence variability conveniently detects spatial patterns of evolutionary rate variation without the introduction of complicated statistical models [29]. We illustrate this advantage by treating mean mutational counts as a time series and invoking spectral analysis. Additionally, we divide all DNA mutations into two labeled classes and perform evolutionary model diagnostics by comparing the prior and posterior expected number of mutations in each class.

The introduction of CTMC induced counting process theory to the field of mathematical evolutionary biology is an important contribution. From the theory, our analytic results provide for algorithms that computationally are more efficient than currently used simulation approaches. This is a very significant advance since in practice properties of the evolutionary counting process need to be evaluated an exceedingly large number of times to account for uncertainty in estimates of evolutionary model parameters or examine thousands of evolutionary traits at the same time [19].

2 Background and notation

Let $\{X_t,\ t \geq 0\}$ be an m-state homogeneous CTMC with infinitesimal generator $\mathbf{\Lambda} = \{\lambda_{ij}\}$, $i, j = 1, \ldots, m$ and finite-time transition probability matrix $\mathbf{P}(t) = e^{\mathbf{\Lambda} t}$ satisfying conditions $\mathbf{\Lambda 1} = \mathbf{0}$ and $\mathbf{P}(t)\mathbf{1} = \mathbf{1}$, where $\mathbf{0}$ and $\mathbf{1}$ are m-dimensional column vectors of zeros and ones respectively. We assume that $\{X_t\}$ is irreducible; this implies the existence of a unique stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ satisfying $\boldsymbol{\pi}\mathbf{\Lambda} = \mathbf{0}^T$, where ^{T} denotes transpose. Suppose now we wish to label a subset of all possible transitions of the chain $\{X_t\}$, where this subset has special biological importance. We specify transitions of interest through a set of ordered index pairs R that labels transitions from state i to state j only if $(i, j) \in R$. Since pairs of equal indices do not define CTMC transitions, R has at most $m(m-1)$ index pairs. The times of labeled transition occurrences define a point process on the positive real line. Figure 1 illustrates how a sample path of a three-state CTMC can generate, for example, realizations of two different point processes (black dots). In the first plot, all possible transitions are labeled. The second plot shows a process that keeps track of only transitions defined by set $R = \{(1, 3), (2, 1)\}$.

Let N_t be the total number of labeled transitions in time interval $(0, t]$. In the Examples section, we show that in order to characterize the evolutionary counting process

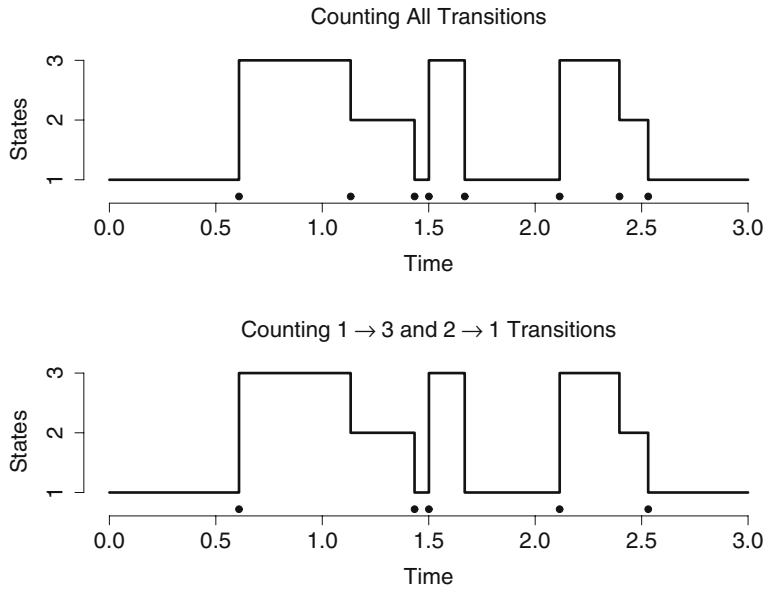

Fig. 1 Examples of continuous-time Markov chain (CTMC) induced counting processes. The *first plot* depicts a sample path of a three-state CTMC. The counting process is formed by the total number of all transitions of the Markov chain. The *second plot* shows the same three-state Markov chain path, but now only transitions from 1 to 3 and from 2 to 1 are labeled. The number of such transitions in time interval $[0, t)$ forms a different counting process

N_t on a phylogenetic tree we need means to compute joint probabilities

$$q_{ij}(n, t) = \Pr\left(N_t = n, X_t = j \mid X_0 = i\right), \tag{1}$$

and restricted factorial moments

$$m_{ij}^{[k]}(t) = \mathrm{E}\left(N_t^{[k]} 1_{\{X_t = j\}} \mid X_0 = i\right), \tag{2}$$

where $N^{[k]} = N(N-1)\cdots(N-k+1)$. Quantities in Eqs. (1) and (2) implicitly depend on the labeling set R that together with the infinitesimal generator matrix $\mathbf{\Lambda}$ define the evolutionary counting process $\{N_t,\ t \geq 0\}$.

3 Two-state CTMC induced counting processes

Consider a two-state Markov chain $\{X_t\}$ with generator

$$\mathbf{\Lambda} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2, \end{pmatrix} \tag{3}$$

where $\lambda_1, \lambda_2 > 0$, and a counting process $\{N_t\}$ that labels all transitions. Specifically, $R = \{(1, 2), (2, 1)\}$. Without loss of generality, throughout this section we assume that $X_0 = 1$. Then, since X_t cycles between only two states, the state of the chain at

time t is fully determined by number of transitions N_t. In particular, if N_t is an odd integer, then X_t must be 2. Similarly if N_t is even, X_t must occupy state 1. Therefore, to compute the joint probabilities

$$\begin{aligned} q_{11}(n,t) &= \Pr(N(t)=n|X_0=1)1_{\{n \text{ is even}\}} \quad \text{and} \\ q_{12}(n,t) &= \Pr(N(t)=n|X_0=1)1_{\{n \text{ is odd}\}}, \end{aligned} \tag{4}$$

where $1_{\{\cdot\}}$ is the indicator function, it suffices to determine marginal probabilities

$$q_n(t) = \Pr(N_t = n|X_0 = 1). \tag{5}$$

For a two-state system, CTMC finite-time transition probabilities satisfy $p_{ij}(\Delta t) = \lambda_i \Delta t + o(\Delta t)$ when $i \neq j$ and $p_{ii}(\Delta t) = 1 - \lambda_i \Delta t + o(\Delta t)$ otherwise. Consequently, the following recursive equations hold for the marginal probabilities in Eq. (5),

$$\begin{aligned} q_{2k-1}(t+\Delta t) &= q_{2k-1}(t)(1-\lambda_2\Delta t) + q_{2k-2}(t)\lambda_1\Delta t + o(\Delta t), \quad \text{and} \\ q_{2k}(t+\Delta t) &= q_{2k}(t)(1-\lambda_1\Delta t) + q_{2k-1}(t)\lambda_2\Delta t + o(\Delta t), \end{aligned} \tag{6}$$

where $k = 1, 2, \ldots$ and $q_0(t) = e^{-\lambda_1 t}$. Dividing both sides of Eq. (6) by Δt and sending it to 0 yields recursive differential equations

$$\begin{aligned} \frac{d}{dt}q_{2k-1}(t) &= -\lambda_2 q_{2k-1}(t) + \lambda_1 q_{2k-2}(t) \quad \text{and} \\ \frac{d}{dt}q_{2k}(t) &= -\lambda_1 q_{2k}(t) + \lambda_2 q_{2k-1}(t), \end{aligned} \tag{7}$$

subject to initial conditions $q_n(0) = 0$ for $n > 0$.

Let $f_n(s) = \mathcal{L}[q_n(t)](s)$, where $\mathcal{L}$ is the Laplace transform. Applying $\mathcal{L}$ to both sides of Eq. (7) produces the following algebraic equations

$$f_{2k-1}(s) = \frac{\lambda_1}{s+\lambda_2} f_{2k-2}(s) \quad \text{and} \quad f_{2k}(t) = \frac{\lambda_2}{s+\lambda_1} f_{2k-1}(s). \tag{8}$$

Keeping in mind that $f_0(s) = \mathcal{L}\left[e^{-\lambda_1 t}\right](s) = \frac{1}{s+\lambda_1}$, we deduce that

$$f_{2k-1}(s) = \frac{\lambda_1^k \lambda_2^{k-1}}{(s+\lambda_1)^k (s+\lambda_2)^k} \quad \text{and} \quad f_{2k}(s) = \frac{\lambda_1^k \lambda_2^k}{(s+\lambda_1)^{k+1}(s+\lambda_2)^k}. \tag{9}$$

If $\lambda_1 = \lambda_2$, then $f_n(s) = \frac{\lambda_1^n}{(s+\lambda_1)^{n+1}}$ and $q_n(t) = \mathcal{L}^{-1}[f_n(s)](t) = e^{-\lambda_1}\frac{\lambda_1^n t^n}{n!}$. Therefore, when the rates of leaving states 1 and 2 are equal, the number of Markov chain

transitions N_t is Poisson distributed as intuition predicts. In the case of unequal rates, $\lambda_1 \neq \lambda_2$, we expand Eq. (9) into partial fractions,

$$\begin{aligned} f_{2k-1}(s) &= \sum_{i=1}^{k} \frac{A_{k-i}^k}{(s+\lambda_1)^i} + \sum_{i=1}^{k} \frac{B_{k-i}^k}{(s+\lambda_2)^i} \quad \text{and} \\ f_{2k}(s) &= \sum_{i=1}^{k+1} \frac{C_{k-i+1}^k}{(s+\lambda_1)^i} + \sum_{i=1}^{k} \frac{D_{k-i}^k}{(s+\lambda_2)^i} \end{aligned} \tag{10}$$

and apply the method of derivatives to determine the partial fraction coefficients

$$\begin{aligned} A_i^k &= \frac{1}{i!}\frac{d^i}{ds^i}\left[(s+\lambda_1)^k f_{2k-1}(s)\right]_{s=-\lambda_1}, \\ C_i^k &= \frac{1}{i!}\frac{d^i}{ds^i}\left[(s+\lambda_1)^{k+1} f_{2k}(s)\right]_{s=-\lambda_1}, \\ B_i^k &= \frac{1}{i!}\frac{d^i}{ds^i}\left[(s+\lambda_2)^k f_{2k-1}(s)\right]_{s=-\lambda_2} \quad \text{and} \\ D_i^k &= \frac{1}{i!}\frac{d^i}{ds^i}\left[(s+\lambda_2)^k f_{2k}(s)\right]_{s=-\lambda_2}. \end{aligned} \tag{11}$$

Inverse Laplace transformation of Eq. (10) and differentiation in (11) leads to a closed-form solution to the differential Eq. (7)

$$\begin{aligned} q_{2k-1}(t) &= \sum_{i=1}^{k} A_{k-i}^k \frac{t^{i-1}}{(i-1)!} e^{-\lambda_1 t} + \sum_{i=1}^{k} B_{k-i}^k \frac{t^{i-1}}{(i-1)!} e^{-\lambda_2 t} \quad \text{and} \\ q_{2k}(t) &= \sum_{i=1}^{k+1} C_{k-i+1}^k \frac{t^{i-1}}{(i-1)!} e^{-\lambda_1 t} + \sum_{i=1}^{k} D_{k-i}^k \frac{t^{i-1}}{(i-1)!} e^{-\lambda_2 t}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} A_i^k &= \binom{k+i-1}{i} \frac{(-1)^i \lambda_1^k \lambda_2^{k-1}}{(\lambda_2-\lambda_1)^{k+i}}, \quad C_i^k = \binom{k+i-1}{i} \frac{(-1)^i \lambda_1^k \lambda_2^k}{(\lambda_2-\lambda_1)^{k+i}}, \\ B_i^k &= \binom{k+i-1}{i} \frac{(-1)^i \lambda_1^k \lambda_2^{k-1}}{(\lambda_1-\lambda_2)^{k+i}}, \quad D_i^k = \binom{k+i}{i} \frac{(-1)^i \lambda_1^k \lambda_2^k}{(\lambda_1-\lambda_2)^{k+i+1}}. \end{aligned} \tag{13}$$

To further explore the properties of the evolutionary counting process for two-state Markov chains, we proceed with deriving the probability generating function of N_t,

$$g(r,t) = \sum_{n=0}^{\infty} r^n q_n(t), \; |r| \leq 1. \tag{14}$$

Starting from Eq. (7), we construct a forward differential equation for the generating function (14),

$$\frac{\partial}{\partial t} g(r,t) = (r-1)\left[\lambda_1 \sum_{k=0}^{\infty} r^{2k} q_{2k}(t) + \lambda_2 \sum_{k=0}^{\infty} r^{2k+1} q_{2k+1}(t)\right]. \tag{15}$$

Identifying $\sum_{k=0}^{\infty} r^{2k} q_{2k}(t) = \frac{1}{2}[g(r,t) + g(-r,t)]$ and $\sum_{k=0}^{\infty} r^{2k+1} q_{2k+1}(t) = \frac{1}{2}[g(r,t) - g(-r,t)]$, we derive a forward equation for the mirror image of the generating function, $g(-r,t)$, and combine it with (15) to arrive at a simple linear system of ordinary differential equations

$$\begin{aligned} \frac{\partial}{\partial t} g(r,t) &= \frac{1}{2}(r-1)[(\lambda_1+\lambda_2)g(r,t) - (\lambda_2-\lambda_1)g(-r,t)] \quad \text{and} \\ \frac{\partial}{\partial t} g(-r,t) &= \frac{1}{2}(r+1)[(\lambda_2-\lambda_1)g(r,t) - (\lambda_1+\lambda_2)g(-r,t)]. \end{aligned} \tag{16}$$

The solution to this system with initial conditions $g(r,0) = g(-r,0) = 1$ yields the generating function of N_t given X_t starts in state 1,

$$g(r,t) = \frac{(r-1)\lambda_1 - \alpha_2}{\alpha_1 - \alpha_2} e^{\alpha_1 t} + \frac{\alpha_1 - (r-1)\lambda_1}{\alpha_1 - \alpha_2} e^{\alpha_2 t}, \tag{17}$$

where $\alpha_{1,2} = \frac{1}{2}\left[-\lambda_1 - \lambda_2 \pm \sqrt{(\lambda_1+\lambda_2)^2 + 4(r^2-1)\lambda_1\lambda_2}\right]$. This analytic formula for $g(r,t)$ allows for simple derivations of factorial moments

$$\mathrm{E}\left(N^{[k]}(t) \mid X_0 = 1\right) = \frac{\partial^k}{\partial r^k} g(r,t)|_{r=1}, \tag{18}$$

bypassing the repetitive one-step calculations relied on in [34]. Since our objective is to develop algorithms for computing factorial moments over phylogenetic trees, we are not only interested in moments (18), but also in restricted factorial moments. We once again exploit the even-odd symmetry of the two-state model and express the restricted factorial moments as

$$\begin{aligned} m_{11}^{[k]}(t) &= \frac{1}{2}\frac{\partial^k}{\partial r^k}\left[g(r,t) + g(-r,t)\right]_{r=1} \quad \text{and} \\ m_{12}^{[k]}(t) &= \frac{1}{2}\frac{\partial^k}{\partial r^k}\left[g(r,t) - g(-r,t)\right]_{r=1}. \end{aligned} \tag{19}$$

Clearly, if $X_0 = 2$ we can simply exchange λ_1 and λ_2 in the above derivations to arrive at the formulas for $q_{21}(n,t)$, $q_{22}(n,t)$, $m_{21}^{[k]}(t)$, and $m_{22}^{[k]}(t)$.

Forward transitions

Since in independent origins questions it is necessary to count only directed transitions that represent morphological innovations over evolutionary history, we derive the joint probabilities and restricted factorial moments for directed transitions in the two-state model. Let $X_0 = 1$ and N_t^f be the number of forward transitions from state 1 to state 2, i.e. $R = \{(1, 2)\}$, during the time interval $(0, t]$. The joint probabilities for N_t^f can be recovered from the probability mass function of N_t,

$$\begin{aligned} &\Pr\left(N_t^f = k, X_t = 1 \mid X_0 = 1\right) = q_{2k}(t) \quad \text{and} \\ &\Pr\left(N_t^f = k, X_t = 2 \mid X_0 = 1\right) = \begin{cases} q_{2k-1}(t) & k > 0, \\ 0 & k = 0. \end{cases} \end{aligned} \tag{20}$$

We can also express the restricted mean number of forward transitions as

$$\begin{aligned} \mathrm{E}\left(N_t^f 1_{\{X_t=1\}} \mid X_0 = 1\right) &= \frac{1}{2}\sum_{k=1}^{\infty} 2k\, q_{2k}(t) = \frac{1}{2} m_{11}^{[1]}(t) \text{ and} \\ \mathrm{E}\left(N_t^f 1_{\{X_t=2\}} \mid X_0 = 1\right) &= \sum_{k=1}^{\infty} k\, q_{2k-1}(t) \\ &= \frac{1}{2}\left[\sum_{k=1}^{\infty}(2k-1) q_{2k-1}(t) + \sum_{k=1}^{\infty} q_{2k-1}(t)\right] \\ &= \frac{1}{2}\left[m_{12}^{[1]}(t) + p_{12}(t)\right]. \end{aligned} \tag{21}$$

We obtain higher moments and formulas for backward transitions in a similar fashion.

4 Multi-state CTMC induced counting processes

General theory

Let us now turn to the general case, where $\{X_t\}$ can attain m arbitrary states, and N_t counts transitions between pairs in a predefined set R. In this first subsection, we review previous results from theoretical developments of CTMC induced counting processes. Since these developments are found elsewhere, we omit most of the proofs and provide only missing relevant details [3,5,18]. We then proceed with new developments and demonstrate that reversibility, enjoyed by the majority of evolutionary Markov models, permits one to compute the restricted factorial moments of N_t using an eigen decomposition of the CTMC infinitesimal generator.

As in the two-state case, we start with a forward equation for the joint probabilities $q_{ij}(n, t)$. During infinitesimal time period Δt, the number of labeled transitions N_t does not change if an unlabeled transition into state j is made. Alternatively, N_t

increases by one if a labeled transition into state j occurs. This intuitive reasoning translates into the following relationship

$$q_{ij}(n, t + \Delta t) = \sum_{k:(k,j)\notin R} q_{ik}(n, t)\lambda_{kj} + \sum_{k:(i,j)\in R} q_{ik}(n-1, t)\lambda_{kj} + o(\Delta t), \quad (22)$$

where $i, j = 1, \ldots, m$, $n \geq 1$, and $q_{ij}(n, t)$ is the probability of X_t changing its state from i to j in time t with n labeled transitions. Dividing Eq. (22) by Δt, sending it to 0, and introducing matrices $\mathbf{Q}(n, t) = \{q_{ij}(n, t)\}$, $\mathbf{\Lambda}_R = \{\lambda_{ij} \times 1_{\{(i,j)\in R\}}\}$, and $\mathbf{\Lambda}_{\overline{R}} = \{\lambda_{ij} \times 1_{\{(i,j)\notin R\}}\}$, we arrive at the matrix differential equation

$$\frac{d}{dt}\mathbf{Q}(n, t) = \mathbf{Q}(n, t)\mathbf{\Lambda}_{\overline{R}} + \mathbf{Q}(n-1, t)\mathbf{\Lambda}_R. \quad (23)$$

Starting with $\mathbf{Q}(0, t) = e^{\mathbf{\Lambda}_{\overline{R}}t}$ and the Laplace transform of Eq. (23), we obtain that for $n \geq 1$,

$$\mathbf{F}(n, s) = \mathbf{F}(n-1, s)\mathbf{\Lambda}_R \left(s\boldsymbol{I} - \mathbf{\Lambda}_{\overline{R}}\right)^{-1}, \quad (24)$$

where $\mathbf{F}(n, s) = \mathcal{L}\left[\mathbf{Q}(n, t)\right](s)$, defined on its region of convergence $\mathrm{Re}(s) > 0$, and $\mathbf{I}$ is an $m \times m$ identity matrix. Since $\mathcal{L}[\mathbf{Q}(0, t)] = \mathcal{L}[e^{\mathbf{\Lambda}_{\overline{R}}t}] = (s\mathbf{I} - \mathbf{\Lambda}_{\overline{R}})^{-1}$, the solution of the recursive Eq. (24) is

$$\mathbf{F}(n, s) = \left(s\boldsymbol{I} - \mathbf{\Lambda}_{\overline{R}}\right)^{-1} \left[\mathbf{\Lambda}_R \left(s\boldsymbol{I} - \mathbf{\Lambda}_{\overline{R}}\right)^{-1}\right]^n. \quad (25)$$

Recalling that $\mathrm{Re}(s) > 0$ and $\lambda_{ii} = -\sum_{j\neq i} \lambda_{ij}$ implies that $|s - \lambda_{ii}| \geq |\mathrm{Re}(s) + \sum_{j\neq i} \lambda_{ij}| \geq \sum_{j\neq i} \lambda_{ij} \geq \sum_{j\neq i} \lambda_{ij} \times 1_{\{(i,j)\notin R\}}$ for $i = 1, \ldots, m$. Therefore, matrix $s\mathbf{I} - \mathbf{\Lambda}_{\overline{R}}$ is strictly diagonally dominant and invertible. When $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_R$ commute, Eq. (25) simplifies to $\mathbf{F}(n, s) = \mathbf{\Lambda}_R^n \left(s\boldsymbol{I} - \mathbf{\Lambda}_{\overline{R}}\right)^{-n-1}$, making it possible to obtain the inverse Laplace transform of $\mathbf{F}(s, n)$ analytically, $\mathbf{Q}(n, t) = \frac{(\mathbf{\Lambda}_R t)^n}{n!} e^{\mathbf{\Lambda}_{\overline{R}}t}$. If $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_R$ do not commute, a closed form solution for $\mathbf{Q}(n, t)$ does not seem feasible. However, in many evolutionary applications, probabilities $\mathbf{Q}(n, t)$ are needed only for small n. In such cases, formulas (23) and (25) remain useful and practical. For example, if we decide to label all possible transitions, then

$$q_{ij}(0, t) = \begin{cases} e^{\lambda_{ii}t} & i = j, \\ 0 & i \neq j \end{cases} \quad (26)$$

and the inverse Laplace transform of $\mathbf{F}(1, s)$ yields

$$q_{ij}(1, t) = \begin{cases} \frac{\lambda_{ij}}{\lambda_{ii}-\lambda_{jj}} \left(e^{\lambda_{ii}t} - e^{\lambda_{jj}t}\right) & i \neq j \text{ and } \lambda_{ii} \neq \lambda_{jj}, \\ \lambda_{ij} t e^{\lambda_{jj}t} & i \neq j \text{ and } \lambda_{ii} = \lambda_{jj}, \\ 0, & i = j. \end{cases} \quad (27)$$

Alternatively, Siepel et al. [28] propose to use an embedded discrete-time Markov chain to arrive at a recursive algorithm for computing $\mathbf{Q}(n,t)$ when $R = \{(i,j) : i,j = 1,\ldots,m, i \neq j\}$.

Similarly to the two-state case, we proceed with the matrix probability generating function of N_t,

$$\mathbf{G}(r,t) = \sum_{n=0}^{\infty} r^n \mathbf{Q}(n,t). \tag{28}$$

Using matrix differential Eq. (23), we arrive at the forward differential equation for $\mathbf{G}(r,t)$,

$$\frac{\partial}{\partial t}\mathbf{G}(r,t) = \mathbf{G}(r,t)\left(\mathbf{\Lambda}_{\overline{R}} + r\mathbf{\Lambda}_R\right), \tag{29}$$

subject to initial condition $\mathbf{G}(r,0) = \mathbf{I}$. Since the righthand side of this equation depends on t only through $\mathbf{G}(r,t)$, we immediately recognize that the solution of this equation is the matrix exponential

$$\mathbf{G}(r,t) = e^{(\mathbf{\Lambda}_{\overline{R}} + r\mathbf{\Lambda}_R)t}. \tag{30}$$

We would like to highlight that in contrast with engineering and ion channel applications, the relatively small size of the state space in evolutionary Markov models does not prohibit numerical estimation of the matrix exponential (30). However, it is still desirable to avoid this computationally intensive calculations whenever possible.

Factorial moments

We now turn to the problem of calculating restricted factorial moments $\mathbf{M}^{[k]}(t)$. In principle, the restricted factorial moments can be recovered by differentiating the matrix probability generating function,

$$\mathbf{M}^{[k]}(t) = \frac{\partial^k}{\partial r^k}\mathbf{G}(r,t)\Big|_{r=1}. \tag{31}$$

However, as with the inverse Laplace transform of (25), such differentiation is only possible analytically when matrices $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_R$ commute. In this situation $\mathbf{M}^{[k]}(t) = (\mathbf{\Lambda}_R t)^k e^{\mathbf{\Lambda} t}$. Ball and Mine justly point out that formulas (30) and (31) are not very practical for numerical calculation of the restricted factorial moments as matrices $\mathbf{\Lambda}_R$ and $\mathbf{\Lambda}$ usually do not commute [3]. Therefore, we follow the authors' suggestion and use an integral representation of the restricted factorial moments. We start by differentiating Eq. (29) k times with respect to r and evaluating both sides of the equation at $r = 1$. This produces the following recursive differential equations for the restricted factorial moments,

$$\frac{\partial}{\partial t}\mathbf{M}^{[k]}(t) = \mathbf{M}^{[k]}(t)\mathbf{\Lambda} + k\mathbf{M}^{[k-1]}(t)\mathbf{\Lambda}_R, \tag{32}$$

where $\mathbf{M}^{[0]}(t) = \mathbf{P}(t) = e^{\mathbf{\Lambda} t}$ is the finite-time transition probability matrix. Multiplying both sides of Eq. (32) by integrating factor $e^{-\mathbf{\Lambda} t}$ and integrating with respect to t, we obtain a recursive integral formula for the restricted factorial moments,

$$\mathbf{M}^{[k]}(t) = k \int_0^t \mathbf{M}^{[k-1]}(\theta) \mathbf{\Lambda}_R e^{\mathbf{\Lambda}(t-\theta)} d\theta. \tag{33}$$

To make further progress for evolutionary models, we assume that $\mathbf{\Lambda}$ is diagonalizable with eigen decomposition $\mathbf{\Lambda} = \mathbf{U}\mathbf{H}\mathbf{U}^{-1}$, where $\mathbf{H}$ is a diagonal matrix with the real eigenvalues $h_1, \ldots, h_m$ of $\mathbf{\Lambda}$ along its diagonal and eigenvectors of $\mathbf{\Lambda}$ forming the columns of $\mathbf{U}$. Then, transition probability matrix $\mathbf{P}(t) = e^{\mathbf{\Lambda} t} = \mathbf{U} e^{\mathbf{H} t} \mathbf{U}^{-1}$, where $e^{\mathbf{H} t}$ is a diagonal matrix composed of elements $e^{h_1 t}, \ldots, e^{h_m t}$. With this eigen decomposition of the transition probability matrix, we arrive at its spectral representation

$$e^{\mathbf{\Lambda} t} = \sum_{i=1}^{m} \mathbf{B}_i e^{h_i t}, \tag{34}$$

where $\mathbf{B}_i = \mathbf{U}\mathbf{E}_i\mathbf{U}^{-1}$ and $\mathbf{E}_i$ is a matrix with zero entries everywhere, except at the ii-th entry, which is one. Almost all evolutionary models derive reversible CTMCs. Reversibility implies that the infinitesimal generator is similar to a symmetric matrix and hence is diagonalizable with real eigenvalues [9]. Therefore our seemingly strong diagonalizability assumption does in fact hold for the majority of evolutionary applications.

From Eq. (33) we obtain

$$\mathbf{M}^{[1]}(t) = \int_0^t e^{\mathbf{\Lambda}\theta} \mathbf{\Lambda}_R e^{\mathbf{\Lambda}(t-\theta)} d\theta \tag{35}$$

for the first moment. Then using spectral representation (34) we arrive at the following convenient to evaluate expression

$$\mathbf{M}^{[1]}(t) = \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{B}_i \mathbf{\Lambda}_R \mathbf{B}_j I_{ij}(t), \tag{36}$$

where

$$I_{ij}(t) = \begin{cases} t e^{h_i t} & \text{if } h_i = h_j, \\ \dfrac{e^{h_i t} - e^{h_j t}}{h_i - h_j} & \text{if } h_i \neq h_j. \end{cases} \tag{37}$$

Continuing the application of spectral decomposition and integration yields the following expression for restricted factorial moments

$$\mathbf{M}^{[k]}(t) = k! \sum_{i_1=1}^{m} \cdots \sum_{i_{k+1}=1}^{m} \mathbf{B}_{i_1} \left(\prod_{l=2}^{k+1} \mathbf{\Lambda}_R \mathbf{B}_{i_l} \right) I_{i_1,\ldots,i_{k+1}}(t), \tag{38}$$

for $k = 2, \ldots,$ where

$$I_{i_1,\ldots,i_{k+1}}(t) = \int_0^t \int_0^{t_k} \cdots \int_0^{t_2} e^{h_{i_1} t_1 + \sum_{l=2}^{k} h_{i_l}(t_l - t_{l-1}) + h_{i_{k+1}}(t - t_k)} dt_1 \cdots dt_k \tag{39}$$

For example, the second factorial moment, often sought for statistical testing, is

$$\mathbf{M}^{[2]}(t) = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \mathbf{B}_i \mathbf{\Lambda}_R \mathbf{B}_j \mathbf{\Lambda}_R \mathbf{B}_k I_{ijk}(t), \tag{40}$$

where

$$I_{ijk}(t) = \begin{cases} \dfrac{t^2}{2} e^{h_i t} & h_i = h_j = h_k, \\ \dfrac{t e^{h_j t}}{h_j - h_k} - \dfrac{e^{h_j t} - e^{h_k t}}{(h_j - h_k)^2} & h_i = h_j, h_j \neq h_k, \\ \dfrac{1}{h_i - h_j} \left(t e^{h_i t} - \dfrac{e^{h_j t} - e^{h_k t}}{h_j - h_k} \right) & h_i \neq h_j, h_i = h_k, \\ \dfrac{1}{h_i - h_j} \left(\dfrac{e^{h_i t} - e^{h_k t}}{h_i - h_k} - t e^{h_j t} \right) & h_i \neq h_j, h_j = h_k, \\ \dfrac{1}{h_i - h_j} \left(\dfrac{e^{h_i t} - e^{h_k t}}{h_i - h_k} - \dfrac{e^{h_j t} - e^{h_k t}}{h_j - h_k} \right) & \text{otherwise.} \end{cases} \tag{41}$$

These derivations generalize the work of Hobolth et al. [12]. These authors use an eigen decomposition of the infinitesimal generator and expressions (37) and (41) to compute the mean and the variance of the number of mutations between a fixed pair of states. In other words, they consider a particular counting process with set R consisting of exactly one pair of indices. As illustrated in the Examples section, such a labeling scheme is not applicable to many evolutionary applications, where R consists of more than one pair.

In deriving restricted factorial moments $\mathbf{M}^{[k]}$, we replace differentiation in (31) with integral Eq. (33). An alternative approach to computing derivatives of a matrix exponential also with the help of integration is described in [26]. These authors rely on an eigen decomposition of the infinitesimal generator. The decomposition coupled with the Cauchy integral formula helps the authors reduce differentiation of a matrix exponential to evaluating one dimensional complex variable integrals. Although computationally these two approaches are equivalent, our method is more accessible as it does not require any knowledge of complex analysis. Further, Narayana and Neuts show how to calculate the first two restricted factorial moments without an eigen decomposition of the CTMC infinitesimal generator [15]. These authors employ series expansions based on the uniformization method. Uniformization can become highly inefficient when the rates of CTMC transitions vary considerably. In evolutionary

applications, this variation arises most often in the context of amino acid evolution, where selection imposes strong constraints on the fixation of mutations [1,11].

When information about the trait value is missing at some tips of a phylogenetic tree, marginal factorial moments of the evolutionary counting process,

$$\mu_i^{[k]}(t) = \mathrm{E}\left(N_t^{[k]} \mid X_0 = i\right), \quad i = 1, \ldots, m, \tag{42}$$

are needed for computations. For an irreducible CTMC, Ball and Mine derive a closed-form expression for marginal first factorial moment vector $\boldsymbol{\mu}^{[1]}(t) = \left(\mu_1^{[1]}(t), \ldots, \mu_m^{[1]}(t)\right)^T$, using properties of the CTMC fundamental matrix [3]. In general, marginal factorial moments can be recovered from the restricted factorial moments developed here as

$$\boldsymbol{\mu}^{[k]}(t) = \mathbf{M}^{[k]}(t)\mathbf{1}. \tag{43}$$

5 Examples

Markov models and phylogenies

We first describe how a CTMC model together with a phylogenetic tree structure defines a probabilistic model for discrete trait evolution [7]. Consider a rooted binary tree τ with internal nodes labeled as $1, \ldots, k-1$ starting with the root and terminal nodes (tips) labeled as $k, \ldots, 2k-1$. We observe evolutionary trait values, $\mathbf{d} = (d_k, \ldots, d_{2k-1})$, only at the tips. All edges, commonly called branches, of the tree have weights/lengths that denote evolutionary times between the bifurcation events. We label this collection of weights as $\mathbf{t}$. To define the likelihood of observed data $\mathbf{d}$, we first augment $\mathbf{d}$ with missing trait values, $\mathbf{b} = (b_1, \ldots, b_{k-1})$, at the internal nodes of the tree τ. We then assume that given the state at a parent node, its children trait values evolve independently along the two descending branches according to a CTMC with infinitesimal generator $\boldsymbol{\Lambda}$. Therefore, setting an initial distribution $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)$ for root trait values b_1, the likelihood of the augmented data is

$$\Pr(\mathbf{d}, \mathbf{b} | \boldsymbol{\Lambda}, \mathbf{t}, \boldsymbol{\nu}, \tau) = \nu_{b_1} \prod_{(i,j)} p_{b_i b_j}(t_{ij}) \prod_{(k,l)} p_{b_k d_l}(t_{kl}), \tag{44}$$

where (i, j) spans all internal node parent-child pairs, (k, l) ranges over all parent-child pairs, where the child is a tip of the tree. Parameters t_{ij} and t_{kl} are lengths of branches connecting children with their parents, and $p_{ij}(t) = \left\{e^{\boldsymbol{\Lambda} t}\right\}_{ij}$ are the CTMC finite-time transition probabilities. We find the phylogenetic likelihood of the observed data by summing over values of the missing data

$$\Pr(\mathbf{d} | \boldsymbol{\Lambda}, \mathbf{t}, \boldsymbol{\nu}, \tau) = \sum_{\mathbf{b}} \nu_{b_1} \prod_{(i,j)} p_{b_i b_j}(t_{ij}) \prod_{(k,l)} p_{b_k d_l}(t_{kl}). \tag{45}$$

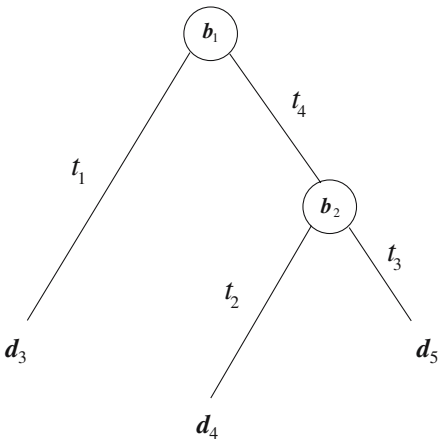

Fig. 2 Phylogenetic tree describing the evolutionary relationship among three organisms with observed trait states d_3, d_4, and d_5. Internal node variables b_1 and b_2 represent unknown states of the trait at times when ancestral organisms split into different lineages

This summation proceeds over all possible trait values at the internal nodes of τ. For example, the likelihood of the tip data on the tree depicted in Fig. 2 is

$$\Pr(\mathbf{d} \mid \mathbf{\Lambda}, \mathbf{t}, \boldsymbol{\nu}, \tau) = \sum_{b_1=1}^{m} \sum_{b_2=1}^{m} \nu_{b_1} p_{b_1 d_3}(t_1) p_{b_1 b_2}(t_4) p_{b_2 d_4}(t_2) p_{b_2 d_5}(t_3). \tag{46}$$

We view (45) as an integrated likelihood, where $\prod_{(k,l)} p_{b_k d_l}(t_{kl})$ is integrated over a *prior* distribution on $\mathbf{b}$

$$\Pr(\mathbf{b}) = \nu_{b_1} \prod_{(i,j)} p_{b_i b_j}(t_{ij}). \tag{47}$$

To ease further presentation we omit implicit conditioning on $\mathbf{\Lambda}$, $\mathbf{t}$, $\boldsymbol{\nu}$, and τ in (47) and hereafter as we assume that these model parameters are fixed. It is also possible to compute the *posterior* distribution of the internal nodes

$$\Pr(\mathbf{b} \mid \mathbf{d}) = \Pr(b_1 \mid \mathbf{d}) \prod_{(i,j)} \Pr(b_j \mid b_i, \mathbf{d}), \tag{48}$$

where $\Pr(b_1 \mid \mathbf{d})$ and $\Pr(b_j \mid b_i, \mathbf{d})$ are functions of the transition probabilities $\mathbf{P}(t) = \{p_{ij}(t)\}$ [19,23]. The conditional independence reflected in expressions (45), (47), and (48) allows for efficient computation of the phylogenetic likelihood and related quantities through the specialized sum-product algorithm, also known as the pruning algorithm [6,7].

Prior and posterior distributions of the number of evolutionary changes

We start with a hypothetical example that resembles a typical problem from evolutionary developmental biology. Consider the tree in Fig. 2 and suppose that we observe

a binary trait at the tips of this tree. Let N_τ be the total number of times the trait changes its state during its evolution over tree τ. We would like to know the probability mass function of N_τ with and without conditioning on the observed data. If we define conditional probabilities

$$\hat{q}_{ij}(n,t) = q_{ij}(n,t)/p_{ij}(t) = \Pr(N_t = n, \mid X_0 = i, X_t = j), \tag{49}$$

then

$$\begin{aligned}\Pr(N_\tau = n) = &\sum_{n_1+\cdots+n_4=n} \sum_{b_1=1}^{m} \sum_{b_2=1}^{m} q_{b_1}(n_1,t_1)\hat{q}_{b_1 b_2}(n_4,t_4) \\ &\times q_{b_2}(n_2,t_2) q_{b_2}(n_3,t_3)\Pr(b_1,b_2),\end{aligned} \tag{50}$$

where $q_i(n,t) = \Pr(N_t = n \mid X_0 = i)$ and vector $(n_1,\ldots,n_4)$ ranges over all possible ways n transitions can be distributed among the branches of τ. Such distribution can be accomplished by generating all possible integer partitions of n into 4 parts, permuting these partitions, and keeping only unique vectors $(n_1,\ldots,n_4)$. When there is no data present at the tips, the probabilities of the terminal branches in (50) do not involve conditioning on the Markov chain end state. Similar to the above derivations, through conditioning on the internal node states, we arrive at the posterior probability mass function of N_τ,

$$\begin{aligned}\Pr(N_\tau = n \mid \mathbf{d}) = &\sum_{n_1+\cdots+n_4=n} \sum_{b_1=1}^{m} \sum_{b_2=1}^{m} \hat{q}_{b_1 d_3}(n_1,t_1)\hat{q}_{b_1 b_2}(n_4,t_4) \\ &\times \hat{q}_{b_2 d_4}(n_2,t_2)\hat{q}_{b_2 d_5}(n_3,t_3)\Pr(b_1,b_2 \mid \mathbf{d}).\end{aligned} \tag{51}$$

Introduction of conditional probabilities $\hat{q}_{ij}(t)$ serves only for notational convenience, since transition probabilities in (49) cancel out in (50) and (51) during multiplication by the prior and posterior probabilities of internal node states. Moreover, it is easy to see that while calculating $\Pr(N_\tau = n)$ and $\Pr(N_\tau = n \mid \mathbf{d})$, we can efficiently distribute multiplication along the phylogenetic tree similar to the pruning algorithm mentioned above.

We now consider a numerical example, where we set branch lengths for the tree in Fig. 2 to $t_1 = 0.3$, $t_2 = 0.2$, $t_3 = 0.1$, and $t_4 = 0.1$. We also assume $\lambda_1 = 1$ and $\lambda_2 = 2$ using parameterization (3) of the two-state CTMC generator and the stationary distribution of the chain at the root of τ. Given such model parameters, we compute prior probabilities $\Pr(N_\tau = n)$ and posterior probabilities $\Pr(N_\tau = n \mid \mathbf{d}_1)$ and $\Pr(N_\tau = n \mid \mathbf{d}_2)$ for $n = 0,\ldots,5$ under two different data patterns $\mathbf{d}_1 = (1,2,2)$ and $\mathbf{d}_2 = (2,2,1)$. Plots in the bottom row of Fig. 3 show the first six entries of the probability mass function of N_τ (vertical bars). Without observing any data at the tips (the first column of Fig. 3) the probability mass function of N_τ exhibits Poisson-like shape with high probability of zero changes reflecting the short branch lengths of tree τ. The first data pattern strongly supports one change in the evolutionary history of the trait (the second column), while under the second data pattern probabilities of one and two total number of changes are approximately equal (the third column).

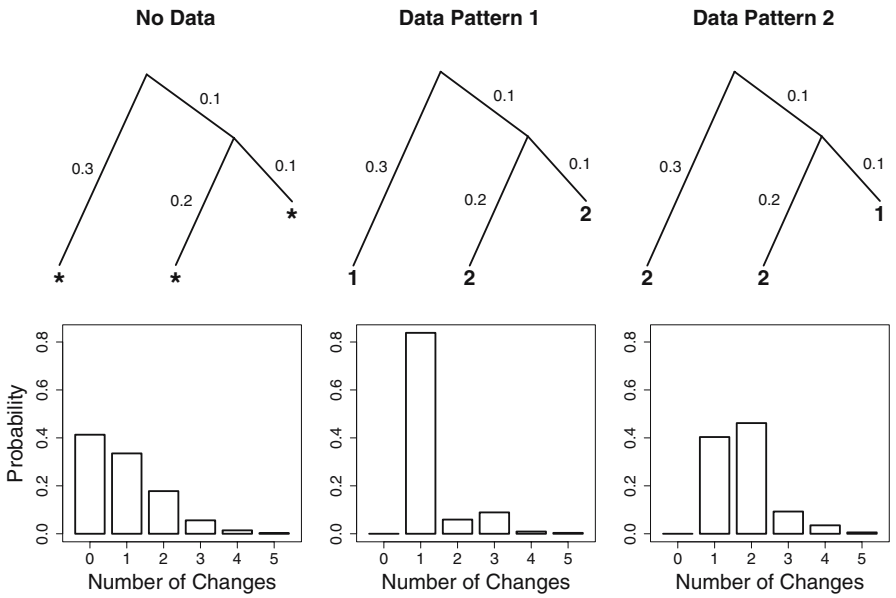

Fig. 3 Effect of observed data on distribution of the number of changes in the evolutionary history of a binary trait. The *top plots* show the same phylogenetic tree with no observed data at its tips (*left*) and two different observed data patterns at the tips (*middle* and *right*). The *bottom plots* depict probabilities of 0–5 changes (*vertical bars*) for each data pattern in the top plots

In evolutionary developmental biology, researchers are often interested in how many times a trait changed its state during an evolutionary history. For example, one common hypothesis may postulate that a trait changed its state no more than one time in its evolutionary history. The alternative to this hypothesis implies multiple independent origins of a new trait state. We can envision formally testing these hypotheses statistically in a Bayesian framework. Since we now know how to compute quantities $\Pr(N_\tau \leq 1)$ and $\Pr(N_\tau \leq 1 \,|\, \mathbf{d})$, it is straightforward to integrate them over the prior and posterior distribution of model parameters respectively and set up a Bayes factor test [13].

Counting mutations in DNA sequences

We now demonstrate that counting mutations in DNA sequences provides an effective data exploratory and model diagnostic tool. We start with a multiple sequence alignment of N DNA sequences $\mathbf{Y} = \{Y_{nl}\}$, $n = 1, \ldots, N$, $l = 1 \ldots, L$. The L columns of the alignment, called sites, are observations generated by the CTMC evolutionary process. To simplify mathematics involved in probabilistic and statistical treatment of DNA evolution, it is often assumed that sites $\mathbf{Y}_1, \ldots, \mathbf{Y}_L$ are independent and identically distributed (iid). Since both of these assumptions are very unlikely to hold in nature, relaxing them remains an active area of research in the theory of molecular evolution [24,32,33]. A common approach to assigning nonidentical distributions

to different sites is to divide them into a small number of classes and allow model parameters to vary among but not within the classes [31]. We describe one such site partitioning that plays an important role in protein coding regions.

In protein coding regions, 61 out of 64 possible triplets of DNA, called codons, are translated into 20 amino acids. Therefore, different codons may encode the same amino acid. Such redundancy in the genetic code allows the three codon positions to mutate at different rates. The third codon position experiences mutations more frequently than the other two positions, since a change in this position does not generally lead to a change at the amino acid level. Differences between mutation rates of the first and the second positions are less pronounced with the first position changing slightly faster. We design a simulation study that demonstrates how mean mutational counts help detect heterogeneity of the evolutionary rate among codon positions without having to explicitly model the heterogeneity first. We also use mutational counts to check the adequacy of the nucleotide model that we use in our calculations.

Using the method of Rambaut and Grassly [25] we simulate two different nucleotide alignments of three sequences assuming the evolutionary relationship among them described by the tree in Fig. 2. Branch lengths are the same as in the first example. Both data sets are simulated using an HKY parameterization [10] of the CTMC infinitesimal generator,

$$\mathbf{\Lambda} = \begin{pmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & - \end{pmatrix}, \tag{52}$$

where $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$ is the stationary distribution of the chain, α is called a "transition" rate, and β is called a "transversion" rate. Nucleotides divide into two classes: purines $\{A, G\}$ and pirimidines $\{C, T\}$. Mutations are called "transitions" if they do not change the class assignment of a nucleotide and "transversions" otherwise. We rescale matrix $\mathbf{\Lambda}$ such that $\sum_u \lambda_{uu}\pi_u = -1$, a common procedure often needed for identifiability in statistical phylogenetics [7]. The rescaling leads to a decrease in number of free model parameters and allows the reparameterization of $\mathbf{\Lambda}$ in terms of its stationary distribution $\boldsymbol{\pi}$ and "transition/transversion" rate ratio $\kappa = \alpha/\beta$. In both simulations we set all stationary probabilities to 0.25 and $\kappa = 4.0$.

Sites in the first alignment are simulated under a model with iid sites. The second data set is simulated using a codon partitioning (CP) model. In this model three CTMC generators are used for independently (but not identically) distributed sites in each of the three codon positions, $\mathbf{\Lambda}_c = r_c\mathbf{\Lambda}$, $c = 1, 2, 3$. Scaling factors are set to $r_1 = 1.5, r_2 = 1.0$, and $r_3 = 3.0$. Before examining the simulated datasets, we assume no knowledge of possible heterogeneity and perform calculations under the simpler model that assumes iid alignment sites.

We compute the mean number of mutations at each site in both alignments. For each site l we condition on nucleotides $\mathbf{Y}_l$ observed at the site,

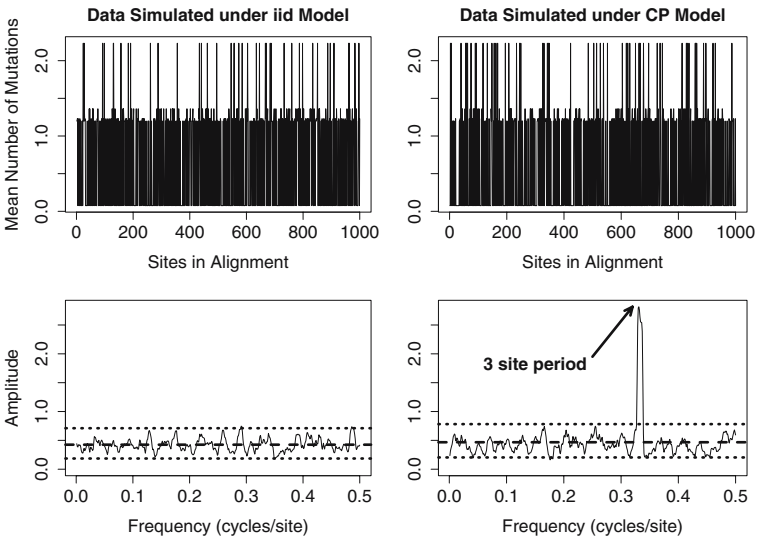

Fig. 4 Spectral analysis of site-specific mutational counts. The *top two plots* depict posterior mean number of mutations for each site in the simulated alignments (*vertical bars*). The *bottom plots* show smoothed spectrums of the mutational counts accompanied by the spectrum of white noise (*dashed lines*) and its 95% confidence bounds (*dotted lines*). In the *bottom right plot* an *arrow* marks the spectrum peak corresponding to a period of three sites

$$\mathrm{E}(N_\tau \mid \mathbf{Y}_l) = \sum_{b_1=1}^{s}\sum_{b_2=1}^{s}\Big[\hat{m}^{[1]}_{b_1 Y_{1l}}(t_1) + \hat{m}^{[1]}_{b_2 Y_{2l}}(t_2) + \hat{m}^{[1]}_{b_2 Y_{3l}}(t_3) + \hat{m}^{[1]}_{b_1 b_2}(t_4)\Big]\Pr(b_1, b_2 \mid \mathbf{Y}_l), \tag{53}$$

where

$$\hat{m}^{[1]}_{ij}(t) = m^{[1]}_{ij}(t)/p_{ij}(t) = \mathrm{E}(N_t \mid X_0 = i, X_t = j). \tag{54}$$

For a general tree, formula (53) suggests that computing the mean number of labeled transition reduces to

$$\sum_{\mathbf{b}} \hat{m}^{[1]}_{b_i b_j}(t_{ij})\Pr(\mathbf{b} \mid \mathbf{d}), \tag{55}$$

where i and j are nodes of the tree connected by a branch with length t_{ij}. We accomplish computations of these quantities by bookkeeping of local calculations on the tree similar to an algorithm by Schadt et al. [27] for calculating phylogenetic likelihood derivatives.

In our calculations of the mean mutational counts, we deliberately further misspecify the evolutionary model by setting "transition/transversion" rate ratio $\kappa = 1.0$ and branch lengths $\mathbf{t} = (0.28, 0.21, 0.12, 0.09)$ instead of values 4.0 and $(0.3, 0.2, 0.1, 0.1)$ used in the simulation. In the top plot of Fig. 4, we indicate the

Table 1 Mean number of "transitions" and "transversions". The prior and posterior (Post.) expectations are calculated for the number of "transitions" and "transversions" in the two alignments simulated under the independent and identically distributed sites (iid) and codon partitioning (CP) models

κ	"Transitions"			"Transversions"		
	Prior	iid Post.	CP Post.	Prior	iid Post.	CP Post.
1.0	233.3	333.7	313.0	466.7	342.3	316.4
2.0	350.0	395.6	373.6	350.0	281.1	257.3
4.0	466.7	455.4	433.0	233.3	244.8	221.4

mean number of mutations for each site (vertical bars) in the simulated alignments. Initially the site-specific mean mutational counts of the iid and CP data show no obvious differences in their patterns of variation. We estimate spectral densities (spectrums) of site-specific mutational counts treating the latter as time series. The bottom two plots of Fig. 4 show mutational count spectrums, smoothed with a Danniel window filter for both simulated alignments [4]. The smoothed spectrum for the iid alignment sites demonstrates no sign of periodicity as all amplitudes fall between the 95% bounds for white noise (dotted lines). The spectrum of the CP sites has a strong peak at frequency $\frac{1}{3}$ that corresponds to a period of three sites, the size of a codon. Therefore, despite the crude misspecification of the substitution model, site-specific mutational counts enable us to detect a repeated pattern of variation among codon positions.

Next, we examine whether discrepancies between the prior and posterior expected number of "transitions" and "transversions" can help us illuminate the deliberate misspecification of the "transition/transversion" rate ratio κ. We first define "transition", $R_1 = \{(1, 2), (2, 1), (3, 4), (4, 3)\}$, and "transverstion", $R_2 = \{(i, j) : i, j = 1, \ldots, 4, i \neq j\} \setminus R_1$, labeling sets. Then, we calculate the mean number of "transitions" and "transversions" for $\kappa = 1, 2, 4$ and $\mathbf{t} = (0.28, 0.21, 0.12, 0.09)$ without conditioning on the data and conditioning on the alignment sites simulated under the iid and CP models (see Table 1). A priori we expect $\kappa/2$ times as many "transitions" than "transversions", because there are eight possible "transversions" and only four possible "transitions". Table 1 reveals that only for the "true" value of $\kappa = 4$ the prior and posterior expectations agree for both mutational classes and for both simulation conditions. On the contrary, when we set κ to 1 or 2, the observed data "surprise" us by too many "transitions" and too few "transversions" relative to our prior expectations. Therefore, mutational counts can serve as diagnostic tools to measure discrepancy between data and a model chosen for analyzing these data. Prior or posterior predictive model checks offer a formal statistical framework for quantifying such discrepancies [14].

6 Conclusion

In this paper, we study properties of counting processes induced by CTMC models for discrete trait evolution. A two-state CTMC is an important model often used

in evolutionary developmental biology. For this model we derive closed-form expressions for the probability mass and probability generating functions. In the multi-state generalization we show that, when the CTMC generator is diagonalizable, it is possible to obtain closed-form solutions at least for the first couple of moments of the counting process. Similarity of the generator to a symmetric and, hence, diagonalizable matrix is not a very restrictive assumption since the majority of evolutionary models are reversible.

In our derivations we allow for an arbitrary labeling of Markov chain transitions that the evolutionary counting process registers. This flexibility is important as evolutionary biologists often divide mutations into classes and search for over-representation of mutations that belong to a certain class. For example, excessive number of nucleotide mutations that do not result in a change at the amino acid level (synonymous mutations) indicates natural selection [20]. Instead of fitting molecular data to complicated models of evolution with different rates of synonymous and nonsynonymous changes, we can use simpler models and compute the mean number of mutations in each mutational class a posteriori as Nielsen proposes [19]. However, this author relies upon simulations to compute the mean mutational counts raising concerns about computational efficiency.

Current simulation approaches estimate properties of evolutionary counting processes with very inefficient rejection sampling algorithms [19]. Although these methods can often provide sufficient approximations of the desired quantities for small state-space CTMCs and datasets, the computational time needed for analysis of large datasets grows prohibitive. Our analytic results open the door for fast computations of important properties of evolutionary counting processes. In our derivations we essentially consider a counting process on one branch of a phylogenetic tree. From these derivations, it is straightforward to develop efficient recursive algorithms, similar to Felsenstein's [6] pruning that combine local, one-branch calculations and compute properties of evolutionary counting processes along the whole phylogenetic tree.

In our examples we provide applications of counting processes to evolutionary biology problems. We first consider binary trait evolution along a phylogenetic tree. We compute the truncated probability mass function of the number of changes that occurred during the evolutionary history with and without trait values observed at the tips of the tree. In the second example we show that a spectral analysis of site-specific mean mutational counts can be used as a simple and effective method of detecting hidden patterns of variation in DNA sequence evolution. We also demonstrate that dividing mutations into classes via labeling of CTMC transitions allows for easy checks of evolutionary model adequacy. We believe that counting processes will see many interesting applications in evolutionary biology and hope that results presented in this paper will make using evolutionary counting processes more practical for analyzing large datasets and testing complex evolutionary hypotheses.

Acknowledgments We are grateful to Dr. Todd Oakley for introducing us to the independent origins questions and, as a result, stimulating our interest in evolutionary counting processes. We also would like to thank Dr. Kenneth Lange for thought provoking discussions and for bringing to our attention his work with Dr. Eric Schadt on differentiation of matrix exponentials.

References

1. Adachi, J., Hasegawa, M.: Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**, 459–468 (1996)
2. Ball, F.: Empirical clustering of bursts of openings in Markov and semi-Markov models of single channel gating incorporating time interval omission. Adv. Appl. Probab. **29**, 909–946 (1997)
3. Ball, F., Milne, R.K.: Simple derivations of properties of counting processes associated with Markov renewal processes. J. Appl. Probab. **42**, 1031–1043 (2005)
4. Chatfield, C.: The Analysis of Time Series: An Introduction. Chapman & Hall, London (2004)
5. Darroch, J.N, Morris, K.W.: Some passage-time generating functions for discrete-time and continuous-time finite Markov chains. J. Appl. Probab. **4**, 496–507 (1967)
6. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **13**, 93–104 (1981)
7. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates Inc., Sunderland (2004)
8. Fitch, W.M., Bush, R.M., Bender, C.A., Cox, N.J.: Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94**, 7712–7718 (1997)
9. Guttorp, P.: Stochastic Modeling of Scientific Data. Chapman & Hall, Suffolk (1995)
10. Hasegawa, M., Kishino, H., Yano, T.: Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**, 160–174 (1985)
11. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**, 10915–10919 (1992)
12. Hobolth, A., Jensen, J.L.: Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat. Appl. Gen. Mol. Biol. **4**, Article 18 (2005)
13. Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)
14. Meng, X.L.: Posterior predictive P-values. Ann. Stat. **22**, 1142–1160 (1994)
15. Narayana, S., Neuts, M.F.: The first two moment matrices of the counts for the Markovian arrival process. Stoch. Models **8**, 459–477 (1992)
16. Neuts, M.F.: A versatile Markovian point process. J. Appl. Probab. **16**, 764–779 (1979)
17. Neuts, M.F.: Models based on the Markovian arrival process. IEICE Trans. Commun. **E75-B**, 1255–1265 (1992)
18. Neuts, M.F.: Algorithmic Probability: a Collection of Problems. Chapman and Hall, London (1995)
19. Nielsen, R.: Mapping mutations on phylogenies. Syst. Biol. **51**, 729–739 (2002)
20. Nielsen, R.: Molecular signatures of natural selection. Ann. Rev. Gen. **39**, 197–218 (2005)
21. Oakley, T.H., Cunningham, C.W.: Molecular phylogenetic evidence for the independent evolutionary origin of an arthropod compound eye. Proc. Natl. Acad. Sci. USA **99**, 1426–1430 (2002)
22. Pagel, M.: Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. B **255**, 37–45 (1994)
23. Pagel, M.: The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. **48**, 612–622 (1999)
24. Pollock, D.D., Taylor, W.R., Goldman, N.: Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**, 187–198 (1999)
25. Rambaut, A., Grassly, N.C.: Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**, 235–238 (1997)
26. Schadt, E., Lange, K.: Codon and rate variation models in molecular phylogeny. Mol. Biol. Evol. **19**, 1534–1549 (2002)
27. Schadt, E.E., Sinsheimer, J.S., Lange, K.: Computational advances in maximum likelihood methods for molecular phylogeny. Genome Res. **8**, 222–233 (1998)
28. Siepel, A., Pollard, K.S., Haussler, D.: New methods for detecting lineage-specific selection. In: Proceedings of the 10th international conference on research in computational molecular biology, pp. 190–205 (2006)
29. Suchard, M.A., Weiss, R.E., Dorman, K.S., Sinsheimer, J.S.: Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. J. Am. Stat. Assoc. **98**, 427–437 (2002)
30. Templeton, A.R.: Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics **144**, 1263–1270 (1996)
31. Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**, 306–314 (1994)

32. Yang, Z.: A space-time process model for the evolution of DNA sequences. Genetics **139**, 993–1005 (1995)
33. Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M.K.: Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**, 431–449 (2000)
34. Zheng, Q.: On the dispersion index of a Markovian molecular clock. Math. Biosci. **172**, 115–128 (2001)