PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 3985–3995
doi:10.1098/rstb.2008.0176
Published online 7 October 2008

Fast, accurate and simulation-free stochastic mapping

Vladimir N. Minin[1,*] **and Marc A. Suchard**[2,3,4]

[1] *Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA*
[2] *Department of Biomathematics, and* [3] *Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-1766, USA*
[4] *Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1766, USA*

Mapping evolutionary trajectories of discrete traits onto phylogenies receives considerable attention in evolutionary biology. Given the trait observations at the tips of a phylogenetic tree, researchers are often interested where on the tree the trait changes its state and whether some changes are preferential in certain parts of the tree. In a model-based phylogenetic framework, such questions translate into characterizing probabilistic properties of evolutionary trajectories. Current methods of assessing these properties rely on computationally expensive simulations. In this paper, we present an efficient, simulation-free algorithm for computing two important and ubiquitous evolutionary trajectory properties. The first is the mean number of trait changes, where changes can be divided into classes of interest (e.g. synonymous/non-synonymous mutations). The mean evolutionary reward, accrued proportionally to the time a trait occupies each of its states, is the second property. To illustrate the usefulness of our results, we first employ our simulation-free stochastic mapping to execute a posterior predictive test of correlation between two evolutionary traits. We conclude by mapping synonymous and non-synonymous mutations onto branches of an HIV intrahost phylogenetic tree and comparing selection pressure on terminal and internal tree branches.

Keywords: stochastic mapping; counting processes; evolutionary rewards; posterior predictive diagnostics

1. INTRODUCTION AND BACKGROUND

Reconstructing evolutionary histories from present-day observations is a central problem in quantitative biology. Phylogenetic estimation is one example of such reconstruction. However, phylogenetic reconstruction alone does not provide a full picture of an evolutionary history, because evolutionary paths (mappings) describing trait states along the phylogenetic tree remain hidden. Although one is rarely interested in detailed reconstruction of such mappings, certain probabilistic properties of the paths are frequently used in evolutionary hypotheses testing (Nielsen 2002; Huelsenbeck *et al.* 2003; Leschen & Buckley 2007). For example, given a tree and a Markov model of amino acid evolution, one can compute the expected number of times a transition from a hydrophobic to a hydrophilic state occurs, conditional on the observed amino acid sequence alignment. Such expectations can inform researchers about model adequacy and provide insight into the features of the evolutionary process overlooked by standard phylogenetic techniques (Dimmic *et al.* 2005).

Nielsen (2002) introduced stochastic mapping of trait states on trees and employed this new technique in a model-based evolutionary hypothesis testing context. The author starts with a discrete evolutionary trait X that attains m states. He further assumes that this trait evolves according to an evolutionary model described by a parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ consists of a tree τ with n tips and branch lengths $\boldsymbol{T}=(t_1,\ldots,t_{B_n})$, root distribution $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_m)$ and a continuous-time Markov chain (CTMC) generator $\boldsymbol{Q}=\{q_{ij}\}$ for $i,j=1,\ldots,m$. Let mapping $\boldsymbol{M_\theta}=(\{X_{1t}\},\ldots,\{X_{B_n t}\})$ be a collection of CTMC trajectories along all branches of τ and $H(\boldsymbol{M_\theta})$ be a real-valued summary of $\boldsymbol{M_\theta}$. Clearly, even when parameters θ are fixed, $h(\boldsymbol{M_\theta})$ remains a random variable. Nielsen (2002) proposed to test evolutionary hypotheses using prior and posterior expectations $E[H(\boldsymbol{M_\theta})]$ and $E[H(\boldsymbol{M_\theta})|\boldsymbol{D}]$, where $\boldsymbol{D}=(D_1,\ldots,D_n)$ are trait values observed at the n tips of τ. Since these expectations are deterministic functions of $\boldsymbol{\theta}$ and $\boldsymbol{D}$, they can be used as discrepancy measures for posterior predictive p-value calculations (Meng 1994; Gelman *et al.* 1996).

A major advantage of Nielsen's stochastic mapping framework is its ability to account for uncertainty in model parameters, including phylogenies. A major limitation of stochastic mapping is its current implementation that relies on time-consuming simulations. In describing his method for calculating $E[H(\boldsymbol{M_\theta})]$ and $E[H(\boldsymbol{M_\theta})|\boldsymbol{D}]$, Nielsen (2002) wrote 'In general, we can not evaluate sums in equations 5 and 6 directly, because the set [of all possible mappings] is not of finite size.' However, the infinite number of possible mappings does not prevent one from explicitly

---

[*] Author for correspondence (vminin@u.washington.edu).

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

This journal is © 2008 The Royal Society

calculating $E[H(\boldsymbol{M}_\theta)]$ for some choices of H. For example, if

$$H(\boldsymbol{M}_\theta) = \begin{cases} 1 & \text{if } \boldsymbol{M}_\theta \text{ is consistent with } \boldsymbol{D} \\ 0 & \text{if } \boldsymbol{M}_\theta \text{ is inconsistent with } \boldsymbol{D}, \end{cases} \tag{1.1}$$

then $E[H(\boldsymbol{M}_\theta)] = \Pr(\boldsymbol{D})$, the familiar phylogenetic likelihood that can be evaluated without simulations (Felsenstein 2004). Therefore, hope remains that other choices of H may also permit evaluation of $E[H(\boldsymbol{M}_\theta)]$ and $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$ without simulations.

In this paper, we consider a class of additive mapping summaries of the form

$$H(\boldsymbol{M}_\theta) = \sum_{b\in\Omega} h(\{X_{bt}\}), \tag{1.2}$$

where $h(\{X_{bt}\})$ is a one-dimensional summary of the Markov chain path along branch b and Ω is an arbitrary subset of all branches of τ. Moreover, we restrict our attention to the two most popular choices of function h. Let $\mathcal{L} \subset \{1, \dots, m\}^2$ be a set of ordered index pairs that label transitions of trait X. For each Markov path $\{X_t\}$ and interval $[0, t)$, we count the number of labelled transitions in this interval and arrive at

$$h_1(\{X_t\}) = N_{\mathcal{L}} \text{ - number of state transitions labelled by set } \mathcal{L}, \tag{1.3}$$

where we omit dependence on $\boldsymbol{\theta}$ and t for brevity. Although our second choice of h is more abstract, it is motivated by Huelsenbeck *et al.* (2003), who used Nielsen's stochastic mapping algorithm to calculate the mean dwelling time of a trait in a particular state. Let $\boldsymbol{w} = (w_1, \dots, w_m)$ be a set of rewards assigned to each trait state. Trait X is 'rewarded' the amount $t \times w_i$ for spending time t in state i. We obtain the total reward of Markov path $\{X_t\}$ by summing up all rewards that X accumulates during interval $[0, t)$,

$$h_2(\{X_t\}) = R_{\boldsymbol{w}} \text{ - evolutionary reward defined by vector } \boldsymbol{w}. \tag{1.4}$$

To obtain dwelling times of X in a predefined set of trait states, we set $w_i = 1$ if i belongs to the set of interest and $w_i = 0$ otherwise.

For these two choices of function h, we provide an algorithm for exact, simulation-free computation of $E[H(\boldsymbol{M}_\theta)]$ and $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$. Similar to phylogenetic likelihood calculations of $\Pr(\boldsymbol{D})$, this algorithm relies on the eigen decomposition of $\boldsymbol{Q}$ and requires traversing τ. Our work generalizes and unifies exact calculations previously developed for stochastic mapping (Holmes & Rubin 2002; Guindon *et al.* 2004; Dutheil *et al.* 2005). Despite the restricted form of the considered mapping summaries, our results cover nearly all current applications of stochastic mapping. We conclude with two applications of stochastic mapping illustrating the capabilities of exact computation. In our first example, we examine coevolution of two binary traits and demonstrate that a previously developed simulation-based test of independent evolution can be executed without simulations. We then turn to a large codon Markov state space, on which simulation-based stochastic mapping generally experiences severe computational limitations. Using our exact computations, we study temporal patterns of synonymous and non-synonymous mutations in intrahost HIV evolution.

2. LOCAL, ONE BRANCH CALCULATIONS

In this section, we provide mathematical details needed for calculating expectations of stochastic mapping summaries on one branch of a phylogenetic tree. We first motivate the need for such local computations by making further analogies between calculations of $\Pr(\boldsymbol{D})$ and expectations of stochastic mapping summaries. The additive form of H reduces calculation of $E[H(\boldsymbol{M}_\theta)]$ and $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$ to compute branch-specific expectations $E[h(\{X_{bt}\})]$ and $E[h(\{X_{bt}\})\,|\,\boldsymbol{D}]$. Recall that according to the most phylogenetic models, trait X evolves independently on each branch of τ, conditional on trait states at all internal nodes of τ. This conditional independence is the key behind the dynamic programming algorithm that allows for efficient calculation of $\Pr(\boldsymbol{D})$ (Felsenstein 1981). For this likelihood calculation algorithm, it suffices to compute finite-time transition probabilities $\boldsymbol{P}(t) = \{p_{ij}(t)\}$, where

$$p_{ij}(t) = \Pr(X_t = j\,|\,X_0 = i), \tag{2.1}$$

for arbitrary branch length t. Similarly, to obtain $E[h(\{X_t\})]$ and $E[h(\{X_t\})\,|\,\boldsymbol{D}]$, we require means of computing local expectations $\boldsymbol{E}(h, t) = \{e_{ij}(h, t)\}$, where

$$e_{ij}(h, t) = E\left[h(\{X_t\})1_{\{X_t = j\}}\,|\,X_0 = i\right] \tag{2.2}$$

and $1_{\{\cdot\}}$ is the indicator function. After illustrating how to compute $\boldsymbol{E}(N_{\mathcal{L}}, t)$ and $\boldsymbol{E}(R_{\boldsymbol{w}}, t)$ without resorting to simulations, we provide an algorithm that efficiently propagates local expectations $\boldsymbol{E}(h, t)$ and finite-time transition probabilities $\boldsymbol{P}(t)$ along τ to arrive at $E[h(\{X_t\})]$ and $E[h(\{X_t\})\,|\,\boldsymbol{D}]$.

(a) *Expected number of labelled Markov transitions*

Abstracting from phylogenetics, let $N_{\mathcal{L}}(t)$ count the number of labelled transitions of a CTMC $\{X_t\}$ during time interval $[0, t)$. It follows from the theory of Markov chain-induced counting processes that

$$\boldsymbol{E}(N_{\mathcal{L}}, t) = \int_0^t \mathrm{e}^{\boldsymbol{Q}z}\boldsymbol{Q}_{\mathcal{L}}\mathrm{e}^{\boldsymbol{Q}(t-z)}\,\mathrm{d}z, \tag{2.3}$$

where $\boldsymbol{Q}_{\mathcal{L}} = \{q_{ij} \times 1_{\{(i,j)\in\mathcal{L}\}}\}$ (Ball & Milne 2005). Since most evolutionary models are locally reversible, we can safely assume that $\boldsymbol{Q}$ is diagonalizable with eigen decomposition $\boldsymbol{Q} = \boldsymbol{U} \times \mathrm{diag}(d_1, \dots, d_m) \times \boldsymbol{U}^{-1}$, where eigenvectors of $\boldsymbol{Q}$ form the columns of $\boldsymbol{U}$, $d_1, \dots, d_m$ are the real eigenvalues of $\boldsymbol{Q}$, and $\mathrm{diag}(d_1, \dots, d_m)$ is a diagonal matrix with elements $d_1, \dots, d_m$ on its main diagonal. Such analytic or numeric diagonalization procedure permits calculation of finite-time transition probabilities $\boldsymbol{P}(t) = \boldsymbol{U} \times \mathrm{diag}(\mathrm{e}^{d_1 t}, \dots, \mathrm{e}^{d_m t}) \times \boldsymbol{U}^{-1}$, needed for likelihood calculations (Lange 2004). Minin & Suchard (2008) have shown that one can use the same eigen decomposition of $\boldsymbol{Q}$ to calculate local expectations

$$\boldsymbol{E}(N_{\mathcal{L}}, t) = \sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{S}_i\boldsymbol{Q}_{\mathcal{L}}\boldsymbol{S}_j I_{ij}(t), \tag{2.4}$$

where $\boldsymbol{S}_i = \boldsymbol{U}\boldsymbol{E}_i\boldsymbol{U}^{-1}$, $\boldsymbol{E}_i$ is a matrix with zero entries everywhere except at the *ii*-th entry, which is one, and

$$I_{ij}(t) = \begin{cases} t\mathrm{e}^{d_i t} & \text{if } d_i = d_j, \\ \dfrac{\mathrm{e}^{d_i t} - \mathrm{e}^{d_j t}}{d_i - d_j} & \text{if } d_i \neq d_j. \end{cases} \tag{2.5}$$

(b) *Expected Markov rewards*

For the reward process $R_{\boldsymbol{w}}(t)$, we define a matrix cumulative distribution function $\boldsymbol{V}(x,t) = \{V_{ij}(x,t)\}$, where

$$V_{ij}(x,t) = \Pr(R_{\boldsymbol{w}}(t) \leq x,\ X_t = j \mid X_0 = i). \tag{2.6}$$

Neuts (1995, ch. 5) demonstrated that local reward expectations can be expressed as

$$\boldsymbol{E}(R_{\boldsymbol{w}}, t) = -\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{V}^*(s,t)\Big|_{s=0}, \tag{2.7}$$

where

$$\boldsymbol{V}^*(s,t) = \int_0^\infty \mathrm{e}^{-sx}\,\mathrm{d}\boldsymbol{V}(x,t) = \mathrm{e}^{[\boldsymbol{Q} - \mathrm{diag}(w_1,\ldots,w_m)s]t} \tag{2.8}$$

is the Laplace–Stieltjes transform of $\boldsymbol{V}(x,t)$. It is easy to see that the matrix exponential in equation (2.8) satisfies the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}^*(s,t) = \boldsymbol{V}^*(s,t)[\boldsymbol{Q} - \mathrm{diag}(w_1,\ldots,w_m)s]. \tag{2.9}$$

Differentiating this matrix differential equation with respect to *s*, exchanging the order of differentiation and evaluating both sides of the resulting equation at $s = 0$, we arrive at the differential equation for local expectations

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}(R_{\boldsymbol{w}}, t) = \boldsymbol{E}(R_{\boldsymbol{w}}, t)\boldsymbol{Q} + \mathrm{e}^{\boldsymbol{Q}t}\mathrm{diag}(w_1,\ldots,w_m), \tag{2.10}$$

where $\boldsymbol{E}(R_{\boldsymbol{w}}, 0)$ is the $m \times m$ zero matrix. Multiplication of both sides of equation (2.10) by integrating factor $\mathrm{e}^{-\boldsymbol{Q}t}$ from the right and integration with respect to *t* produces the solution,

$$\boldsymbol{E}(R_{\boldsymbol{w}}, t) = \int_0^t \mathrm{e}^{\boldsymbol{Q}z}\mathrm{diag}(w_1,\ldots,w_m)\mathrm{e}^{\boldsymbol{Q}(t-z)}\,\mathrm{d}z. \tag{2.11}$$

Similarity between equations (2.3) and (2.11) invites calculation of the expected Markov rewards via spectral decomposition of $\boldsymbol{Q}$,

$$\boldsymbol{E}(R_{\boldsymbol{w}}, t) = \sum_{i=1}^m \sum_{j=1}^m \boldsymbol{S}_i \mathrm{diag}(w_1,\ldots,w_m)\boldsymbol{S}_j I_{ij}(t). \tag{2.12}$$

In summary, formulae (2.4) and (2.12) provide a recipe for exact calculations of local expectations for the number of labelled transitions and rewards.

3. ASSEMBLING PIECES TOGETHER OVER A TREE

(a) *Notation for tree traversal*

Let us label the internal nodes of τ with integers $\{1,\ldots,n-1\}$ starting from the root of the tree. Recall that we have already arbitrarily labelled the tips of τ with integers $\{1,\ldots,n\}$. Let $\mathcal{I}$ be the set of internal branches and $\mathcal{E}$ be the set of terminal branches of τ. For each branch $b \in \mathcal{I}$, we denote the internal node labels of the parent and child of branch *b* by $p(b)$ and $c(b)$, respectively. We use the same notation for each terminal branch *b* except $p(b)$, which is an internal node index, while $c(b)$ is a tip index. Let $\boldsymbol{i} = (i_1,\ldots,i_{n-1})$ denote the internal node trait states. Then, the complete likelihood of unobserved internal node states and the observed states at the tips of τ is

$$\Pr(\boldsymbol{i}, \boldsymbol{D}) = \pi_{i_1} \prod_{b \in \mathcal{I}} p_{i_{p(b)} i_{c(b)}}(t_b) \prod_{b \in \mathcal{E}} p_{i_{p(b)} D_{c(b)}}(t_b). \tag{3.1}$$

We form the likelihood of the observed data by summing over all possible states of internal nodes,

$$\Pr(\boldsymbol{D}) = \sum_{i_1=1}^m \cdots \sum_{i_{n-1}=1}^m \pi_{i_1} \prod_{b \in \mathcal{I}} p_{i_{p(b)} i_{c(b)}}(t_b) \prod_{b \in \mathcal{E}} p_{i_{p(b)} D_{c(b)}}(t_b). \tag{3.2}$$

Clearly, when data on the tips are not observed, the prior distribution of internal nodes becomes

$$\Pr(\boldsymbol{i}) = \pi_{i_1} \prod_{b \in \mathcal{I}} p_{i_{p(b)} i_{c(b)}}(t_b). \tag{3.3}$$

(b) *Posterior expectations of mapping summaries*

Consider an arbitrary branch b^* connecting parent internal node $p(b^*)$ to its child $c(b^*)$. First, we introduce restricted moments,

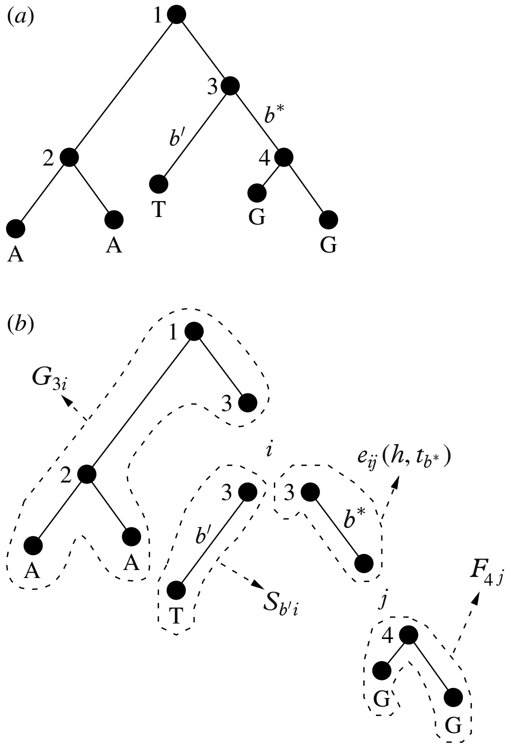
$$E\left[h(\{X_{b^*t}\})1_{\boldsymbol{D}}\right] = E\left[h(\{X_{b^*t}\}) \mid \boldsymbol{D}\right] \times \Pr(\boldsymbol{D}). \tag{3.4}$$

The expectation (3.4) integrates over all evolutionary mappings consistent with $\boldsymbol{D}$ on the tips of τ. Invoking the law of total expectation and the definition of conditional probability, we deduce

$$\begin{aligned} E\left[h(\{X_{b^*t}\})1_{\boldsymbol{D}}\right] &= E[h(\{X_{b^*t}\}) \mid \boldsymbol{D}] \times \Pr(\boldsymbol{D}) \\ &= \sum_{\boldsymbol{i}} E[h(\{X_{b^*t}\}) \mid \boldsymbol{i}, \boldsymbol{D}] \times \Pr(\boldsymbol{i}, \boldsymbol{D}) \\ &= \sum_{\boldsymbol{i}} E\left[h(\{X_{b^*t}\}) \mid i_{p(b^*)}, i_{c(b^*)}\right] \pi_{i_1} \prod_{b \in \mathcal{I}} p_{i_{p(b)} i_{c(b)}}(t_b) \\ &\quad \times \prod_{b \in \mathcal{E}} p_{i_{p(b)} D_{c(b)}}(t_b) \\ &= \sum_{\boldsymbol{i}} e_{i_{p(b^*)} i_{c(b^*)}}(h, t_{b^*}) \pi_{i_1} \prod_{b \in \mathcal{I} \setminus \{b^*\}} p_{i_{p(b)} i_{c(b)}}(t_b) \\ &\quad \times \prod_{b \in \mathcal{E} \setminus \{b^*\}} p_{i_{p(b)} D_{c(b)}}(t_b). \end{aligned} \tag{3.5}$$

The last expression in derivation (3.5) illustrates that in order to calculate the posterior restricted moment (3.4) along branch $b^* \in \mathcal{I}$, we merely need to replace finite-time transition probability $p_{i_{p(b^*)} i_{c(b^*)}}(t_{b^*})$ with local expectation $e_{i_{p(b^*)} i_{c(b^*)}}(h, t_{b^*})$ in the likelihood formula (3.2). Similarly, if $b^* \in \mathcal{E}$, we substitute $e_{i_{p(b^*)} D_{c(b^*)}}(h, t_{b^*})$ for $p_{i_{p(b^*)} D_{c(b^*)}}(t_{b^*})$ in (3.2). Given matrices $\boldsymbol{P}(t_b)$ for $b \neq b^*$ and $\boldsymbol{E}(h, t_{b^*})$, we can sum over internal node states using Felsenstein's pruning algorithm to arrive at the restricted mean $E[h(\{X_{b^*t}\})1_{\boldsymbol{D}}]$ and then divide this quantity by $\Pr(\boldsymbol{D})$ to obtain $E[h(\{X_{b^*t}\}) \mid \boldsymbol{D}]$.

This procedure is efficient for calculating the posterior expectations of mapping summaries for one branch of τ. However, in practice, we need to calculate the mapping expectations over many branches and consequently execute the computationally intensive



$$E[h(\{X_{b^*t}\})1_{\boldsymbol{D}}] = \sum_{i=1}^{m}\sum_{j=1}^{m} G_{3i}S_{b'i}e_{ij}(h, t_{b^*})F_{4j}$$

Figure 1. Sandwich formula illustration. (*a*) An example phylogenetic tree in which we label internal nodes numerically and two branches b^* and b'. We break this tree at nodes 3 and 4 into the subtrees shown in (*b*). Assuming that the trait states are i and j at nodes 3 and 4, respectively, we mark each subtree by the corresponding quantity needed for calculating the posterior expectation of a mapping summary on branch b^*.

pruning algorithm many times. A similar problem is encountered in maximum-likelihood phylogenetic estimation during differentiation of the likelihood with respect to branch lengths. An algorithm, reviewed by Schadt *et al.* (1998), allows for computationally efficient, repeated replacement of one of the finite-time transition probabilities with an arbitrary function of the corresponding branch length in equation (3.2). This algorithm finds informal use since the 1980s in pedigree analysis (Cannings *et al.* 1980) and is implemented in most maximum-likelihood phylogenetic reconstruction software such as PAUP (J. Huelsenbeck 2007, personal communication).

Let $\boldsymbol{F}_u = (F_{u1}, \ldots, F_{um})$ be the vector of forward, often called partial or fractional, likelihoods at node u. Element F_{ui} is the probability of the observed data at only the tips that descend from the node u, given that the state of u is i. If u is a tip, then we initialize partial likelihoods via $F_{ui} = 1_{\{i=D_u\}}$. In case of missing or ambiguous data, D_u denotes the subset of possible trait states, and forward likelihoods are set to $F_{ui} = 1_{\{i\in D_u\}}$. During the first, upward traversal of τ, we compute forward likelihoods for each internal node u using the recursion

$$F_{ui} = \left[\sum_{j=1}^{m} F_{c(b_1)j}p_{ij}(t_{b_1})\right] \times \left[\sum_{j=1}^{m} F_{c(b_2)j}p_{ij}(t_{b_2})\right], \qquad (3.6)$$

where b_1 and b_2 are indices of the branches descending from node u and $c(b_1)$ and $c(b_2)$ are the corresponding children of u. We denote the directional likelihoods in square brackets of equation (3.6) by S_{b_1i} and S_{b_2i} and record $\boldsymbol{S}_b = (S_{b1}, \ldots, S_{bm})$ together with $\boldsymbol{F}_u$. Finally, we define backward likelihoods $\boldsymbol{G}_u = (G_{u1}, \ldots, G_{um})$, where G_{ui} is the probability of observing state i at node u together with other tip states on the subtree of τ obtained by removing all lineages downstream of node u. A second, downward traversal of τ yields $\boldsymbol{G}_u$ given the precomputed $\boldsymbol{S}_b$.

For each branch b^*, we can sandwich $p_{ij}(t_{b^*})$ among the forward, directional and backward likelihoods and write

$$\Pr(\boldsymbol{D}) = \sum_{i=1}^{m}\sum_{j=1}^{m} G_{p(b^*)i}S_{b'i}p_{ij}(t_{b^*})F_{c(b^*)j}, \qquad (3.7)$$

where b' is the second branch descending from the parent node $p(b^*)$. Therefore, with $\boldsymbol{F}_u$, $\boldsymbol{S}_b$ and $\boldsymbol{G}_u$ precomputed for all nodes and branches of τ, we can replace $p_{ij}(t_{b^*})$ with any other quantity for any arbitrary branch b without repeatedly traversing τ. In particular, the posterior restricted moment for branch b^* can be expressed as

$$E[h(\{X_{b^*t}\})1_{\boldsymbol{D}}] = \sum_{i=1}^{m}\sum_{j=1}^{m} G_{p(b^*)i}S_{b'i}e_{ij}(h, t_{b^*})F_{c(b^*)j}. \qquad (3.8)$$

In figure 1, we use an example tree to illustrate the correspondence between each quantity in sandwich formula (3.8) and the part of the tree involved in this quantity computation. We summarize all steps that lead to the computation of the global mean $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$ in algorithm 1. Note that $\Pr(\boldsymbol{D})$, needed to transition between conditional and restricted expectations in formula (3.4), is computed with virtually no additional cost in step (iv) of the algorithm.

(c) *Pulley principle for evolutionary mappings*

Suppose that we are interested in a mean mapping summary $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$, obtained as a sum of local mapping summaries over *all branches* of the phylogenetic tree τ. We would like to know whether quantity $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$ changes when we move the root of τ to a different location.

Recall that reversibility of the Markov chain $\{X_t\}$ makes $\Pr(\boldsymbol{D})$ invariant to the root placement in τ if the root distribution $\boldsymbol{\pi}$ is the stationary distribution of $\{X_t\}$ (Felsenstein 1981). Felsenstein's pulley principle rests on the detailed balance condition

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t) \qquad (3.9)$$

and Chapman–Kolmogorov relationship

$$p_{ij}(t_1 + t_2) = \sum_{k=1}^{m} p_{ik}(t_1)p_{kj}(t_2). \qquad (3.10)$$

Applying both formulae (3.9) and (3.10) once in equation (3.2) allows one to move the root to any position along the two root branches without changing $\Pr(\boldsymbol{D})$. Therefore, $\Pr(\boldsymbol{D})$ is invariant to moving the root to any position on any branch of τ since we can repeatedly apply the detailed balance condition and the Chapman–Kolmogorov equation.

Invariance of $\Pr(\boldsymbol{D})$ to root placement together with formula (3.4) suggests that root position invariance of

Algorithm 1. Calculating posterior expectations $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$.

```
(i)   obtain an eigen decomposition of the infinitesimal generator $\boldsymbol{Q}$
(ii)  use this decomposition to compute $\boldsymbol{P}(t_b)$ for each branch $b$ of $\tau$
(iii) employing the same eigen decomposition, compute $\boldsymbol{E}(h, t_{b^*})$ for each branch $b^*$ in the set of interest $\Omega$ using either equation (2.4) or equation (2.12)
(iv)  traverse $\tau$ once and calculate $\boldsymbol{F}_u$ and $\boldsymbol{S}_b$ for each node $u$ and each branch $b$. Compute data likelihood $\Pr(\boldsymbol{D})$ as the dot product of $\boldsymbol{F}_{\text{root}}$ and root distribution $\pi$
(v)   traverse $\tau$ the second time and calculate backward likelihoods $\boldsymbol{G}_u$ for all nodes $u$
(vi)  for each $b^* \in \Omega$, apply equation (3.8) to obtain $E[h(\{X_{b^*t}\})1_{\boldsymbol{D}}]$
(vii) return $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}] = 1/\Pr(\boldsymbol{D})\sum_{b^* \in \Omega} E[h(\{X_{b^*t}\})1_{\boldsymbol{D}}]$
```

conditional expectations $E[H(\boldsymbol{M}_\theta)\,|\,\boldsymbol{D}]$ holds if and only if invariance of joint expectations $E[H(\boldsymbol{M}_\theta)1_{\boldsymbol{D}}]$ is satisfied. Consider a two-tip phylogeny with branches of length t_1 and t_2 leading to observed trait states D_1 and D_2, respectively. According to formulae (1.2) and (3.5), we may expect that

$$\begin{aligned} E[H(\boldsymbol{M}_\theta)1_{\boldsymbol{D}}] &= \sum_{k=1}^{m} \pi_k e_{kD_1}(h, t_1) p_{kD_2}(t_2) \\ &\quad + \sum_{k=1}^{m} \pi_k e_{kD_2}(h, t_2) p_{kD_1}(t_1) \\ &= \sum_{k=1}^{m} \pi_{D_1} e_{D_1 k}(h, t_1) p_{kD_2}(t_2) \\ &\quad + \sum_{k=1}^{m} \pi_{D_1} e_{kD_2}(h, t_2) p_{D_1 k}(t_1) \\ &= \pi_{D_1} e_{D_1 D_2}(h, t_1 + t_2), \end{aligned} \tag{3.11}$$

depends only on the sum $t_1 + t_2$. Therefore, we can move the root anywhere on this phylogeny without altering the expectations if $\{X_t\}$ is reversible. It is easy to see that repeated application of derivation (3.11) readily allows for extension of the root invariance principle to n-tip phylogenies.

In derivation (3.11), we use identities

$$e_{ij}(h, t_1 + t_2) = \sum_{k=1}^{m} \left[e_{ik}(h, t_1) p_{kj}(t_2) + e_{kj}(h, t_2) p_{ik}(t_1) \right] \tag{3.12}$$

and

$$\pi_i e_{ij}(h, t) = \pi_j e_{ji}(h, t). \tag{3.13}$$

Equation (3.12) splits computation of the expected summary on interval [0, $t_1 + t_2$) into calculations bound to intervals $[0, t_1)$ and $[t_1, t_1 + t_2)$ with the help of the total expectation law and Markov property. This derivation parallels that of Chapman–Kolmogorov equation (3.10). Identity (3.13) requires more care as it *does not hold* for all choices of function h. Using equation (2.12), detailed balance condition (3.13) holds for $h_2 = R_{\boldsymbol{w}}$. However, equation (2.3) suggests that we only can guarantee the detailed balanced condition (3.13) for $h_1 = N_{\mathcal{L}}$ when $(i, j) \in \mathcal{L}$ if and only if $(j, i) \in \mathcal{L}$.

(d) *Prior expectations of mapping summaries*

In many applications of stochastic mapping, one wishes to compare the prior to posterior expectations of summaries (Nielsen 2002). In this section, we derive the formulae necessary for computing prior expectations. Similar to our derivation of posterior expectations, we begin by considering an arbitrary branch b^* and use the law of total expectation to arrive at

$$E[h(\{X_{b^*t}\})] = \begin{cases} \sum_{\boldsymbol{i}} e_{i_{p(b^*)} i_{c(b^*)}}(h, t_{b^*}) \pi_{i_1} \prod_{b \in \mathcal{I} \setminus \{b^*\}} p_{i_{p(b)} i_{c(b)}}(t_b) & \text{if } b^* \in \mathcal{I} \\ \sum_{\boldsymbol{i}} e_{i_{p(b^*)}}(h, t_{b^*}) \pi_{i_1} \prod_{b \in \mathcal{I}} p_{i_{p(b)} i_{c(b)}}(t_b) & \text{if } b^* \in \mathcal{E}, \end{cases} \tag{3.14}$$

where $e_i(h, t) = \sum_{j=1}^{m} e_{ij}(h, t)$ is the marginal local expectation of the mapping summary.

Identity $\boldsymbol{P}(t)\mathbf{1} = \mathbf{1}$ allows us to eliminate summation over some internal node states in formula (3.14) and consider only those internal nodes that lie on the path connecting the root of τ and $c(b^*)$. If π is the stationary distribution of $\{X_t\}$, then formula (3.14) together with identities $\pi^T \boldsymbol{P}(t) = \pi^T$ and $\boldsymbol{P}(t)\mathbf{1} = \mathbf{1}$ simplifies prior local expectations even further,

$$\begin{aligned} E[h(\{X_{b^*t}\})] &= \sum_i \sum_j \pi_i e_{ij}(h, t_{b^*}) \\ &= \pi^T \boldsymbol{E}(h, t_{b^*}) \mathbf{1} = \begin{cases} \pi^T \boldsymbol{Q}_{\mathcal{L}} \mathbf{1} t_{b^*} & \text{if } h = N_{\mathcal{L}} \\ \displaystyle\sum_{i=1}^{m} \pi_i w_i t_{b^*} & \text{if } h = R_{\boldsymbol{w}} \end{cases} \end{aligned} \tag{3.15}$$

The fact that prior local expectations at stationarity compute with virtually no additional burden has immediate practical implications. In the context of the posterior predictive model checking, researchers often need to simulate L independent and identically distributed (iid) realizations $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_L$ of data at the tips of τ and then calculate the summary-based discrepancy measure $1/L \sum_{l=1}^{L} E[h(\{X_t\})\,|\,\boldsymbol{D}_l]$ (Nielsen 2002). Since we know the 'true' model under simulation, we can approximate this discrepancy measure with the prior expectation

$$\begin{aligned} \frac{1}{L} \sum_{l=1}^{L} E[h(\{X_t\})\,|\,\boldsymbol{D}_l] &\approx \sum_{\boldsymbol{D}^*} E[h(\{X_t\})\,|\,\boldsymbol{D}^*] \Pr(\boldsymbol{D}^*) \\ &= E[h(\{X_t\})], \end{aligned} \tag{3.16}$$

where $\boldsymbol{D}^*$ ranges over all possible trait values at the tips of τ. In molecular evolution applications, L is of the order of 10^3–10^5, and hence approximation (3.16) should work well.

(e) *Comparison with simulation-based stochastic mapping*

We comment earlier that the Monte Carlo algorithm for stochastic mapping is an alternative and a very popular way to compute expectations of mapping summaries. This algorithm consists of two major steps. In the first step, the tree is traversed once to compute $\boldsymbol{F}_u$ for each node. In the second step, internal node states $\boldsymbol{i}$ are simulated conditional on $\boldsymbol{D}$ (Pagel 1999). Then, conditional on $\boldsymbol{i}$, one simulates CTMC trajectories on each branch of the tree and computes summaries of

Table 1. Efficiency and accuracy of stochastic mapping. (For each number of iterations, we report the median number of rejected CTMC trajectories over the entire phylogenetic tree per iteration and the sum of absolute errors (SAE) of simulation-based estimates of the mean number of synonymous mutations along branches of the phylogenetic tree.)

	slow evolving site		fast evolving site	
iterations	rejections/ iteration	SAE	rejections/ iteration	SAE
100	100	0.0598	38 845	0.4624
500	105	0.0255	39 247	0.3319
1000	102	0.0259	42 075	0.2905
10 000	106	0.0205	40 805	0.2809

interest. This second step is repeated N times producing a Monte Carlo sample of mapping summaries whose averages approximate the branch-specific expectations.

The running time of the algorithm depends on N and the computational efficiency of generating CTMC trajectories. The number of samples N, required for accurate estimation, varies with $\boldsymbol{Q}$ and $\boldsymbol{T}$. Unfortunately, this aspect of stochastic mapping is largely ignored in the literature and by practitioners. Although more than one way exists to simulate trajectories conditional on starting and ending states (Rodrigue *et al.* 2008), rejection sampling is the most common (Nielsen 2002). Assessing the efficiency of rejection sampling is complicated, because the efficiency depends not only on the choice of CTMC parameters, but also on the observed data patterns $\boldsymbol{D}$. We illustrate these difficulties using a CTMC on a state space of 64 codon triplets. We take two sites from an alignment of 129 HIV sequences that we discuss later in one of our examples. We call the first site slow evolving as it obtains only three different codons. The second, fast evolving site emits nine different codons. We take one-parameter slice of our Markov chain Monte Carlo (MCMC) sample and run rejection sampling to estimate the expected number of mutations for all branches of τ. For each trial, we record the total number of rejections on all branches of τ and the absolute errors of the estimated mean number of synonymous mutations summed over all branches. We summarize the results of this experiment in table 1. We see a 400× increase in the number of rejections required to simulate trajectories conditional on the fast site compared to equivalent simulations based on the slow site. The Monte Carlo error decreases as the number of Monte Carlo iterations increases, but not as fast as one would hope.

In summary, simulation-based stochastic mapping requires simulation of CTMC trajectories; this is not a trivial computational task. Assessing accuracy of methods is cumbersome and difficult to automate. Our algorithm 1 replaces both simulation components from stochastic mapping calculations and therefore should be a preferred way of calculating expectations of mapping summaries.

4. EXAMPLES

(a) *Detecting coevolution via dwelling times*

In this section, we reformulate a previously developed simulation-based method for detection of correlated evolution (Huelsenbeck *et al.* 2003) in terms of a Markov reward process. We consider two primate evolutionary traits, oestrus advertisement (EA) and multimale mating system (MS), analysed by Pagel & Meade (2006). These authors first use cytochrome *b* molecular sequences to estimate the posterior distribution of the phylogenetic relationship among 60 Old World monkeys and ape species. Using 500 MCMC samples from the posterior distribution of phylogenetic trees, Pagel and Meade run another MCMC chain, this time with a reversible jump component that explores a number of CTMC models involving EA and MS traits and assess the models' posterior probabilities to learn about EA/MS coevolution. The authors find support in favour of a hypothesis stating that EA presence correlates with MS presence. The trait data are shown in figure 2 together with a phylogenetic tree, randomly chosen from the posterior sample. While this is not the case for Pagel & Meade (2006), reversible jump MCMC for model selection can be difficult to implement, especially as the number of trait states grows. Methods that simply require data fitting under the null model are warranted. Consequentially, we revisit this dataset and apply posterior predictive model diagnostics to test the hypothesis of independent evolution between EA and MS traits.

Our null model assumes that EA and MS evolve independently as 2 two-state Markov chains $X_t^{(1)}$, $X_t^{(2)} \in \{0, 1\}$, where 0 and 1, respectively, stand for trait absence and presence. Let the infinitesimal generators of the EA and MS CTMCs be

$$\boldsymbol{Q}_{\mathrm{EA}} = \begin{pmatrix} -\alpha_0 & \alpha_0 \\ \alpha_1 & -\alpha_1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{Q}_{\mathrm{MS}} = \begin{pmatrix} -\beta_0 & \beta_0 \\ \beta_1 & -\beta_1 \end{pmatrix}. \tag{4.1}$$

We form a product Markov chain $Y_t = \left(X_t^{(1)}, X_t^{(2)}\right)$ on the state space $\{(0,0),(0,1),(1,0),(1,1)\}$ that keeps track of presence/absence of the two traits simultaneously assuming that they evolve independently. The generator of the product chain is obtained via the Kronecker sum ($\oplus$),

$$\boldsymbol{\Phi} = \boldsymbol{Q}_{\mathrm{MS}} \oplus \boldsymbol{Q}_{\mathrm{EA}} = \begin{pmatrix} -(\alpha_0+\beta_0) & \beta_0 & \alpha_0 & 0 \\ \beta_1 & -(\alpha_0+\beta_1) & 0 & \alpha_0 \\ \alpha_1 & 0 & -(\alpha_1+\beta_0) & \beta_0 \\ 0 & \alpha_1 & \beta_1 & -(\alpha_1+\beta_1) \end{pmatrix}. \tag{4.2}$$

The Kronecker sum representation extends to general finite state-space Markov chains and to an arbitrary number of independently evolving traits (Neuts 1995). Computationally, this representation is advantageous, because eigenvalues and eigenvectors of a potentially high-dimensional product chain generator derive analytically from eigenvalues and eigenvectors of low-dimensional individual generators (Laub 2004).

To test the independent evolution model fit via posterior predictive diagnostics, we need a discrepancy measure (Meng 1994). Following Huelsenbeck *et al.* (2003), we employ mean dwelling times to form a

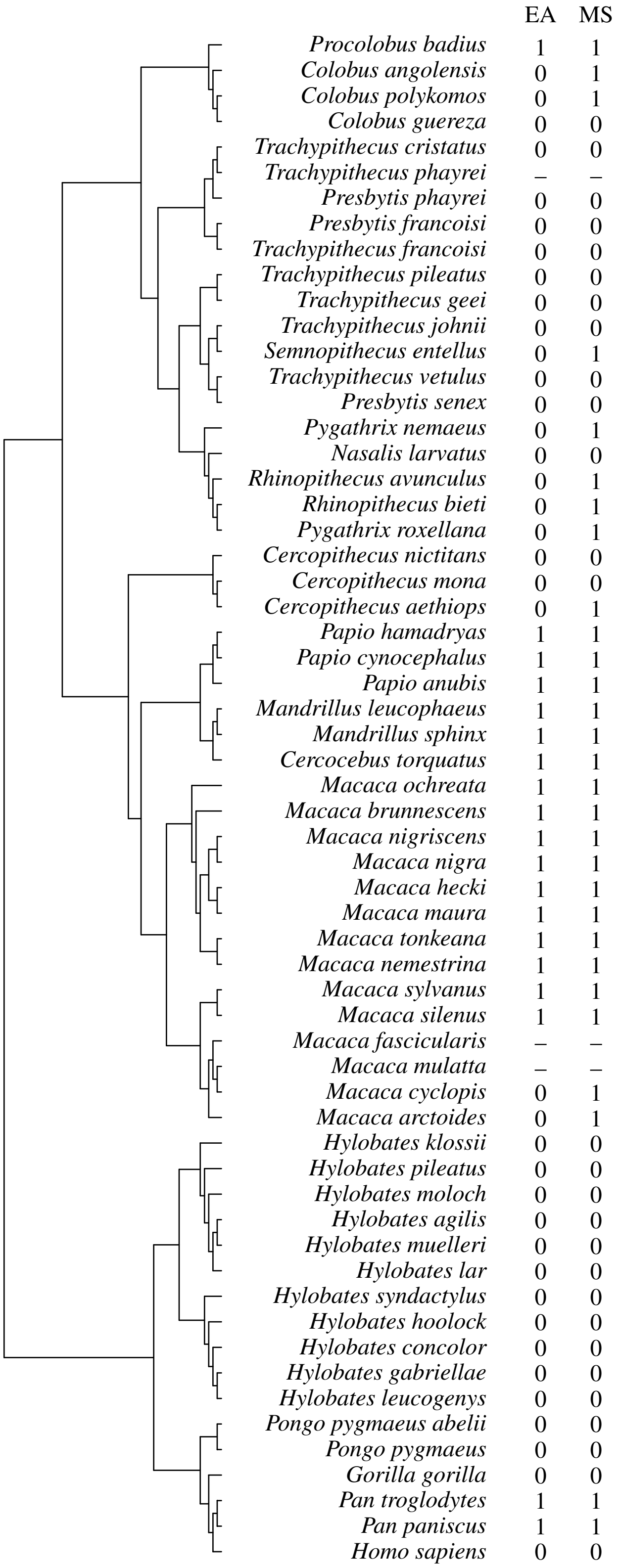

EA MS
Procolobus badius 1 1
Colobus angolensis 0 1
Colobus polykomos 0 1
Colobus guereza 0 0
Trachypithecus cristatus 0 0
Trachypithecus phayrei – –
Presbytis phayrei 0 0
Presbytis francoisi 0 0
Trachypithecus francoisi 0 0
Trachypithecus pileatus 0 0
Trachypithecus geei 0 0
Trachypithecus johnii 0 0
Semnopithecus entellus 0 1
Trachypithecus vetulus 0 0
Presbytis senex 0 0
Pygathrix nemaeus 0 1
Nasalis larvatus 0 0
Rhinopithecus avunculus 0 1
Rhinopithecus bieti 0 1
Pygathrix roxellana 0 1
Cercopithecus nictitans 0 0
Cercopithecus mona 0 0
Cercopithecus aethiops 0 1
Papio hamadryas 1 1
Papio cynocephalus 1 1
Papio anubis 1 1
Mandrillus leucophaeus 1 1
Mandrillus sphinx 1 1
Cercocebus torquatus 1 1
Macaca ochreata 1 1
Macaca brunnescens 1 1
Macaca nigriscens 1 1
Macaca nigra 1 1
Macaca hecki 1 1
Macaca maura 1 1
Macaca tonkeana 1 1
Macaca nemestrina 1 1
Macaca sylvanus 1 1
Macaca silenus 1 1
Macaca fascicularis – –
Macaca mulatta – –
Macaca cyclopis 0 1
Macaca arctoides 0 1
Hylobates klossii 0 0
Hylobates pileatus 0 0
Hylobates moloch 0 0
Hylobates agilis 0 0
Hylobates muelleri 0 0
Hylobates lar 0 0
Hylobates syndactylus 0 0
Hylobates hoolock 0 0
Hylobates concolor 0 0
Hylobates gabriellae 0 0
Hylobates leucogenys 0 0
Pongo pygmaeus abelii 0 0
Pongo pygmaeus 0 0
Gorilla gorilla 0 0
Pan troglodytes 1 1
Pan paniscus 1 1
Homo sapiens 0 0

Figure 2. Primate trait data. We plot a phylogenetic tree, randomly chosen from the posterior sample, of 60 primate species. Branches of the sampled tree are not drawn to scale, nor is the tree ultrametric. Taxa names and trait values (‘0’, absence; ‘1’, presence; ‘–’, missing) for oestrus advertisement (EA) and multimale mating system (MS) are depicted at the tips of the tree.

discrepancy measure. Let $Z=\sum_{b=1}^{B} t_b$ be the tree length of τ. We define the mean dwelling times $Z_i^{(1)}$ and $Z_i^{(2)}$ of traits $X_t^{(1)}$ and $X_t^{(2)}$ in state i, and the mean dwelling time Z_{ij} of the product chain Y_t in state (i,j) for $i,j=0,1$. More formally, we set

$$Z_i^{(1)}=E\left[R_{\boldsymbol{w}_i}\,|\,\boldsymbol{D}^{(1)}\right],\qquad Z_i^{(2)}=E\left[R_{\boldsymbol{w}_i}\,|\,\boldsymbol{D}^{(2)}\right]\quad\text{and}$$
$$Z_{ij}=E\left[R_{\boldsymbol{w}_{ij}}\,|\,\boldsymbol{D}^{(1)},\boldsymbol{D}^{(2)}\right],\tag{4.3}$$

where $\boldsymbol{w}_1=(1,0)$, $\boldsymbol{w}_2=(0,1)$, $\boldsymbol{w}_{00}=(1,0,0,0)$, $\boldsymbol{w}_{01}=(0,1,0,0)$, $\boldsymbol{w}_{10}=(0,0,1,0)$, $\boldsymbol{w}_{11}=(0,0,0,1)$ and $\boldsymbol{D}^{(1)}$, $\boldsymbol{D}^{(2)}$ are observations of the two traits on the tips of τ.

Using the dwelling times, we define ‘expected’ ε_{ij} and ‘observed’ η_{ij} fractions of time the two traits spend in states i and j,

$$\varepsilon_{ij}=\frac{Z_i^{(1)}}{Z}\times\frac{Z_j^{(2)}}{Z}\quad\text{and}\quad\eta_{ij}=\frac{Z_{ij}}{Z}.\tag{4.4}$$

We use quotation marks because both quantities are not observed. To quantify the deviation from the null hypothesis of independence seen in the data, we introduce the discrepancy measure,

$$\Delta(\boldsymbol{Q}_{\mathrm{EA}},\boldsymbol{Q}_{\mathrm{MS}},\boldsymbol{D})=\sum_{i=0}^{1}\sum_{j=0}^{1}(\varepsilon_{ij}-\eta_{ij})^2.\tag{4.5}$$

This measure returns small values under the null and implicitly depends on τ and branch lengths $\boldsymbol{T}$. We account for this dependence and phylogenetic uncertainty by averaging our results over a finite sample from the posterior distribution of τ and $\boldsymbol{T}$, obtained from the molecular sequence data.

We use the software package BayesTraits to accomplish this averaging and to produce a MCMC sample from the posterior distribution of $\boldsymbol{Q}_{\mathrm{EA}}$ and $\boldsymbol{Q}_{\mathrm{MS}}$, assuming the null model of independent evolution of the two traits (Pagel *et al.* 2004). Each iteration sample from the output of BayesTraits consists of $(\tau,\boldsymbol{T},\alpha_0,\alpha_1,\beta_0,\beta_1)$ drawn from their posterior distribution. Given these model parameters, we generate a new dataset $\boldsymbol{D}^{\mathrm{rep}}$ and compute the observed $\Delta(\boldsymbol{Q}_{\mathrm{EA}},\boldsymbol{Q}_{\mathrm{MS}},\boldsymbol{D})$ and predicted $\Delta(\boldsymbol{Q}_{\mathrm{EA}},\boldsymbol{Q}_{\mathrm{MS}},\boldsymbol{D}^{\mathrm{rep}})$ discrepancies for each iteration. We then compare their marginal distributions by plotting their corresponding histograms (figure 3). In figure 3, we also plot the observed against predicted discrepancies to display the correlation between these two random variables. The apparent disagreement between observed and predicted discrepancies is a manifestation of poor fit of the independent model of evolution. The observed discrepancy consistently exceeds the predicted quantity. To illustrate the performance of predictive diagnostics when the independent model fits data well, we simulate trait data under this model along one of the 500 *a posteriori* supported phylogenetic trees. Figure 3*c*,*d* depicts the result of the posterior model diagnostics applied to the simulated data. In contrast to the observed primate data, the simulated data do not exhibit disagreement between the observed and predicted discrepancies.

Disagreement between the observed and predicted discrepancies can be quantified using a tail probability, called a posterior predictive p value,

$$\mathrm{ppp}=\Pr\left(\Delta(\boldsymbol{Q}_{\mathrm{EA}},\boldsymbol{Q}_{\mathrm{MS}},\boldsymbol{D}^{\mathrm{rep}})>\Delta(\boldsymbol{Q}_{\mathrm{EA}},\boldsymbol{Q}_{\mathrm{MS}},\boldsymbol{D})\,|\,\boldsymbol{D},H_0\right),\tag{4.6}$$

where the tail probability is taken over the posterior distribution of the independent model. In practice,

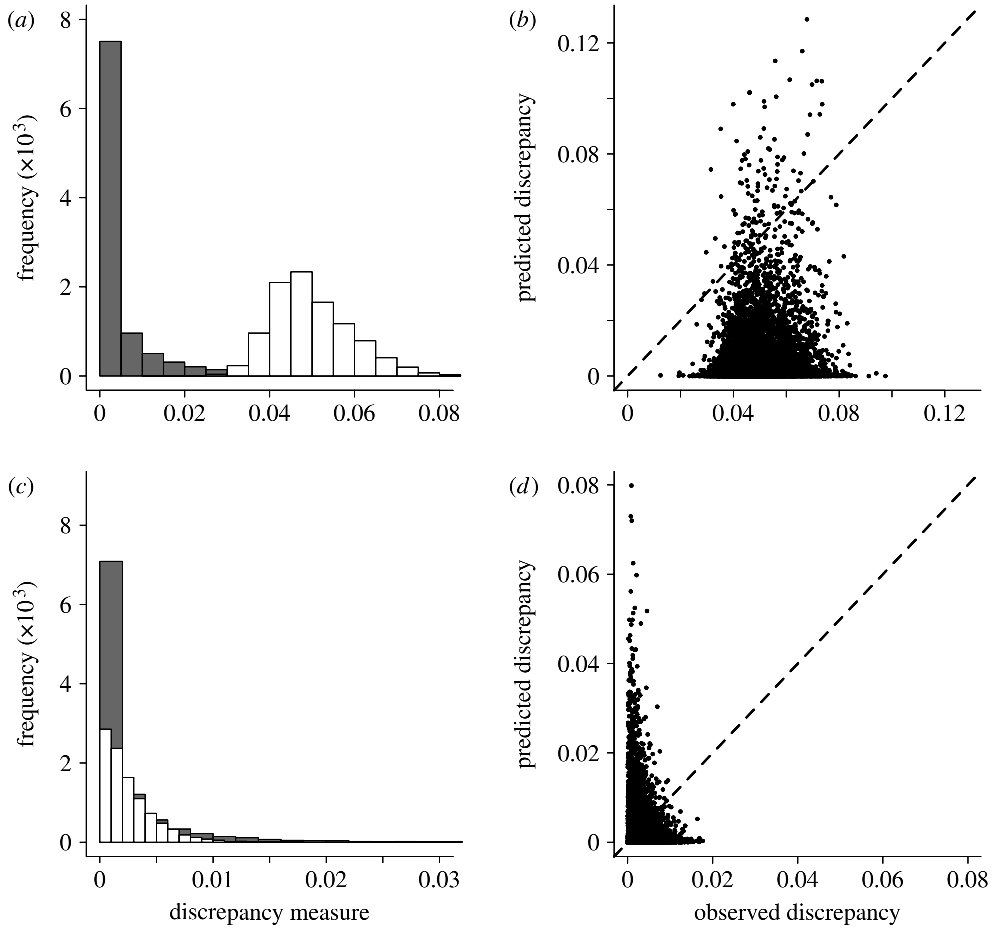


Figure 3. Testing coevolution. (*a*,*c*) Plots depicting observed (white bars) and predicted (grey bars) distributions of the discrepancy measure for the (*a*,*b*) primate and (*c*,*d*) simulated data. (*b*,*d*) The scatter plots of these distributions.

given N MCMC iterations, one estimates posterior predictive p values via

$$\text{ppp} \approx \frac{1}{N}\sum_{g=1}^{N} 1_{\left\{\Delta\left(\boldsymbol{Q}_{\text{EA}}^{(g)}, \boldsymbol{Q}_{\text{MS}}^{(g)}, \boldsymbol{D}^{\text{rep},g}\right) > \Delta\left(\boldsymbol{Q}_{\text{EA}}^{(g)}, \boldsymbol{Q}_{\text{MS}}^{(g)}, \boldsymbol{D}\right)\right\}}, \tag{4.7}$$

where $\boldsymbol{Q}_{\text{EA}}^{(g)}, \boldsymbol{Q}_{\text{MS}}^{(g)}$ are the parameter values realized at iteration g, and $\boldsymbol{D}^{\text{rep},g}$ is a dataset simulated using these parameter values. Following this recipe, we estimate ppp for the primate data and the artificial data. The disagreement between the 'observed' and 'predicted' discrepancies is reflected in a low ppp = 0.0128. By contrast, the ppp = 0.3139, for the simulated data, supports agreement between the observed and predicted distributions of Δ.

(b) *Mapping synonymous and non-synonymous mutations*

In this section, we consider the important task of mapping synonymous and non-synonymous mutations onto branches of a phylogenetic tree. Our point of departure is a recent ambitious analysis of HIV intrahost evolution by Lemey *et al.* (2007), who use sequence data originally reported by Shankarappa *et al.* (1999). The authors attempt to estimate branch-specific synonymous and non-synonymous mutation rates and then project these measurements onto a time axis. This projection enables one to relate the time evolution of selection processes with clinical covariates. Lemey *et al.* (2007) find fitting codon models computationally prohibitive in this case. Instead, they first fit a DNA model to the intrahost HIV sequences, obtain a posterior sample of phylogenies with branch lengths, and then use these phylogenies to fit a codon model to the same DNA sequence alignment. Instead of fitting two different models to the data, we propose to use just a DNA model and exploit mapping summaries to predict synonymous and non-synonymous mutation rates.

Suppose we observe a multiple DNA sequence alignment $\boldsymbol{D} = (D_1, \dots, D_L)$ of a protein-coding region with L sites and that all $C = L/3$ codons are aligned to each other such that the coding region starts at site 1 of $\boldsymbol{D}$ (in other words, there is no frameshift). Following the codon partitioning model of Yang (1996), we assume that sites corresponding to all first codon positions, $\boldsymbol{D}_1, \boldsymbol{D}_4, \dots, \boldsymbol{D}_{L-2}$, evolve according to a standard HKY model with generator $\boldsymbol{Q}_{\text{HKY}}(\kappa_1, \pi_1)$, where κ_1 is the transition–transversion ratio and π_1 is the stationary distribution, both just for the first codon position appropriately constrained. Similarly, we define CTMC generators $\boldsymbol{Q}_{\text{HKY}}(\kappa_2, \pi_2)$ and $\boldsymbol{Q}_{\text{HKY}}(\kappa_3, \pi_3)$ for the other two codon positions with independent parameters. Assuming that all L nucleotide sites in $\boldsymbol{D}$ evolve independently together with the three codon position HKY models induces a product Markov chain model on the space of codons $(AAA, AAG, \dots, TTT)$, where codons are arranged in lexicographic order with respect to our nucleotide order $A < G < C < T$, the generator of this product CTMC is

$$\boldsymbol{Q}_{\text{codon}} = \boldsymbol{Q}_{\text{HKY}}(\kappa_1, \pi_1) \oplus \boldsymbol{Q}_{\text{HKY}}(\kappa_2, \pi_2) \oplus \boldsymbol{Q}_{\text{HKY}}(\kappa_3, \pi_3). \tag{4.8}$$

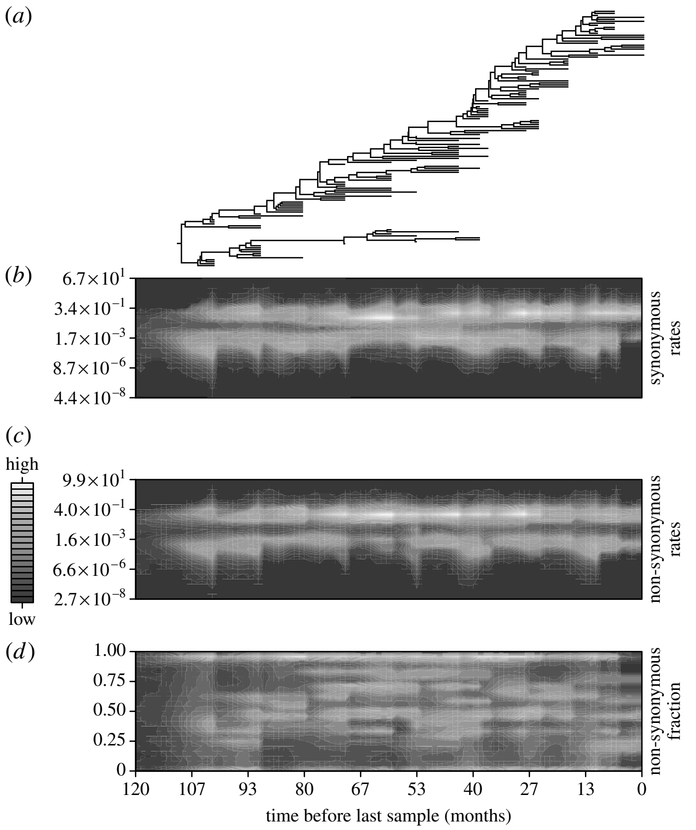


Figure 4. Time evolution of synonymous and non-synonymous rates. (*a*) A representative phylogeny of 129 intrahost HIV sequences. The three heat maps depict the marginal posterior densities of the (*b*) synonymous and (*c*) non-synonymous rates, and (*d*) the proportion of non-synonymous mutations over time.

With this Markov chain on the codon space, we define a labelling $\mathcal{L}(s)$ that contains all possible pairs of codons that translate into the same amino acid. All other codon pairs are collected into a labelling set $\mathcal{L}(n)$. Clearly, transitions between elements of $\mathcal{L}(s)$ constitute synonymous mutations, and non-synonymous mutations are represented by transitions between elements of $\mathcal{L}(n)$. In this manner, counting processes map synonymous and non-synonymous mutations onto specific branches of τ. We consider HIV sequences from patient 1 of Shankarappa *et al.* (1999) and approximate the posterior distribution of our DNA model parameters $\Pr(\boldsymbol{Q}_{\text{codon}}, \tau, \boldsymbol{T} \mid \boldsymbol{D})$ using MCMC sampling implemented in the software package BEAST (Drummond & Rambaut 2007). The serially sampled HIV sequences permit us to estimate the branch lengths $\boldsymbol{T}$ in units of clock time, months in this case. For each saved MCMC sample, we compute branch-specific rates of synonymous and non-synonymous mutations,

$$r_b(s) = \frac{\frac{1}{C}\sum_{c=1}^{C} E\left[N_{\mathcal{L}(s)}(t_b) \mid \boldsymbol{D}_{c:c+2}\right]}{t_b} \quad \text{and}$$

$$r_b(n) = \frac{\frac{1}{C}\sum_{c=1}^{C} E\left[N_{\mathcal{L}(n)}(t_b) \mid \boldsymbol{D}_{c:c+2}\right]}{t_b}, \tag{4.9}$$

where we denote data at codon site c by $\boldsymbol{D}_{c:c+2}$. We also record the fraction $r_b(n)/(r_b(s)+r_b(n))$ of non-synonymous mutations. Similar to Lemey *et al.* (2007), we summarize these measurements by projecting them on the time axis. To this end, we form a finite time grid and produce a density profile of the synonymous and non-synonymous rates, and of the non-synonymous mutation fractions for each time interval between grid points (figure 4). Both synonymous and non-synonymous rate density profiles are consistently bimodal across time. Interestingly, the modes also stay appreciably constant. The density profile of the non-synonymous mutation fractions is multimodal and fairly complex. There are a considerable number of branches that exhibit strong negative $((r_b(n))/(r_b(s)+r_b(n)) \sim 0)$ and positive $((r_b(n))/(r_b(s)+r_b(n)) \sim 1)$ selection. In general, when comparing branches that occur earlier in the tree to those that occur later, the non-synonymous mutation fraction has first a modest upward trend through time and then descends to lower values, consistent with other patterns of evolutionary diversity reported by Shankarappa *et al.* (1999).

Intrigued by the multimodality observed in figure 4, we investigate this issue further. Lemey *et al.* (2007) consider several branch categories in their analysis, e.g. internal and terminal. We decide to test whether differences exist between selection forces acting on internal and terminal branches of τ. We define the fractions of non-synonymous mutations on internal and terminal branches as

$$\rho_{\mathcal{E}}(\boldsymbol{Q}_{\text{codon}}, \tau, \boldsymbol{T}, \boldsymbol{D}) = \frac{\sum_{b\in\mathcal{E}} \sum_{c=1}^{C} E\left[N_{\mathcal{L}(n)}(t_b) \mid \boldsymbol{D}_{c:c+2}\right]}{\sum_{b\in\mathcal{E}} \sum_{c=1}^{C} \left\{E\left[N_{\mathcal{L}(s)}(t_b) \mid \boldsymbol{D}_{c:c+2}\right] + E\left[N_{\mathcal{L}(n)}(t_b) \mid \boldsymbol{D}_{c:c+2}\right]\right\}} \tag{4.10}$$

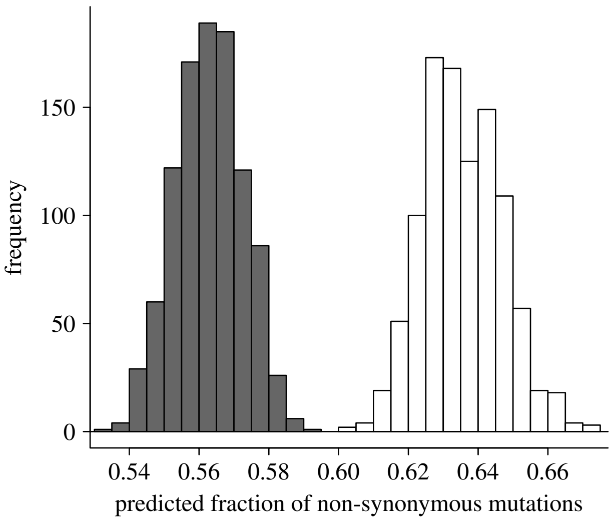


Figure 5. Bimodality of the fraction of non-synonymous mutations. We plot the predicted fraction of non-synonymous mutations computed for terminal (grey bars) and internal (white bars) branches.

and

$$\rho_{\mathcal{I}}(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}) = \frac{\sum_{b\in\mathcal{I}}\sum_{c=1}^{C} E\left[N_{\mathcal{L}(n)}(t_b)\,|\,\boldsymbol{D}_{c:c+2}\right]}{\sum_{b\in\mathcal{I}}\sum_{c=1}^{C}\left\{E\left[N_{\mathcal{L}(s)}(t_b)\,|\,\boldsymbol{D}_{c:c+2}\right]+E\left[N_{\mathcal{L}(n)}(t_b)\,|\,\boldsymbol{D}_{c:c+2}\right]\right\}}. \tag{4.11}$$

We plot their posterior histograms in figure 5. These histograms do not overlap, suggesting different fractions of non-synonymous mutations for internal and external branches. To test this hypothesis more formally, we form a discrepancy measure

$$\Delta(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}) = \rho_{\mathcal{E}}(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}) - \rho_{\mathcal{I}}(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}). \tag{4.12}$$

As in our previous example, we compare the observed discrepancy $\Delta(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D})$ with the expected discrepancy $\Delta(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}^{\text{rep}})$, where $\boldsymbol{D}^{\text{rep}}$ is a multiple sequence alignment simulated under the codon partitioning model with parameters $\boldsymbol{Q}_{\text{codon}}$, τ and $\boldsymbol{T}$. Evoking approximation (3.16) and recalling that our model assumes the same substitution rates for each branch of τ, we deduce that

$$\Delta(\boldsymbol{Q}_{\text{codon}},\tau,\boldsymbol{T},\boldsymbol{D}^{\text{rep}}) \approx 0, \tag{4.13}$$

for all parameter values $\boldsymbol{Q}_{\text{codon}}$, $\boldsymbol{T}$ and τ and replicated data $\boldsymbol{D}^{\text{rep}}$. Plugging in our new discrepancy measures into equation (4.7), we find that $\text{ppp} < 0.001$. Therefore, our posterior predictive test suggests that there is a significant heterogeneity in selective pressure among the branches of τ.

5. DISCUSSION

In this paper, we develop a computationally efficient framework for mapping evolutionary trajectories onto phylogenies. Although we aim to keep this mathematical framework fairly general, our main interest with evolutionary mappings lies in computing the mean number of labelled trait transitions and the mean evolutionary reward that depends linearly on the time a trait occupies each of its states. These two mapping summaries have been the most promising building blocks for constructing statistical tests.

We build upon our earlier work involving single branch calculations for Markov-induced counting processes (Minin & Suchard 2008). In our extension, we introduce single branch calculations for evolutionary reward processes and devise algorithms to extend single branch calculations to mapping expectations of counting and reward processes onto branches across an entire phylogeny. Our main result generalizes Felsenstein's pruning algorithm that forms the workhorse of modern phylogenetic computation. The generalized pruning algorithm warrants two comments about its efficiency for performing simulation-free stochastic mapping. First, when the infinitesimal rates are unknown, a traditionally slow component of phylogenetic inference is the eigen decomposition of their matrix. Fortunately, this decomposition finds immediate reuse in our algorithm to calculate posterior expectations of mappings. Second, the algorithm requires only two traversals of the phylogenetic tree, and is therefore at most two times slower than the standard likelihood evaluation algorithm. In practice, we find our algorithm approximately 1.5 times slower than the likelihood calculation. We achieve this advantage because during the second traversal *n* terminal branches are not visited. Finally, the Felsenstein's algorithm analogy yields a pulley principle for stochastic mapping and reduction in computation for prior expectations.

Our examples demonstrate how our novel algorithm facilitates phylogenetic exploratory analysis and hypothesis testing. First, we use simulation-free stochastic mapping of occupancy times to re-implement Huelsenbeck *et al.*'s (2003) posterior predictive test of independent evolution. In our second example, we attempt to recover synonymous and non-synonymous mutation rates without resorting to codon models and instead use an independent codon partitioning model. We overcome this gross model misspecification with stochastic mapping, find intriguing multimodality of synonymous and non-synonymous rates, and use a posterior predictive model check to test differences in selection pressures between terminal and internal branches. We stress that our predictions are only as good as the model we use. For example, the terminal/internal branch differences may be due to a general bad fit of our purposely misspecified model to the intrahost HIV data. However, we find this scenario unlikely in light of the recent demonstration of the excellent performance of codon partitioning models during analyses of protein coding regions (Shapiro *et al.* 2006).

Our examples illustrate importance of hypothesis testing in statistical phylogenetics. In recent years, it has become clear that an evolutionary analysis almost never ends with tree estimation. Importantly, phylogenetic inference enables evolutionary biologists to tackle scientific hypotheses, appropriately accounting for ancestry-induced correlation in observed trait values (Huelsenbeck *et al.* 2000; Pagel & Lutzoni 2002). Several authors demonstrate that mapping evolutionary histories onto inferred phylogenies provides a convenient and probabilistically grounded basis for designing statistically rigorous tests of evolutionary hypotheses (Nielsen 2002; Huelsenbeck *et al.* 2003; Dimmic *et al.* 2005). Unfortunately, this important statistical technique has been hampered by the high computational cost of stochastic mapping. Our general mathematical framework and fast algorithms should secure a central place for stochastic mapping in the statistical toolbox of evolutionary biologists.

We would like to thank Philippe Lemey for sharing his HIV codon alignments. We are also grateful to Andrew Meade for his help on using the BayesTraits software and Ziheng Yang and two anonymous reviewers for their constructive criticism. M.A.S. is supported by an Alfred P. Sloan Research Fellowship, John Simon Guggenheim Memorial Fellowship and NIH R01 GM086887.

REFERENCES

Ball, F. & Milne, R. 2005 Simple derivations of properties of counting processes associated with Markov renewal processes. *J. Appl. Prob.* **42**, 1031–1043. (doi:10.1239/jap/1134587814)

Cannings, C., Thompson, E. & Skolnick, M. 1980 Pedigree analysis of complex models. In *Current developments in anthropological genetics*, pp. 251–298. New York, NY: Plenum Press.

Dimmic, M., Hubisz, M., Bustamante, C. & Nielsen, R. 2005 Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* **21**, i126–i135. (doi:10.1093/bioinformatics/bti1032)

Drummond, A. & Rambaut, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)

Dutheil, J., Pupko, T., Marie, A. & Galtier, N. 2005 A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.* **22**, 1919–1928. (doi:10.1093/molbev/msi183)

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **13**, 93–104.

Felsenstein, J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Gelman, A., Meng, X. & Stern, H. 1996 Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* **6**, 733–807.

Guindon, S., Rodrigo, A., Dyer, K. & Huelsenbeck, J. 2004 Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl Acad. Sci. USA* **101**, 12 957–12 962. (doi:10.1073/pnas.0402177101)

Holmes, I. & Rubin, G. M. 2002 An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol* **317**, 753–764. (doi:10.1006/jmbi.2002.5405)

Huelsenbeck, J., Rannala, B. & Masly, J. 2000 Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**, 2349–2350. (doi:10.1126/science.288.5475.2349)

Huelsenbeck, J., Nielsen, R. & Bollback, J. 2003 Stochastic mapping of morphological characters. *Syst. Biol.* **52**, 131–158. (doi:10.1080/10635150390192780)

Lange, K. 2004 *Applied probability*. New York, NY: Springer.

Laub, A. 2004 *Matrix analysis for scientists and engineers*. Philadelphia, PA: SIAM.

Lemey, P., Pond, S. K., Drummond, A., Pybus, O., Shapiro, B., Barroso, H., Taveira, N. & Rambaut, A. 2007 Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* **3**, e29 . (doi:10.1371/journal.pcbi.0030029)

Leschen, R. & Buckley, T. 2007 Multistate characters and diet shifts: evolution of Erotylidae (Coleoptera). *Syst. Biol.* **56**, 97–112. (doi:10.1080/10635150701211844)

Meng, X. 1994 Posterior predictive *p*-values. *Ann. Stat.* **22**, 1142–1160. (doi:10.1214/aos/1176325622)

Minin, V. & Suchard, M. 2008 Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412. (doi:10.1007/s00285-007-0120-8)

Neuts, M. 1995 *Algorithmic probability: a collection of problems*. London, UK: Chapman and Hall.

Nielsen, R. 2002 Mapping mutations on phylogenies. *Syst. Biol.* **51**, 729–739. (doi:10.1080/10635150290102393)

Pagel, M. 1999 The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48**, 612–622. (doi:10.1080/106351599260184)

Pagel, M. & Lutzoni, F. 2002 Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. *Biological evolution and statistical physics*, pp. 148–161. Berlin, Germany: Springer.

Pagel, M. & Meade, A. 2006 Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167**, 808–825. (doi:10.1086/503444)

Pagel, M., Meade, A. & Barker, D. 2004 Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53**, 673–684. (doi:10.1080/10635150490522232)

Rodrigue, N., Philippe, H. & Lartillot, N. 2008 Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics* **24**, 56–62. (doi:10.1093/bioinformatics/btm532)

Schadt, E., Sinsheimer, J. & Lange, K. 1998 Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res.* **8**, 222–233.

Shankarappa, R. *et al.* 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10 489–10 502.

Shapiro, B., Rambaut, A. & Drummond, A. 2006 Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* **23**, 7–9. (doi:10.1093/molbev/msj021)

Yang, Z. 1996 Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587–596. (doi:10.1007/BF02352289)