PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 4041–4047
doi:10.1098/rstb.2008.0140
Published online 7 October 2008

The impact of single substitutions on multiple sequence alignments

Steffen Klaere*, Tanja Gesell and Arndt von Haeseler

Center for Integrative Bioinformatics Vienna, University of Vienna, Medical University Vienna, Veterinary University Vienna, Max F. Perutz Laboratories, Dr Bohrgasse 9, 1030 Wien, Austria

We introduce another view of sequence evolution. Contrary to other approaches, we model the substitution process in two steps. First we assume (arbitrary) scaled branch lengths on a given phylogenetic tree. Second we allocate a Poisson distributed number of substitutions on the branches. The probability to place a mutation on a branch is proportional to its relative branch length. More importantly, the action of a single mutation on an alignment column is described by a doubly stochastic matrix, the so-called one-step mutation matrix. This matrix leads to analytical formulae for the posterior probability distribution of the number of substitutions for an alignment column.

Keywords: maximum likelihood; maximum parsimony; substitution model; posterior probability; tree reconstruction

1. INTRODUCTION

Tree reconstruction is nowadays a matter of routinely applying the available programs, comparing the resulting trees and then concluding what might be the best tree (e.g. Hillis *et al.* 1996; Felsenstein 2004; Schmidt & von Haeseler 2008). With the advent of Baysian approaches, it is possible to model increasingly complex evolutionary scenarios (Huelsenbeck & Ronquist 2001). However, a detailed understanding about what can and cannot be inferred (Baake 1998; Mossel & Steel 2005) from a sequence alignment is still missing. In recent years, some theoretical insights were obtained by studying very simple models of sequence evolutions on binary sequence data (e.g. Erdös *et al.* 1999*a*,*b*). Although these models are not realistic in a biological sense, they have provided some profound insights in the reconstruction process *per se*.

Here, we will introduce another description of the evolutionary process on trees. More precisely, given a phylogenetic tree and an alignment that evolved along the tree, we now ask the following question: how does the alignment change if an additional substitution on an arbitrary branch of the tree takes place? This rather abstract question is motivated by the following biological problem. Consider a collection of morphological traits that are either in an ancestral (0) or derived (1) state. Each derived character state characterizes a monophyletic group and represents a cluster in the tree. For such a data matrix (or alignment), the tree reconstruction problem is easy. However, stochastic effects that act somewhere on the branches of the tree may disturb this signal. This noise is modelled by the assumption of throwing an arbitrary number of changes on the tree and measuring their impact on the otherwise perfect data matrix. To this end, we construct a one-step mutation (OSM) matrix. The matrix description allows a linear algebra view of evolution and comprises distance methods, maximum parsimony as well as maximum likelihood. Moreover, the description reveals a surprising connection to Hadamard matrices that were employed for phylogenetic questions (Hendy & Penny 1989; Steel *et al.* 1998). An immediate application of OSM matrices is the analytical computation of posterior probabilities that count the number of evolutionary changes on a tree. So far, these posterior probabilities have been estimated using Bayesian simulation (Nielsen 2002; Huelsenbeck *et al.* 2003) or by applying the theory of counting processes (Minin & Suchard 2008).

In the following we refrain from most technicalities but rather outline the general ideas by discussing illustrative examples. The technical details and proofs will appear elsewhere.

2. THE BINARY MODEL ON AN *n*-TAXON TREE

(a) *Notation*

We consider a set of n taxa $S=\{1, \dots, n\}$. With S comes some information about the common properties and differences of the taxa, typically displayed in an alignment. In the following a (sequence) *alignment* $\mathbb{A}$ is an $n\times\ell$-array with entries either 0 or 1, where ℓ is the length of the alignment. Each of the ℓ columns (*sites*) $\boldsymbol{a}_j$ of the alignment represents a *pattern* of n homologous characters, where $a_{ij}\in\{0,1\}$ is the state of character j in taxon i. For binary character states, 2^n patterns are possible.

We are interested in the evolution of such patterns along a (rooted) *tree* $T=(V, E)$ with node set V and branch set $E\subset V\times V$ (Semple & Steel 2003). The node set V contains the taxon set S that forms the leaf set. Avoiding the technical details, each branch is uniquely encoded by the subsets X of S that originates from the branch. Such a set X specified by a branch will be called a *cluster*. A leaf is a trivial cluster.

*Author for correspondence (steffen.klaere@univie.ac.at).

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

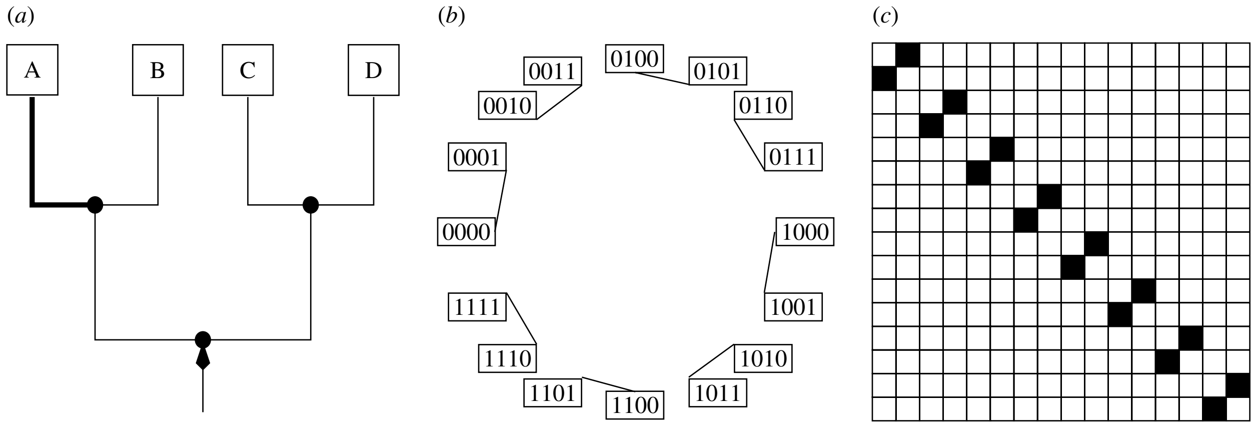

Figure 1. (*a*) The rooted phylogenetic tree T_4. The branch defined by cluster {A} is highlighted. A substitution on this branch gives rise to the unique change in patterns depicted in (*b*) the graph or (*c*) its corresponding adjacency matrix.

Finally, we introduce a function $\lambda: E \rightarrow \mathbb{R}_+$, such that $\lambda(e) > 0$ represents the *length* of a branch $e \in E$. The *tree-length* Λ_T is the sum of the branch lengths. The *relative branch length*

$$p_e = \frac{\lambda(e)}{\Lambda_T} \qquad (2.1)$$

denotes the probability that a substitution hits branch e of the rooted tree T.

(b) *The effect of substitutions on an alignment*

We now describe how a single mutation on the tree changes the current character states at the leaves. Obviously the outcome will depend on the branch where the substitution occurred. Moreover, each of the 2^n possible patterns will be affected differently by such a substitution. Therefore we introduce a $2^n \times 2^n$ matrix that describes the action of a substitution on the patterns for a specific branch.

Figure 1 describes the model on an example tree T_4 with four taxa. For instance, a substitution at the branch defined by cluster {A} changes the pattern 1011 to the pattern 1010 because only the character of taxon A is affected. Please note that order of taxa is (D, C, B, A) for each pattern. All possible changes between the patterns identified by a substitution on branch e_A are depicted in the substitution graph (figure 1*b*). The corresponding adjacency matrix σ_A is displayed in figure 1*c*, where a black square stands for one (the patterns are connected by an edge in the substitution graph) and a white square represents zero. The structure of matrix σ_A constitutes an example of the so-called *permutation matrices* with entries equal to one if the substitution converts one pattern into another (Bona 2004, p. 75).

For each branch we easily construct the corresponding permutation matrix. Without proof we state that each matrix is fix point free (a substitution changes every pattern) and has 2^{n-1} transpositions (applying a substitution twice returns the original pattern). Then it follows that this type of permutation matrix is self-inverse with respect to matrix multiplication and that they form an Abelian group if the identity matrix is included.

We point out that the permutation matrix for a non-trivial cluster is the product of the permutation matrices of its elements. In other words the action of one mutation on a branch e can be replaced by any partition of the cluster associated with e, such that each set of the partition is represented by a branch in the tree. For tree T_4 we obtain six permutation matrices σ_A, σ_B, σ_C, σ_D and $\sigma_{AB} = \sigma_A \cdot \sigma_B$, $\sigma_{CD} = \sigma_C \cdot \sigma_D$.

Because we study symmetric permutation matrices, it is well known that the eigenvalues are either 1 or −1. Further, we observe that all those permutation matrices are diagonalized by the $2^n \times 2^n$-dimensional Hadamard matrix $\boldsymbol{H}_{2^n}$ (Hendy & Penny 1989; Hendy 2005). The 2×2 Hadamard matrix is defined as

$$\boldsymbol{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The $2^n \times 2^n$ Hadamard matrix is then

$$\boldsymbol{H}_{2^n} = \underbrace{\boldsymbol{H}_2 \otimes \cdots \otimes \boldsymbol{H}_2}_{n\text{-times}},$$

where the symbol ⊗ represents the Kronecker product. Hence, for the cluster $\mathcal{C}(e)$ associated with branch e,

$$D_e = \boldsymbol{H}_{2^n}^{-1} \sigma_{\mathcal{C}(e)} \boldsymbol{H}_{2^n}$$

is the diagonal matrix of eigenvalues of $\sigma_{\mathcal{C}(e)}$.

To take the relative contribution of the branch lengths into account, we weight each permutation matrix with p_e as described in (2.1). Such matrices are a special case of the *generalized* permutation matrices. Then the so-called OSM matrix of the tree T_4 is simply the following convex sum:

$$M_4 = p_A \cdot \sigma_A + p_B \cdot \sigma_B + p_C \cdot \sigma_C + p_D \cdot \sigma_D + p_{AB} \cdot \sigma_{AB} + p_{CD} \cdot \sigma_{CD}.$$

Figure 2*c* shows the result of this computation. The substitution graph in figure 2*b* displays the effect of a substitution on all the branches of the tree on the patterns. Two patterns are connected by an edge if a substitution switches between the two patterns.

For an arbitrary phylogenetic tree T on n taxa, the OSM matrix is obtained by

$$M_T = \sum_{e \in E} p_e \sigma_{\mathcal{C}(e)}, \qquad (2.2)$$

where $\mathcal{C}(e)$ is the cluster identified by branch $e \in E$. Owing to its construction, M_T is also diagonalized by the Hadamard matrix.

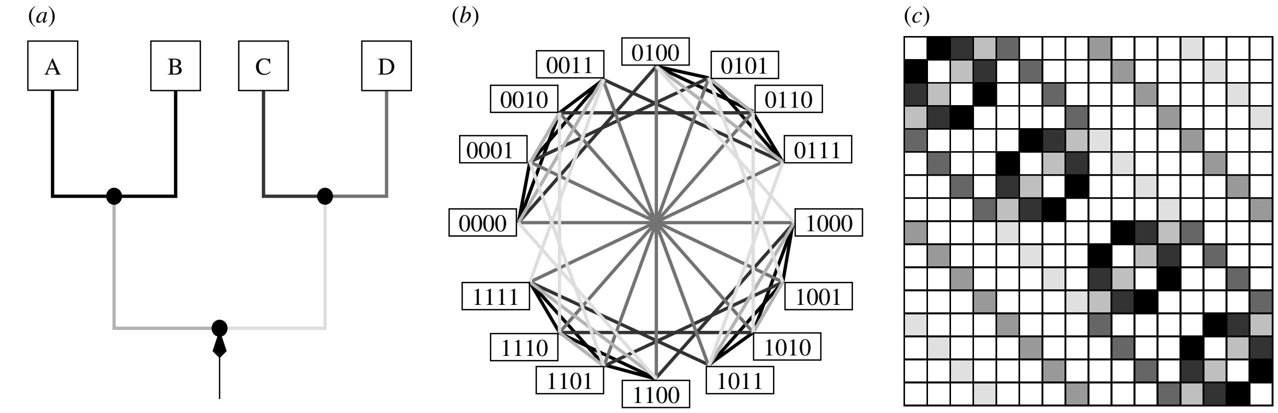

Figure 2. (*a*) Substitutions on different branches of T_4 give rise to branches of (*b*) the graph, and (*c*) the corresponding entries in the OSM matrix. Corresponding components are identified by common shades of grey.

The entry $M_T(\boldsymbol{i}, \boldsymbol{j})$ is positive if the tree T contains a cluster where a substitution on the corresponding branch implies that pattern $\boldsymbol{i}$ is changed to $\boldsymbol{j}$. Hence, each row and each column has $2n-2$ non-zero entries, one entry for each branch in the tree. Thus, the OSM matrix belongs to the class of doubly stochastic Markov transition matrices, where the relative branch lengths are represented exactly once in each row and each column. Consequently, the k-th power M_T^k provides the probabilities to move from one pattern to another in k substitutions. Thus, the repeated application of M_T describes a *random walk* on the state space of the 2^n patterns. This random walk is very different from the standard random walk on the hypercube (Eigen *et al.* 1989). If k is large, then M_T^k will lose the phylogenetic information of the original alignment and will approximate the uniform distribution, that is each pattern occurs at the same frequency.

(c) *Poisson weights*

Our setting does not yet assume a probability distribution for the number of substitutions on the tree. In molecular evolution, one typically assumes that the number of substitutions is Poisson distributed with parameter Λ_T (Uzzell & Corbin 1971). Under this assumption we can compute the average OSM matrix by

$$\overline{M}_T = \sum_{k=0}^{\infty} \frac{\exp(-\Lambda_T)\Lambda_T^k}{k!} M_T^k, \tag{2.3}$$

which is equivalent to

$$\overline{M}_T = \exp(-\Lambda_T) \cdot \exp[\Lambda_T M_T].$$

The exponential of the matrix $\Lambda_T M_T$ is easy to compute, because M_T is a sum of generalized permutation matrices (2.2), which commute with respect to matrix multiplication. Thus, we obtain

$$\begin{aligned} \overline{M}_T &= \exp(-\Lambda_T) \cdot \exp\left[\textstyle\sum_{e \in E} \lambda(e) \cdot \sigma_{C(e)}\right], \\ &= \exp(-\Lambda_T) \cdot \textstyle\prod_{e \in E} \exp[\lambda(e) \cdot \sigma_{C(e)}], \\ &= \exp(-\Lambda_T) \cdot \textstyle\prod_{e \in E} \boldsymbol{H}_{2^n}^{-1} \cdot \exp[\lambda(e) D_e] \cdot \boldsymbol{H}_{2^n}, \\ &= \exp(-\Lambda_T) \cdot \boldsymbol{H}_{2^n}^{-1} \cdot \exp\left[\textstyle\sum_{e \in E} \lambda(e) D_e\right] \cdot \boldsymbol{H}_{2^n}. \end{aligned} \tag{2.4}$$

3. RELATION TO TREE RECONSTRUCTION

The OSM matrix leads to a very general description of character-based phylogenetic inference techniques. Moreover, the explicit model assumptions in maximum likelihood and the implicit assumptions in maximum parsimony are directly comparable.

The OSM matrix and its powers describe the substitution process between arbitrary patterns. However, in classical phylogeny the starting point of a substitution process is ancestral states on trees. In particular, one assumes a stationary distribution $\pi = (\pi_0, \pi_1)$ of character states at the root, and the characters evolve along the tree according to a Markov transition matrix (Tavaré 1986). In our framework, this is equivalent to starting in the *constant* patterns $\boldsymbol{0} = (0, \ldots, 0)$ or $\boldsymbol{1} = (1, \ldots, 1)$ and letting it evolve according to the OSM matrix. This process has a non-stationary pattern distribution π_T^k which starts at $\pi_T^0 = (\pi_0, 0, \ldots, 0, \pi_1)$, i.e. with zero substitutions only constant patterns exist, and in each step the pattern distribution is given by $\pi_T^k = M_T^k \pi_T^0$. If the number of substitutions is not weighted as in equation (2.3), then π_T^k will approach the uniform distribution as k goes to infinity. To overcome the loss of phylogenetic signal, we assume in the following that the number of substitutions on a tree is Poisson distributed with parameter Λ_T. Moreover, we assume that the substitution process is described by the symmetric Cavender–Farris–Neyman mutation model (CFN, Neyman 1971; Farris 1973; Cavender 1978). Under these assumptions, the probability of observing pattern $\boldsymbol{a}$ when starting in a constant pattern is then calculated employing (2.3)

$$\mathbb{P}[\boldsymbol{a} | \{\boldsymbol{0}, \boldsymbol{1}\}] = \pi_0 \overline{M}_T(\boldsymbol{0}, \boldsymbol{a}) + \pi_1 \overline{M}_T(\boldsymbol{1}, \boldsymbol{a}), \tag{3.1}$$

where π_0 and π_1 are taken from the stationary distribution of character states. The resulting probability distribution for all possible patterns is then identical to the standard way of computing the probabilities of pattern (Felsenstein 2004).

(a) *Distance approaches*

Now, we briefly illustrate how to derive distance corrections from the OSM matrix. To this end, we consider the rooted tree with two leaves A, B and branch lengths λ_A and λ_B. Then the corresponding

$$\begin{aligned}\overline{M}_2 &= e^{-\Lambda_2}\cdot\exp\left[\lambda_\mathrm{A}\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}+\lambda_\mathrm{B}\begin{pmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{pmatrix}\right],\\ &= e^{-\Lambda_2}\left(\begin{array}{cc|cc}\cosh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B}\\ \sinh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B}\\ \hline \cosh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B}\\ \sinh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B} & \sinh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B} & \cosh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B}\end{array}\right)\end{aligned} \tag{*}$$

OSM matrix $\boldsymbol{M}_2$ has the following structure:

$$\boldsymbol{M}_2 = \begin{pmatrix}0 & p_\mathrm{A} & p_\mathrm{B} & 0\\ p_\mathrm{A} & 0 & 0 & p_\mathrm{B}\\ p_\mathrm{B} & 0 & 0 & p_\mathrm{A}\\ 0 & p_\mathrm{B} & p_\mathrm{A} & 0\end{pmatrix},$$

where p_A and p_B are computed according to (2.1). From (2.4) it is straightforward to compute $\overline{M}_2$ using (∗).

The first and the last row of $\overline{M}_2$ yield the probabilities to observe one of the four patterns. If evolution starts with character state 0 or 1 at the root of the tree and the character states are in equilibrium ($\pi_0 = \pi_1$), then we quickly compute

$$\mathbb{P}(\mathbf{0},\mathbf{1}) = \frac{1}{2}\left(1+\exp(-2\Lambda)\right)$$

as the probability to observe a constant pattern in an alignment. Similarly we compute the probability to observe different character states between taxa A and B. From this it is straightforward to get the distance correction of the CFN model.

(b) *Maximum likelihood*

The maximum-likelihood principle for an alignment $\mathbb{A}$ and a tree T is easily formulated in terms of the OSM matrix. We introduce as parameter vector $\boldsymbol{\theta}$ the branch lengths of T. Then the probability of $\mathbb{A}$ is given by

$$L(\mathbb{A}|\mathrm{T}) = \prod_{i=1}^{\ell}\mathbb{P}[\boldsymbol{a}_i|\{\mathbf{0},\mathbf{1}\}], \tag{3.2}$$

where the factors on the right-hand side are defined by equation (3.1). The parameter vector $\boldsymbol{\theta}$ enters the equation via the OSM and $\Lambda = \sum\theta_i$ in the obvious way. As usual, we want to find parameter assignments such that (3.2) is maximized.

(c) *Maximum parsimony*

We associate the adjacency matrix $\boldsymbol{A}_\mathrm{OSM}$ (e.g. Cormen *et al.* 2001, §22.1), or simply $\boldsymbol{A}$, with the OSM matrix. $\boldsymbol{A}$ is obtained as the unweighted sum of the permutation matrices $\sigma_{\mathcal{C}(e)}$. Hence, an entry $\boldsymbol{A}_{ij}$ is equal to one when there is a branch in the tree which changes pattern i into pattern j, and is zero otherwise. Finally, we note that $\boldsymbol{A}^k(i,j)$ describes the number of paths of length k between pattern i and j. Each path specifies a series of branches in the tree where a substitution occurred.

Now, fix a column $\boldsymbol{a}_i$ in alignment $\mathbb{A}$, and a tree T. We ask for the minimal number k_min such that $\boldsymbol{A}^{k_\mathrm{min}}(\boldsymbol{a}_i,\mathbf{0})$ or $\boldsymbol{A}^{k_\mathrm{min}}(\boldsymbol{a}_i,\mathbf{1})$ is greater than zero. In other words, for an alignment column $\boldsymbol{a}_i$, the minimal number of mutations on T equals

$$MP(\boldsymbol{a}_i) = \min\{k\in\mathbb{N}|\boldsymbol{A}^k(\boldsymbol{a}_i,\mathbf{0})>0 \text{ or } \boldsymbol{A}^k(\boldsymbol{a}_i,\mathbf{1})>0\}.$$

Thus the minimal number of mutations for an alignment $\mathbb{A} = (\boldsymbol{a}_1,\ldots,\boldsymbol{a}_\ell)$ equals

$$MP(\mathrm{T}) = \sum_{i=1}^{\ell} MP(\boldsymbol{a}_i). \tag{3.3}$$

This is another description of the maximum-parsimony principle.

4. MAPPING SUBSTITUTIONS

From the computation of the powers of the OSM matrix, it is possible to derive the (posterior) probability distribution $\mathrm{ppdf}(k|\boldsymbol{x})$ of the number of mutations that generated an observed pattern $\boldsymbol{x}$, when the process started in pattern $\mathbf{0}$ or $\mathbf{1}$. The posterior probabilities have been estimated before, employing Bayesian simulation methods (Nielsen 2002; Huelsenbeck *et al.* 2003; Minin & Suchard 2008), but an analytic approach has not previously been attempted.

In general, the posterior probabilities $\mathrm{ppdf}(k|\boldsymbol{a})$ for a pattern $\boldsymbol{a}$ are calculated in the following way using (3.1):

$$\mathrm{ppdf}(k|\boldsymbol{a}) = \frac{\mathrm{e}^{-\Lambda_\mathrm{T}}\Lambda_\mathrm{T}^k(\pi_0 M_\mathrm{T}^k(\mathbf{0},\boldsymbol{a}) + \pi_1 M_\mathrm{T}^k(\mathbf{1},\boldsymbol{a}))}{k!\,\mathbb{P}[\boldsymbol{a}|\{\mathbf{0},\mathbf{1}\}]}$$

i.e. we compute for pattern $\boldsymbol{a}$ the proportion of its occurrence after k substitutions.

For the two-taxon case, we observe that

$$\mathbb{P}[00|\mathbf{0}] = \exp(-\Lambda)\cosh\lambda_\mathrm{A}\cosh\lambda_\mathrm{B},$$
$$\mathbb{P}[00|\mathbf{1}] = \exp(-\Lambda)\sinh\lambda_\mathrm{A}\sinh\lambda_\mathrm{B}.$$

Thus, applying hyperbolic identities, we end up with

$$\mathbb{P}[00|\mathbf{0},\mathbf{1}] = \frac{1}{2}\exp(-\Lambda)\cosh\Lambda.$$

Taylor expansion of $\cosh(x)$ around $x=0$ then leads to the posterior probability

$$\mathrm{ppdf}(2k|00) = \frac{\Lambda^{2k}}{(2k)!\cdot\cosh\Lambda}.$$

For the remaining patterns, we obtain the following ppdf:

$$\begin{aligned}\mathrm{ppdf}(2k|11) &= \mathrm{ppdf}(2k|00),\\ \mathrm{ppdf}(2k+1|01) &= \mathrm{ppdf}(2k+1|10)\\ &= \frac{\Lambda^{2k+1}}{(2k+1)!\cdot\sinh\Lambda}.\end{aligned}$$

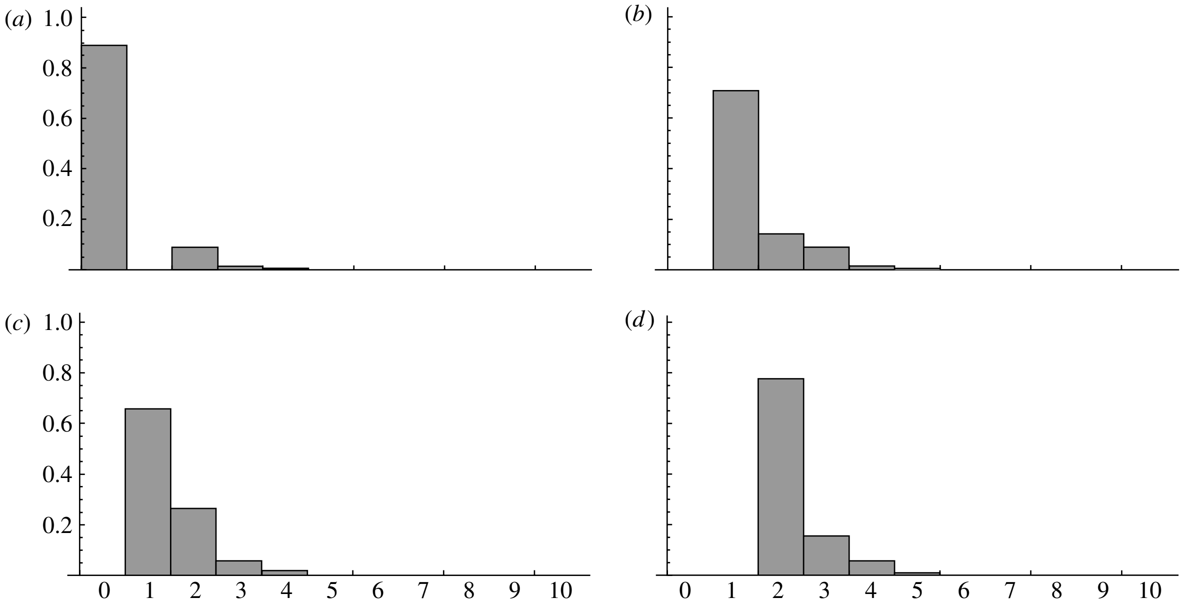

Figure 3. Posterior probabilities for representative patterns of the four-taxon tree T_4 with branch lengths $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0.2$ and $\lambda_{AB} = \lambda_{CD} = 0.1$, and character distribution $\pi_0 = \pi_1 = 1/2$. The selected patterns (*a*) 0000, (*b*) 0001, (*c*) 0011 and (*d*) 0101 represent constant, parsimonious uninformative, compatible and incompatible patterns, respectively.

Thus, on a two-taxon tree, an even number of substitutions leads to a constant pattern when starting in a constant pattern; whereas an odd number of substitutions will be reflected by the non-constant pattern 01 or 10.

Now, it is an easy calculation to obtain closed formulae for the posterior mean number $\mu(\boldsymbol{a})$ of substitutions for pattern $\boldsymbol{a}$ by the following calculations:

$$\mu(00) = \mu(11) = \sum_{k=0}^{\infty} \frac{2k\Lambda^{2k}}{(2k)!\cosh\Lambda} = \Lambda\tanh\Lambda,$$

$$\mu(01) = \mu(10) = \sum_{k=0}^{\infty} \frac{(2k+1)\Lambda^{2k+1}}{(2k+1)!\sinh\Lambda} = \Lambda\coth\Lambda.$$

Only if Λ is large, then the posterior mean number of substitutions will approach the expected number of substitutions per site Λ. For a constant pattern the posterior mean is always smaller than Λ, and for non-constant patterns the posterior mean is larger than Λ.

Similarly we extend the calculations to a four-taxon tree. For instance, consider the four-taxon tree T_4 (figure 1*a*). This tree has two non-trivial clusters {A, B} and {C, D}. We want to compute the posterior probability of the number of substitutions if the constant pattern 0000 is observed. Let us assume that the two character states occur with uniform probability; then we can compute:

$$\begin{aligned}\mathbb{P}[0000|\{\mathbf{0},\mathbf{1}\}] = \frac{1}{16e^{\Lambda}}&(\exp(\lambda_A - \lambda_B + \lambda_C - \lambda_D - \lambda_X)\\ &+ \exp(\lambda_A - \lambda_B - \lambda_C + \lambda_D - \lambda_X)\\ &+ \exp(-\lambda_A + \lambda_B - \lambda_C + \lambda_D - \lambda_X)\\ &+ \exp(\lambda_A + \lambda_B - \lambda_C - \lambda_D + \lambda_X)\\ &+ \exp(-\lambda_A - \lambda_B + \lambda_C + \lambda_D + \lambda_X)\\ &+ \exp(-\lambda_A - \lambda_B - \lambda_C - \lambda_D + \lambda_X)\\ &+ \exp(-\lambda_A + \lambda_B + \lambda_C - \lambda_D - \lambda_X)\\ &+ \exp(\lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_X)),\end{aligned}$$

where $\lambda_X = \lambda_{AB} + \lambda_{CD}$ is the sum of the lengths of the interior branches. Now Taylor expansion leads to the desired posterior probability distribution.

Figure 3 shows the resulting posterior probability distributions for the 16 possible patterns, assuming branch lengths $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0.2$ and $\lambda_{AB} = \lambda_{CD} = 0.1$. The symmetries in the CFN model are reflected in the symmetries of the posterior distributions. Complementary patterns (i.e. 0000 and 1111) show the same distribution. Because the tree is clock-like, the parsimonious uninformative patterns (0001, 0010, 0100, 1000) and their complements show identical distributions, as do the patterns that need at least two substitutions (0101, 0110, 1010, 1001) on T_4. Posterior probabilities may be used to compute, for instance, the number of unvaried sites (Fitch & Ayala 1994), which is exactly the proportion of the constant patterns with zero substitutions. In our example we expect approximately 42 per cent constant patterns of which approximately 90 per cent are unvaried. This is only one application for posterior probabilities of the number of substitutions.

As in the two taxon case, we compute the posterior mean of substitutions for pattern $\boldsymbol{a}$ as

$$\mu(\boldsymbol{a}) = \sum_{k=0}^{\infty} k \cdot \mathrm{ppdf}(k|\boldsymbol{a}).$$

Figure 4*a* shows the posterior mean number of substitutions for the topology of T_4 with branch probabilities $p_A = p_B = p_C = p_D = 0.2$ and $p_{AB} = p_{CD} = 0.1$ for a constant pattern (0000), a pattern compatible with an interior branch (0011) and a pattern incompatible with the tree (0110). The difference between posterior mean and tree lengths is smaller than 0.01 if the tree lengths exceed 10 substitutions per site.

Figure 4*b* displays the posterior means for a tree with branch probabilities $p_A = p_D = 0.47$, $p_B = p_C = 0.02$ and $p_{AB} = p_{CD} = 0.01$. The proportion of p_A and p_D is so large that the incompatible pattern 0110 will be observed more often than the pattern 0011, which is compatible with a branch of the tree. Thus, this tree is

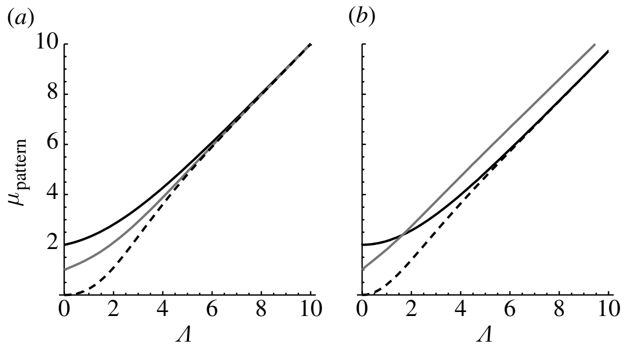

Figure 4. Posterior mean number of substitutions as a function of the tree length Λ for the tree topology T_4. The posterior means for patterns 0000 (dashed line), 0011 (grey solid line) and 0110 (black solid line) are shown. (*a*) The posterior means on the tree with relative branch lengths $p_A = p_B = p_C = p_D = 0.2$ and $p_{AB} = p_{CD} = 0.1$. (*b*) The result for relative branch lengths $p_A = p_D = 0.47$, $p_B = p_C = 0.02$ and $p_{AB} = p_{CD} = 0.01$.

an instance where maximum parsimony will reconstruct the wrong tree (Felsenstein 1978). The figure also shows that the compatible pattern 0011 has a lower posterior mean number of substitutions than 0110 for short tree lengths. However, if the tree length exceeds 1.64 substitutions per site, then the situation is reversed. The posterior mean of the incompatible pattern quickly approaches the tree length, whereas the mean posterior substitutions of the compatible pattern are only close to the tree length if $\Lambda \geq 54$ substitutions per site. In other words, if we observe a compatible pattern, then this pattern has typically experienced more substitutions than the incompatible pattern.

5. SUMMARY AND DISCUSSION

Here, we have presented an alternative description of how to model sequence evolution on a tree. Our approach lifts the commonly used stochastic models of sequence evolution that act on nucleotides to the set of all possible patterns for n taxa.

We have shown that available tree reconstruction principles are included in our description of the process. Moreover, the definition of the OSM matrix leads to analytical formulae to compute the posterior probability distribution of the number of substitutions for each pattern. From this distribution it is then straightforward to compute the posterior mean of the substitutions. The formulation of the substitution process as an OSM matrix leads to the introduction of the Hadamard matrix that allows an easy computation of matrix powers. Recently, Bryant (submitted) has presented a continuous version of the OSM approach and showed its connection to the Hadamard matrix (Hendy 2005).

While we have outlined only the simplest model of sequence evolution, several extensions are easily possible. The OSM approach can be augmented to the Kimura 3st model (Kimura 1981); see figure 5*a* for an illustration. In this framework every substitution class (transition, transversion 1 and transversion 2) uniquely generates a fix-point free $4^n \times 4^n$-dimensional permutation matrix for each branch in a tree. Let $\alpha_1 + \alpha_2 + \alpha_3 = 1$ denote the probabilities for the three substitution classes, then the OSM matrix for the Kimura 3st model is defined as

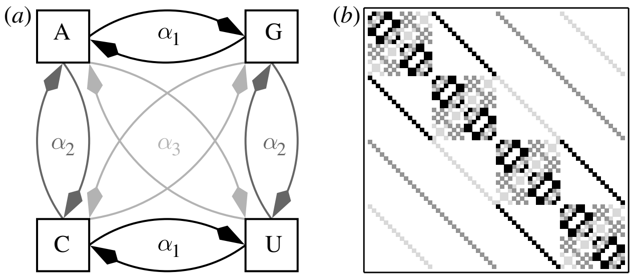

Figure 5. (*a*,*b*) The transition scheme and an example of an OSM matrix under the Kimura 3st model of a rooted triplet tree. All four branch lengths are taken to be equal. Hence, black squares indicate a transition, dark grey squares an α_2 transversion and light grey squares an α_3 transversion.

$$M_T = \sum_{k=1}^{3} \alpha_k \sum_{e \in E} p_e \cdot \sigma^k_{C(e)} \tag{5.1}$$

i.e. we look at the sum of generalized permutation matrices. Figure 5*a* shows an OSM Kimura matrix on a rooted triplet tree. Each row and each column contain 12 = (number of branches) × (number of substitution classes) non-zero entries, where each entry is the product of a mutation class parameter α_i and a branch probability p_e (equation (5.1)). All the results for binary character state models can be expanded to the Kimura 3st model. If one wants to abandon the assumption that evolution proceeds along a tree, then this is also possible within the OSM framework. Consider a set of rooted trees that give rise to a collection $\mathfrak{C}$ of possibly conflicting clusters. The associated OSM matrix is then given by

$$M_{\mathfrak{C}} = \sum_{C \in \mathfrak{C}} p_C \sigma_C,$$

where p_C is the normalized sum of branch lengths of those trees in which the branch depicting C is existent. Here issues such as the meaning of the overall length of the cluster set or the meaning of a root in such sets need to be discussed. This extension bears some similarity to a maximum-likelihood reconstruction of networks (von Haeseler & Churchill 1993).

Another question concerns the Poisson weights for the number substitutions. Generally, the argument is that the process of distributing substitutions along a tree is memoryless and therefore the number of substitutions is Poisson distributed. Our framework permits a different probability distribution to be assigned to the substitution process. One possible weighting scheme could be a contagious distribution, which has been used to evaluate accident data (Kemp 1967). This approach might provide an alternative description of the evolutionary history of an alignment.

In summary, the OSM description offers a variety of potential applications in molecular systematics, which will be explored in the near future.

We are grateful to a number of colleagues who helped to increase our understanding of the model by asking the small questions necessary to get that kind of understanding. Helpful comments by two anonymous reviewers are gratefully acknowledged. Support by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF) is kindly acknowledged. T.G. and A.v.H. are also supported by the Austrian Ministry for Science and Research (GEN-AU Bioinformatics Integration Network II).

REFERENCES

Baake, E. 1998 What can and what cannot be inferred from pairwise sequence comparisons? *Math. Biosci.* **154**, 1–21. (doi:10.1016/S0025-5564(98)10044-5)

Bona, M. 2004 *Combinatorics of permutations*. Boca Raton, FL: Chapman and Hall–CRC.

Bryant, D. Submitted. Hadamard phylogenetic methods and the *n*-taxon process. (http://arXiv.org/abs/0806.1378)

Cavender, J. A. 1978 Taxonomy with confidence. *Math. Biosci.* **40**, 271–280. (doi:10.1016/0025-5564(78)90089-5)

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. 2001 *Introduction to algorithms*, 2nd edn. Cambridge, MA: MIT Press and McGraw-Hill.

Eigen, M., Lindemann, B. F., Tietze, M., Winkler-Oswatitsch, R., Dress, A. & von Haeseler, A. 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**, 673–679. (doi:10.1126/science.2497522)

Erdös, P., Steel, M., Székely, L. & Warnow, T. 1999*a* A few logs suffice to build (almost) all trees (part 1). *Random Struct. Algor.* **14**, 153–184. (doi:10.1002/(SICI)1098-2418(199903)14:2<153::AID-RSA3>3.0.CO;2-R)

Erdös, P., Steel, M., Székely, L. & Warnow, T. 1999*b* A few logs suffice to build (almost) all trees (part 2). *Theor. Comput. Sci.* **221**, 77–118. (doi:10.1016/S0304-3975(99)00028-6)

Farris, J. 1973 A probability model for inferring evolutionary trees. *Syst. Zool.* **22**, 250–256. (doi:10.2307/2412305)

Felsenstein, J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410. (doi:10.2307/2412923)

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

Fitch, W. M. & Ayala, F. J. 1994 The superoxide dismutase molecular clock revisited. *Proc. Natl Acad. Sci. USA* **91**, 6802–6807. (doi:10.1073/pnas.91.15.6802)

Hendy, M. D. 2005 Hadamard conjugation: an analytic tool for phylogenetics. In *Mathematics of evolution and phylogeny* (ed. O. Gascuel), pp. 143–177. Oxford, UK: Oxford University Press.

Hendy, M. D. & Penny, D. 1989 A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**, 297–309. (doi:10.2307/2992396)

Hillis, D. M., Moritz, G. & Mable, B. K. (eds) 1996 *Molecular systematics*, 2nd edn. Sunderland, MA: Sinauer Associates.

Huelsenbeck, J. P. & Ronquist, F. 2001 MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)

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

Kemp, C. D. 1967 On a contagious distribution suggested for accident data. *Biometrics* **23**, 241–255. (doi:10.2307/2528159)

Kimura, M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci. USA* **78**, 454–458. (doi:10.1073/pnas.78.1.454)

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

Mossel, E. & Steel, M. 2005 How much can evolved characters tell us about the tree that generated them? In *Mathematics of evolution and phylogeny* (ed. O. Gascuel), pp. 384–412. Oxford, UK: Oxford University Press.

Neyman, J. 1971 Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics* (ed. J. Y. S. S. Gupta), pp. 1–27. New York, NY: Academic Press.

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

Schmidt, H. A. & von Haeseler, A. 2008 Phylogenetic inference using maximum likelihood methods. In *The phylogenetic handbook* (eds P. Lemey, M. Salemi & A. Vandamme), 2nd edn. Cambridge, UK: Cambridge University Press.

Semple, C. & Steel, M. 2003 *Phylogenetics. Oxford Lectures Series in Mathematics and its Applications*. Oxford, UK: Oxford University Press.

Steel, M., Hendy, M. D. & Penny, D. 1998 Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. *Discret. Appl. Math.* **88**, 367–396. (doi:10.1016/S0166-218X(98)00080-8)

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

Uzzell, T. & Corbin, K. W. 1971 Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089–1096. (doi:10.1126/science.172.3988.1089)

von Haeseler, A. & Churchill, G. A. 1993 Network models for sequence evolution. *J. Mol. Evol.* **37**, 77–85. (doi:10.1007/BF00170465)