Bulletin of Mathematical Biology (2009) 71: 339–351
DOI 10.1007/s11538-008-9364-8

ORIGINAL ARTICLE

Hadamard Phylogenetic Methods and the *n*-taxon Process

David Bryant

Department of Mathematics, University of Auckland, Auckland, New Zealand

Received: 11 December 2007 / Accepted: 18 September 2008 / Published online: 10 October 2008

Abstract The Hadamard transform (Hendy and Penny, Syst. Zool. 38(4):297–309, 1989; Hendy, Syst. Zool. 38(4):310–321, 1989) provides a way to work with stochastic models for sequence evolution without having to deal with the complications of tree space and the graphical structure of trees. Here we demonstrate that the transform can be expressed in terms of the familiar $\mathbf{P}[\tau] = e^{\mathbf{Q}[\tau]}$ formula for Markov chains. The key idea is to study the evolution of vectors of states, one vector entry for each taxa; we call this the *n*-taxon process. We derive transition probabilities for the process. Significantly, the findings show that tree-based models are indeed in the family of (multi-variate) exponential distributions.

Keywords Phylogenetics · Stochastic models · Hadamard conjugation · Spectral decomposition

1. Introduction

1.1. Stochastic models and the Hadamard transform

Stochastic models for sequence evolution now play a part in most phylogenetic analyses. Given a tree, and a set of branch lengths, the models determine a probability distribution for the patterns of nucleotides/amino acids observed at a site in the alignment (the *site patterns*). The real difficulty of these tree-based models is that they are, indeed, based on a tree. The graphical structure of the tree is intrinsic to the probability formulas. Working with abstract spaces made up of discrete graphical structures is difficult for both statisticians and computer scientists. Hence, the *state-of-the-art* optimization methods go little beyond local search strategies.

The Hadamard transform (Hendy and Penny, 1989) circumvents many of these issues. Each tree is encoded as a vector called a *split vector* or *spectrum* $\mathbf{q}$. Hendy and Penny (1989) and Hendy (1989) showed that there is a general formula taking a spectrum to the vector of site pattern probabilities for the tree:

$$\mathbf{p} = \mathbf{H}^{-1} \exp(\mathbf{Hq}). \tag{1}$$

E-mail address: d.bryant@auckland.ac.nz.

The matrix **H** is a Hadamard matrix; see Section 3.2. This formula works for all trees, and once the tree has been coded as a vector, the tree structure plays no further part in the computation. Thus, the Hadamard transform provides a model into which all tree-based models naturally nest. Refer to Felsenstein (2004), Swofford et al. (1996) for excellent introductions to Hadamard transform methods.

The transform is a useful tool for many theoretical investigations. However, there are also important practical advantages of the approach. For example:

1. It provides a way of searching (or potentially, sampling) tree space that does not involve passing from individual tree to individual tree. One can invert the Hadamard transform to construct a spectrum **q** from the observed pattern probabilities and then use **q** to infer a tree.
2. Because the transform provides a model that generalizes trees it can be used to test the hypothesis "does this data actually come from a tree?" An analysis of this sort does not need to be hypothesis driven: phylogenetic network software like SplitsTree (Huson and Bryant, 2006) and Spectronet (Huber et al., 2002) allow one to visualize a spectrum **q** and see where it violates tree based models.

There have been many reformulations of the original Hadamard transform formula, each leading to a slightly different proof of the same result. Early correctness proofs of the transform were based on an interpretation in terms of path sets (Hendy, 1989; Hendy and Penny, 1989); these were recently extended in Hendy and Snir (2008).

The Hadamard transform can be viewed as an example of an Fourier transform on Abelian groups (Evans and Speed, 1993; Steel et al., 1992; Székely et al., 1993a, 1993b). Bryant (2005) used this algebraic machinery to show that the transform can be understood in terms of evolutionary models on phylogenetic networks. Sturmfels and Sullivant (2005) view the transform as a *change of coordinates*, and like Evans and Speed (1993), use it to study invariants on the phylogenetic tree models.

The Hadamard transform was translated into the language of quantum mechanics by Bashford et al. (2004), who showed how the formula follows from properties of Lie symmetries. They start by examining the standard K3ST model for nucleotide substitutions, and then show the process for a single taxon can be extended to a process for multiple taxa building on connections established by Jarvis and Bashford (2001).

Very recently, the Hadamard transform has reappeared in yet another guise. Klaere et al. (2008) derived a process equivalent to the n-taxon process and used it study the distribution of the number of mutations required to generate a single site in the alignment.

There are two important limitations of the Hadamard transform. The first is the running time: A full Hadamard transform takes exponential time and current analyses are limited to a maximum of 30 taxa. This could be remedied using approximations (such as distance based methods like NeighborNet, Bryant and Moulton, 2004) or Monte Carlo strategies.

The second limitation is the restriction on the substitution models with the Hadamard transform to *group-based* models. If we assume that the model is also time reversible, then the only available models are special cases of the K3ST model (see Section 2.1 and Bryant, 2005). This restriction is probably the most important barrier to widespread use of the transform and was one of the motivations behind the reformulation of the transform outlined in this paper. The problem of how to remove this restriction is still open.

1.2. Contribution of this paper

In this paper, we derive a new formulation of the Hadamard transform. Our proof that the transform works does not use path sets or Fourier transforms, at least not explicitly. The transform is established using basic matrix analysis that does not go much beyond the tools that are routinely used in phylogenetic analysis.

The key idea in the current paper is the *n-taxon process*. Consider a phylogenetic tree with times marked. At any particular time, every taxon has a unique ancestral lineage and this lineage has a unique state. Let $\mathbf{v}_t$ denote the vector of ancestral states, so that $\mathbf{v}_t[i]$ is the state of the ancestor of taxa i at time t. The n-taxon process is the continuous time Markov chain that describes the evolution of these vectors over time. This process is similar to the phylogenetic branching process described in Jarvis and Bashford (2001) and in Bashford et al. (2004), though we will not be exploiting the links with quantum multibody theory that these papers have established.

We derive the transition probability matrix $\mathbf{P}[\tau]$ for this process for a given vector τ of branch lengths. It is simply the exponential $\exp(\mathbf{Q}[\tau])$ of a linear combination of rate matrices for the branches. The formula for the Hadamard conjugation falls straight out of the formula for $\mathbf{P}[\tau]$, recast in appropriate notation. In fact, the vector $\mathbf{q}$ as defined in Hendy and Penny (1989) is simply the first column of the rate matrix $\mathbf{Q}[\tau]$. In this way, we obtain a proof for the correctness of the Hadamard transform that involves little more machinery than matrix exponentials.

The entry $\mathbf{P}[\tau]_{\mathbf{uv}}$ gives the probability that the final state is $\mathbf{v}$ (these are the observed states for each taxa) given that the initial vector of states is $\mathbf{u}$. In the standard model for phylogenetic analysis, the root is ancestral to all taxa so all of the entries in the initial vector would have the same state. This may not apply to problems from population genetics.

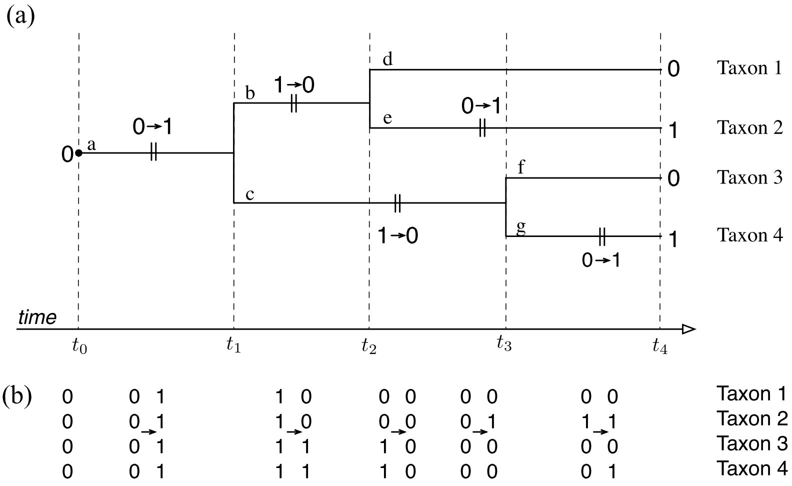
In this paper, we examine only two models, the binary symmetric model and the K3ST model. In practice, these are the only two models used with the Hadamard conjugation. The results here could be generalized to general group based models (Evans and Speed, 1993; Székely et al., 1993b; Bryant, 2005).

2. From tree based models to the n-taxon process

2.1. Tree based models

We begin with a brief outline of the standard models used for sequence evolution; for more details and extensive references, see Bryant et al. (2005), Felsenstein (2004), Swofford et al. (1996).

We assume that different sites in a sequence evolve independently from each other so we need only consider the evolution of a single site. A state is drawn at the root from a fixed distribution π. The evolution of the site then proceeds along the branches from the root to the leaves of the tree. Along each branch, substitutions occur according to a continuous time Markov chain as specified by its (*instantaneous*) *rate matrix* Q.



0	0 1	1 0	0 0	0 0	0 0	Taxon 1
0	0 1	1 0	0 0	0 1	1 1	Taxon 2
0	0 1	1 1	1 0	0 0	0 0	Taxon 3
0	0 1	1 1	1 0	0 0	0 1	Taxon 4

Fig. 1 (a) An example of state evolution on a tree under the binary symmetric model. The horizontal axis is proportional to time. The state 0 was drawn at the root, and substitutions occurred on edges a, b, c, e, g, giving the pattern 0101 at the leaves. (b) The corresponding n-taxon process. At each stage, the value of the process gives the ancestral states for each taxon. This changes five times from left to right.

We consider only two choices for Q; one for the *binary symmetric model* and the other for the *K3ST* model. The respective rate matrices are

$$\begin{aligned} Q_{(2)} &= \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \\ Q_{(4)} &= \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & -\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & -\alpha-\beta-\gamma \end{pmatrix}. \end{aligned} \qquad (2)$$

The values α, β, γ are positive reals chosen so that $\alpha+\beta+\gamma=1$. These parameters control the rates of three types of substitution as represented in Fig. 2. Type I substitutions correspond to DNA transitions; types II and III are types of transversion. Different submodels are obtained by making different rates equal to each other.

Both the binary symmetric model and the K3ST model have uniform stationary distributions. This is used for the distribution π at the root.

Example 1. We illustrate the binary symmetric model on a four taxa tree in Fig. 1(a). The root distribution is uniform: in this case 0 was drawn (with probability $1/2$). Substitutions occurred on edges a, b, c, e, g, giving the pattern 0101 at the leaves.

Let $P_{ij}(t)$ denote the probability that the state at the end of a branch of length t is j given that the state at the beginning of the branch is i. These *transition probabilities* are

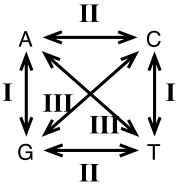


Fig. 2 The three transition types under the K3ST model.

given by the matrix exponential

$$P(t) = \exp(Qt) = I + Qt + Q^2\frac{t^2}{2!} + Q^3\frac{t^3}{3!} + \cdots. \tag{3}$$

The standard technique for computing this exponential (at least in phylogenetics) is to first diagonalize the matrix Q as

$$Q = V\Lambda V^{-1} \tag{4}$$

where Λ is a diagonal matrix, and then use the identity

$$\exp(Qt) = V\exp(\Lambda t)V^{-1}, \tag{5}$$

noting that $\exp(\Lambda t)$ is a diagonal matrix with values $\exp(\Lambda_{ii}t)$ down the diagonal.

Example 2. In the binary symmetric model, $Q = \left(\begin{smallmatrix}-1 & 1\\ 1 & -1\end{smallmatrix}\right)$. We have

$$Q = \begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}0 & 0\\ 0 & -2\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}^{-1}$$

giving

$$P(t) = \begin{pmatrix}\frac{1}{2}+\frac{1}{2}e^{-2t} & \frac{1}{2}-\frac{1}{2}e^{-2t}\\ \frac{1}{2}-\frac{1}{2}e^{-2t} & \frac{1}{2}+\frac{1}{2}e^{-2t}\end{pmatrix}. \tag{6}$$

These probabilities are often presented with t scaled by a constant μ. Here, we assume that time has been scaled so that $\mu = 1$.

2.2. *The n-taxon process*

We introduce an alternative way of describing the evolution of sequences along a tree. The main advantage of the approach is that the dependence on tree structure, and all of the complications that it introduces can be encoded away.

We will be working with a continuous time Markov chain on a much larger state space. Suppose that we have n taxa. Consider the set of all vectors assigning a state (e.g., a binary value or a nucleotide) to every taxa. In the binary case, there are 2^n of these vectors; in the 4-state case, there are 4^n. The set of these vectors will be the state space of our process, however, we will refer to these vectors as *values* of the process in order to minimize confusion with binary or nucleotide states.

Consider again the tree based model. At a particular time t, let $\mathbf{v}_t$ denote the vector of states of the *ancestors* for each of the n taxa. The *n-taxa process* is the continuous time Markov chain describing the evolution of these vectors.

Example 3. In Fig. 1, the initial value of the process is that the ancestors of all taxa have state 0. Hence, $\mathbf{v}_{t_0} = [0, 0, 0, 0]$. Between time t_0 and t_1, there is a substitution (on branch a). This occurs in the population that is ancestral to all of the taxa, so the effect is to change the value of the n-taxa process to $\mathbf{v}_{t_1} = [1, 1, 1, 1]$. On branch b, there is another substitution, though this is ancestral only to taxa 1 and 2. Thus, at time t_2, we have $\mathbf{v}_{t_2} = [0, 0, 1, 1]$. The process continues, until at the present time (t_4), the value of the process equals the observed states.

In a (slight) abuse of terminology, we will say that a taxon is a *descendant* of a branch if it is a descendant of the population/ancestors represented by that branch. For example, in Fig. 1, taxa 3 and 4 are the descendants of branch c.

The n-taxon process is a continuous time Markov chain, but it is not time-homogeneous. The transition probabilities depend on which lineages at a particular time are ancestral to which taxa. In the example, the rates of substitution for the n-taxa process will change at the time points $t_1, t_2, \ldots$. The process is homogeneous during the intervals between these time points. In what follows, we derive the rate matrices for the n-taxon process over each time interval.

We make the following notational conventions to lessen confusion between the different rate matrices involved.

1. We use Q to denote the rate matrix for the underlying (binary symmetric or K3ST) substitution process (binary symmetric or K3ST).
2. We use $\mathbf{Q}^{(i)}$ to denote the rate matrix for the n-taxon process during time interval $[t_{i-1}, t_i]$.
3. We use $\mathbf{R}^{(b)}$ to denote the rate matrix for the n-taxon process restricted to substitutions occurring on a given branch b.

We make extensive use of the *Kronecker product* of matrices. Given an $m \times n$ matrix X and a $p \times q$ matrix Y, the Kronecker product (or *tensor product*) of X and Y is the $mp \times nq$ matrix

$$X \otimes Y = \begin{pmatrix} X_{11}Y & X_{12}Y & \cdots & X_{1n}Y \\ X_{21}Y & X_{22}Y & \cdots & X_{2n}Y \\ \vdots & & \ddots & \vdots \\ X_{m1}Y & \cdots & \cdots & X_{mn}Y \end{pmatrix}. \tag{7}$$

The elements of a Kronecker product can be indexed by vectors so, for example,

$$(X \otimes Y)_{[i,p],[j,q]} = X_{ij}Y_{pq}, \qquad (X \otimes Y \otimes Z)_{[i,p,s],[j,q,t]} = X_{ij}Y_{pq}Z_{st}. \tag{8}$$

See Horn and Johnson (1994) for a detailed introduction to the Kronecker transform. We will make use of the following properties.

Lemma 4. *Let W, X, Y, Z be matrices with appropriate dimensions.*

1. $(X \otimes Y) \otimes Z = X \otimes (Y \otimes Z)$.
2. $(X \otimes Y)(W \otimes Z) = XW \otimes YZ$.
3. *Suppose that X, Y are non-zero. Then $X \otimes Y$ is diagonal if and only if X and Y are diagonal.*

We print matrices formed from Kronecker products in boldface.

3. Transition probabilities for the n-taxon process: binary symmetric case

3.1. Substitutions down a single branch: rate matrix

For the moment, consider only substitutions that occur along one particular branch b in the tree. We ignore substitutions occurring at the same time along other branches. Substitutions occur along branch b at rate 1. These substitutions affect only the taxa that are descendants of the branch b; let A be that set of taxa. A substitution along the branch corresponds to flipping the entries $\mathbf{v}_t[i]$ for which $i \in A$. These are the only substitutions in the n-taxon process restricted to the branch, and these substitutions occur at rate 1. Thus, the rate matrix $\mathbf{R}^{(b)}$ of this restricted process is given by

$$\mathbf{R}^{(b)}_{\mathbf{uv}} = \begin{cases} 1 & \text{if } \mathbf{u}[i] \neq \mathbf{v}[i] \text{ exactly when } i \in A; \\ -1 & \text{if } \mathbf{u} = \mathbf{v}; \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Example 5. In the example in Fig. 1, we have $Q = \left(\begin{smallmatrix} -1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. All elements of $\mathbf{R}^{(b)}$ are zero, except the diagonal elements (all -1) and the elements

$$\mathbf{R}^{(b)}_{[0,0,0,0],[1,1,0,0]}, \mathbf{R}^{(b)}_{[0,0,0,1],[1,1,0,1]}, \ldots, \mathbf{R}^{(b)}_{[1,1,1,1],[0,0,1,1]}, \tag{10}$$

$$\mathbf{R}^{(b)}_{[1,1,0,0],[0,0,0,0],}, \mathbf{R}^{(b)}_{,[1,1,0,1],[0,0,0,1]}, \ldots, \mathbf{R}^{(b)}_{[0,0,1,1],[1,1,1,1]}, \tag{11}$$

which all equal 1.

We reexpress $\mathbf{R}^{(b)}$ in terms of the Kronecker product of simple 2×2 matrices. Define $E = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

Lemma 6. *Let A be the set of taxa that are descendants of the population represented by branch b. For each $i = 1, 2, \ldots, n$ set $M^{(i)} = E$ if $i \in A$ and $M^{(i)} = I$ otherwise. Then*

$$\mathbf{R}^{(b)} = M^{(1)} \otimes M^{(2)} \otimes \cdots \otimes M^{(n)} - \mathbf{I}. \tag{12}$$

Example 7. Consider branch c in Fig. 1. As $A = \{3, 4\}$, we have

$$\mathbf{R}^{(c)} = I \otimes I \otimes E \otimes E - \mathbf{I}. \tag{13}$$

3.2. Substitutions down a single branch: transition probabilities

In this section, we show how to diagonalize the matrices $\mathbf{R}^{(b)}$. One of the attractions of the Kronecker product is that we can generally obtain a diagonalization of the product matrix in terms of its factors.

Let $H := H^{(1)} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and $\Lambda = HEH^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Thus HEH^{-1} and HIH^{-1} are both diagonal. We use Kronecker product to construct a matrix that diagonalises $\mathbf{R}^{(b)}$.

The nth order *Hadamard* matrix is defined as

$$\mathbf{H}^{(n)} = H \otimes H \otimes \cdots \otimes H, \tag{14}$$

an n-fold Kronecker product. Note that $(\mathbf{H}^{(n)})^{-1} = 2^{-n}\mathbf{H}^{(n)}$.

Lemma 8. *Let* $\mathbf{H} = \mathbf{H}^{(n)}$, *the nth order Hadamard matrix, and let* $\mathbf{R}^{b}$ *be the rate matrix for the n-taxon process restricted to branch b, binary symmetric case. Let A be the set of taxa that are descendants of branch b. Then*

$$\mathbf{\Lambda}^{(b)} := \mathbf{H}\mathbf{R}^{(b)}\mathbf{H}^{-1} \tag{15}$$

is a diagonal matrix with

$$\mathbf{\Lambda}^{(b)}_{\mathbf{uu}} = (-1)^{|\{i \in A : \mathbf{u}[i]=1\}|} - 1 \tag{16}$$

for all state vectors $\mathbf{u}$.

Proof: Define matrices $M^{(i)}$ as in Lemma 6. Then

$$\mathbf{H}\big(\mathbf{R}^{(b)} + \mathbf{I}\big)\mathbf{H}^{-1} = \big(HM^{(1)}H^{-1}\big) \otimes \cdots \otimes \big(HM^{(n)}H^{-1}\big). \tag{17}$$

If $i \in A$, then $HM^{(i)}H^{-1} = HEH^{-1} = \Lambda$ while if $i \notin A$, we have $HM^{(i)}H^{-1} = I$. The Kronecker product of diagonal matrices is diagonal, so $\mathbf{\Lambda}$ is diagonal.

For the diagonal values, note that

$$\mathbf{\Lambda}^{(b)}_{\mathbf{uu}} = \prod_{i=1}^{n} \big(HM^{(i)}H^{-1}\big)_{\mathbf{u}[i]\mathbf{u}[i]} \tag{18}$$

and that

$$\big(HM^{(i)}H^{-1}\big)_{\mathbf{u}[i]\mathbf{u}[i]} = \begin{cases} -1 & \text{if } i \in A \text{ and } u[i] = 1; \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

□

The transition probabilities down branch b now follow directly from the diagonalization, since

$$\exp\big(\mathbf{R}^{(b)}t\big) = \mathbf{H}^{-1}\exp\big(\mathbf{\Lambda}^{(b)}\big)\mathbf{H} \tag{20}$$

and $\exp(\mathbf{\Lambda}^{(b)})$ is a diagonal matrix with entries $\exp(\mathbf{\Lambda}^{(b)}_{\mathbf{uu}})$ down the diagonal.

3.3. Transition probabilities over multiple lineages

During the intervals between time points, there will be, in general, several lineages evolving independently. Because of this independence, the rate matrix $\mathbf{Q}^{(i)}$ for the substitution process over all lineages is simply the sum of the rate matrices $\mathbf{R}^{(b)}$ for the individual branches present at that time point.

Example 9. In Fig. 1, the rate matrix for the n-taxon process between t_2 and t_3 equals

$$\mathbf{Q}^{(3)} = \mathbf{R}^{(d)} + \mathbf{R}^{(e)} + \mathbf{R}^{(c)} \tag{21}$$

as branches d, e, c are present during this interval.

Between time points t_0 and the present time t_k, we therefore have a sequence of rate matrices $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots, \mathbf{Q}^{(k)}$. Each rate matrix $\mathbf{Q}^{(i)}$ equals the sum of the rate matrices $\mathbf{R}^{(b)}$ for all branches b present during the interval $[t_{i-1}, t_i]$. During each interval, the probability transitions are given by the standard exponential formula

$$\mathbf{P}^{(i)} = \exp\bigl(\mathbf{Q}^{(i)}(t_i - t_{i-1})\bigr), \tag{22}$$

so the transition probabilities between time t_0 and time t_k are given by

$$\begin{aligned} \mathbf{P} &= \mathbf{P}^{(1)}\mathbf{P}^{(2)} \cdots \mathbf{P}^{(k)} \\ &= e^{\mathbf{Q}^{(1)}(t_1-t_0)} e^{\mathbf{Q}^{(2)}(t_2-t_1)} \cdots e^{\mathbf{Q}^{(k)}(t_k-t_{k-1})}. \end{aligned} \tag{23}$$

Now comes a key step in the proof. The rate matrices $\mathbf{R}^{(b)}$ down each branch are all diagonalized by the Hadamard matrix $\mathbf{H}$ and, therefore, so are the sums $\mathbf{Q}^{(i)}$. Since every matrix $\mathbf{Q}^{(i)}$ in the product (23) is diagonalized by the same matrix $\mathbf{H}$, the rate matrices $\mathbf{Q}^{(i)}$ all commute. If two matrices $\mathbf{X}$ and $\mathbf{Y}$ commute, then $\exp(\mathbf{X})\exp(\mathbf{Y}) = \exp(\mathbf{X}+\mathbf{Y})$. Applying this identity to (23) gives

$$\mathbf{P} = \exp\left(\sum_{i=1}^{k} \mathbf{Q}^{(i)}(t_i - t_{i-1})\right). \tag{24}$$

Now examine the sum $\sum_{i=1}^{k} \mathbf{Q}^{(i)}(t_i - t_{i-1})$, a linear combination of the individual branch rate matrices $\mathbf{R}^{(b)}$. The coefficient of each matrix $\mathbf{R}^{(b)}$ is equal to the total length of time that the branch is present: the length of the branch. Let τ denote the vector of branch lengths. We have now established the following theorem.

Theorem 10. *Let* $\mathbf{P}[\tau]$ *be the matrix of transition probabilities in the n-taxon process for the binary symmetric case given a branch length vector τ. Define*

$$\mathbf{Q}[\tau] = \sum_{b} \mathbf{R}^{(b)}\tau_b \tag{25}$$

where b ranges over branches in the tree, $\mathbf{R}^{(b)}$ *is the matrix given in Lemma 6, and τ_b is the length of branch b. Then* $\mathbf{H}\mathbf{Q}[\tau]\mathbf{H}^{-1}$ *is a diagonal matrix and*

$$\mathbf{P}[\tau] = \exp\bigl(\mathbf{Q}[\tau]\bigr). \tag{26}$$

The probability distribution for a tree can be recovered from (26) by noting that at the root, the process is $\mathbf{0} = [0, 0, \ldots, 0]'$ with probability $\pi_0 = 1/2$ and $\mathbf{1} = [1, 1, \ldots, 1]'$ with probability $1/2$. If $\mathbf{u}$ is the pattern of states at the leaves, then the probability of observing $\mathbf{u}$ equals

$$\mathbf{p} = \frac{1}{2}\mathbf{P}_{\mathbf{0u}}[\tau] + \frac{1}{2}\mathbf{P}_{\mathbf{1u}}[\tau]. \tag{27}$$

Interestingly, (26) also applies to the case when there is not a single common ancestor for the taxa, a feature that may well prove useful in population genetics applications.

3.4. Recovering the Hadamard formula

The Hadamard conjugation formula (Hendy and Penny, 1989; Swofford et al., 1996) assumes that one taxon has all zero states and gives the probabilities for patterns on the remaining taxa. We can retrieve the formula almost directly from Theorem 10, giving a new proof for the Hadamard conjugation. This new derivation explains why the zero entry of the vector $\mathbf{q}$ in Hendy and Penny (1989) is chosen to make the sum of all entries zero: the vector $\mathbf{q}$ is simply a row out of the rate matrix $\mathbf{Q}$.

Theorem 11 (Hendy and Penny, 1989). *Suppose that the tree has taxa at the root with state* 0. *For each non-zero vector* $\mathbf{u}$ (*indexed by the remaining taxa*), *let* $\mathbf{q_u}$ *be the length of the branch with descendants* $\{i : \mathbf{u}[i] = 1\}$, *if there is such a branch in the tree*, *and zero otherwise. Let* $\mathbf{q_0}$ *be the negative of the sum of all the branch lengths in the tree. Let* $\mathbf{p_u}$ *be the probability of observing the pattern* $\mathbf{u}$ *at the leaves. Then*

$$\mathbf{p} = \mathbf{H}^{-1}\exp(\mathbf{Hq}). \tag{28}$$

Here the exponential is entry-wise.

Proof: Let $\mathbf{0}$ denote the vector $[0, 0, \ldots, 0]'$. Both $\mathbf{P}$ and $\mathbf{Q}$ are indexed by vectors of states: By $\mathbf{0}$-row or $\mathbf{0}$ column, we mean the row or column with index $\mathbf{0}$. We seek the probabilities $\mathbf{p_u} = \mathbf{P_{0u}}[\tau]$. As $\mathbf{P}[\tau]$ is symmetric, $\mathbf{P_{0u}}[\tau] = \mathbf{P_{u0}}[\tau]$.

The vector $\mathbf{q}$ is the $\mathbf{0}$-column of $\mathbf{Q}[\tau]$ Let $\mathbf{\Lambda} = \mathbf{HQ}[\tau]\mathbf{H}^{-1}$, so that $\mathbf{HQ} = \mathbf{\Lambda H}$. The $\mathbf{0}$-column of $\mathbf{H}$ is all ones, so the $\mathbf{0}$-column of $\mathbf{\Lambda H}$ is made up of the diagonal entries of $\mathbf{\Lambda}$. Hence, the entries in $\mathbf{Hq}$ are the entries along the diagonal of $\mathbf{\Lambda}$. Taking entry-wise exponentials, we have that $\exp(\mathbf{Hq})$ equals the entries along the diagonal of $\exp(\mathbf{\Lambda})$ and so $\exp(\mathbf{Hq})$ is the first column of $\exp(\mathbf{\Lambda})\mathbf{H}$. Hence, $\mathbf{H}^{-1}\exp(\mathbf{Hq})$ is the $\mathbf{0}$ column of $\mathbf{H}^{-1}\exp(\mathbf{\Lambda})\mathbf{H}$, which by Theorem 10 equals $\mathbf{P}$. □

4. Transition probabilities for the *n*-taxon process: K3ST model

We now extend the results of the previous sections to the K3ST model. In the interests of brevity, we only outline the key steps in the derivation.

4.1. Substitutions down a single branch: rate matrix

Under the K3ST model, there are three types of substitution. Instead of defining one matrix E as above, we define three matrices

$$E_I = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad E_{II} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$
$$E_{III} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \tag{29}$$

so that $Q = \alpha E_I + \beta E_{II} + \gamma E_{III} - (\alpha + \beta + \gamma)I$.

Lemma 12. *Let A be the set of taxa that are descendants of the population represented by branch b. For each $i = 1, 2, \ldots, n$ set $M_I^{(i)} = E_I$ if $i \in A$ and $M^{(i)} = I$ otherwise; likewise for $M_{II}^{(i)}$ and $M_{III}^{(i)}$. Then*

$$\mathbf{R}^{(b)} = \alpha M_I^{(1)} \otimes \cdots \otimes M_I^{(n)} + \beta M_{II}^{(1)} \otimes \cdots \otimes M_{II}^{(n)} + \gamma M_{III}^{(1)} \otimes \cdots \otimes M_{III}^{(n)} - \mathbf{I}. \tag{30}$$

The matrix $\mathbf{R}^{(b)}$ is indexed by vectors of states. We number the states $0, 1, 2, 3$ corresponding to A, C, G, T, respectively.

4.2. Substitutions down a single branch: transition probabilities

We use the same trick as before to diagonalize the rate matrix $\mathbf{R}^{(b)}$ in the K3ST case, using the properties of the Kronecker product. Let

$$H = H^{(2)} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \tag{31}$$

then define $\Lambda_I = H^{-1}E_I H$, $\Lambda_{II} = H^{-1}E_{II}H$ and $\Lambda_{III} = H^{-1}E_{III}H$. Then $\Lambda_I = \mathrm{diag}(1, -1, 1, -1)$, $\Lambda_{II} = \mathrm{diag}(1, 1, -1, -1)$, and $\Lambda_{III} = \mathrm{diag}(1, -1, -1, 1)$.

For this case, we let $\mathbf{H}$ denote the n-fold product $H \otimes H \otimes \cdots \otimes H$, which is equal to the $2n$th order Hadamard matrix.

Lemma 13. *Let $\mathbf{H} = \mathbf{H}^{(2n)}$, the $2n$th order Hadamard matrix, and let $\mathbf{R}^b$ be the rate matrix for the n-taxon process restricted to branch b, K3ST case. Let A be the set of taxa that are descendants of branch b. Then*

$$\mathbf{\Lambda}^{(b)} := \mathbf{H}\mathbf{R}^{(b)}\mathbf{H}^{-1} \tag{32}$$

is a diagonal matrix with

$$\begin{aligned}\mathbf{\Lambda}^{(b)}_{\mathbf{uu}} &= \alpha(-1)^{|\{i\in A:\mathbf{u}[i]=1 \text{ or } 3\}|} + \beta(-1)^{|\{i\in A:\mathbf{u}[i]=2 \text{ or } 3\}|}\\ &\quad + \gamma(-1)^{|\{i\in A:\mathbf{u}[i]=1 \text{ or } 2\}|} - 1 \end{aligned} \tag{33}$$

for all state vectors **u**.

The transition probabilities down branch b now follow directly from the diagonalization, since

$$\exp\bigl(\mathbf{R}^{(b)}t\bigr) = \mathbf{H}\exp\bigl(\mathbf{\Lambda}^{(b)}\bigr)\mathbf{H}^{-1} \tag{34}$$

and $\exp(\mathbf{\Lambda}^{(b)})$ is a diagonal matrix with entries $\exp(\mathbf{\Lambda}^{(b)}_{\mathbf{uu}})$ down the diagonal.

4.3. Transition probabilities over multiple lineages

The progression from rate matrices for branches to rate matrices for the entire n-taxon process is almost identical in the K3ST case as in the binary symmetric model case. During each time interval $[t_{i-1}, t_i]$, the rate matrix $\mathbf{Q}^{(i)}$ for the n-taxon process is the sum of the rate matrices $\mathbf{R}^{(b)}$ for branches present at that time. They are all diagonalized by the $2n$th order Hadamard matrix $\mathbf{H}$, so commute, and we have

$$\mathbf{P} = \exp\left(\sum_{i=1}^{k}\mathbf{Q}^{(i)}(t_i - t_{i-1})\right). \tag{35}$$

Furthermore, given the vector τ of branch lengths, we have

$$\sum_{i=1}^{k}\mathbf{Q}^{(i)}(t_i - t_{i-1}) = \sum_{b}\mathbf{R}^{(b)}\tau_b, \tag{36}$$

establishing the following analogue to Theorem 10.

Theorem 14. *Let* $\mathbf{P}[\tau]$ *be the matrix of transition probabilities in the* n*-taxon process for the binary symmetric case given a branch length vector* τ*. Define*

$$\mathbf{Q}[\tau] = \sum_{b}\mathbf{R}^{(b)}\tau_b \tag{37}$$

where b *ranges over branches in the tree,* $\mathbf{R}^{(b)}$ *is the matrix given in Lemma 12, and* τ_b *is the length of branch* b*. Then* $\mathbf{H}\mathbf{Q}[\tau]\mathbf{H}^{-1}$ *is a diagonal matrix and*

$$\mathbf{P}[\tau] = \exp\bigl(\mathbf{Q}[\tau]\bigr). \tag{38}$$

References

Bashford, J., Jarvis, P.D., Sumner, J., Steel, M.A., 2004. $U(1) \times U(1) \times U(1)$ symmetry of the Kimura 3ST model and phylogenetic branching processes. J. Phys. A Math. Gen. 37, 81–89.

Bryant, D., 2005. Extending tree models to split networks. In: Pachter, L., Sturmfels, B. (Eds.), Algebraic Statistics for Computational Biology, pp. 322–334. Cambridge University Press, Cambridge.

Bryant, D., Galtier, N., Poursat, M.-A., 2005. Likelihood calculations in molecular phylogenetics. In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny, pp. 33–62. Oxford University Press, London.

Bryant, D., Moulton, V., 2004. NeighborNet: An agglomerative algorithm for the construction of planar phylogenetic networks. Mol. Biol. Evol. 21, 255–265.

Evans, S.N., Speed, T.P., 1993. Invariants of some probability models used in phylogenetic inference. Ann. Stat. 21(1), 355–377.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer, Sunderland.

Hendy, M., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38(4), 297–309.

Hendy, M.D., 1989. The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. 38(4), 310–321.

Hendy, M.D., Snir, S., 2008. Hadamard conjugation for the Kimura 3st model: Combinatorial proof using path sets. Trans. Comput. Biol. Bioinform. 5(3), 461–471.

Horn, R.A., Johnson, C.R., 1994. Topics in Matrix Analysis. Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Huber, K.T., Langton, M., Penny, V., Moulton, D., Hendy, M., 2002. Spectronet: A package for computing spectra and median networks. Appl. Bioinform. 1(3), 2041–2059.

Huson, D., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Jarvis, P.D., Bashford, J., 2001. Quantum field theory and phylogenetic branching. J. Phys. A Math. Gen. 34, L703–707.

Klaere, S., Gesell, T., Haeseler, A.v, 2008, in press. The impact of single substitutions on multiple sequence alignments. Proc. R. Soc. Lond. B. 275.

Steel, M.A., Hendy, M.D., Székely, L.A., Erdős, P.L., 1992. Spectral analysis and a closest tree method for genetic sequences. Appl. Math. Lett. 5(6), 63–67.

Sturmfels, B., Sullivant, S., 2005. Toric ideals of phylogenetic invariants. J. Comput. Biol. 12(2), 204–228.

Swofford, D., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics, 2nd edn., pp. 407–514. Sinauer, Sunderland.

Székely, L.A., Erdős, P.L., Steel, M.A., Penny, D., 1993a. A Fourier inversion formula for evolutionary trees. Appl. Math. Lett. 6(2), 13–16.

Székely, L.A., Steel, M.A., Erdős, P.L., 1993b. Fourier calculus on evolutionary trees. Adv. Appl. Math. 14(2), 200–210.