Hadamard Conjugation for the Kimura 3ST Model: Combinatorial Proof Using Path Sets

Michael D. Hendy and Sagi Snir

Abstract—Under a stochastic model of molecular sequence evolution the probability of each possible pattern of characters is well defined. The Kimura's three-substitution-types (K3ST) model of evolution allows analytical expression for these probabilities by means of the Hadamard conjugation as a function of the phylogeny T and the substitution probabilities on each edge of T. In this paper, we produce a direct combinatorial proof of these results using path-set distances, which generalize pairwise distances between sequences. This interpretation provides us with tools that have proved useful in related problems in the mathematical analysis of sequence evolution.

Index Terms—Hadamard conjugation, K3ST model, path-sets, phylogenetic trees, phylogenetic invariants.

1 Introduction

Hadamard conjugation is an analytic formulation of the relationship between the probabilities of expected site patterns of nucleotides for a set of homologous nucleotide sequences and the parameters of some simple models of sequence evolution on a proposed phylogeny T. An important application of these relations is to give a theoretical tool to analyze properties of phylogenetic inference such as the methods of maximum likelihood and maximum parsimony, as well as for generating simulated data, and determining phylogenetic invariants. Hadamard conjugation can also be used directly for phylogenetic inference, inferring either trees with the Closest Tree algorithm [11], [25] or networks using Spectronet [18]. Application of the Hadamard conjugation in maximum likelihood phylogenetic inference under the Kimura's three-substitution-types (K3ST) model was done in [5] and in a related problem, where phylogenetic invariants were used to reconstruct quartet trees under a generalized variant of K3ST [4].

Hadamard conjugation was first introduced in 1989 [10], [13] to analyze two-state character sequences evolving under the Neyman model [22]. Evans and Speed [9] noted that K3ST model [21] for 4-state characters could be modeled by the Klein group $\mathbb{Z}_2 \times \mathbb{Z}_2$. Noting this, Székely et al. [28], [29] extended the two-state analysis to a more general algebraic theory, where substitutions belonged to an arbitrary Abelian group. They then applied this to sequences evolving under the K3ST model. Current applications of Closest Tree and Spectronet [18] are usually applied to the 4-state K3ST model or its derivatives, the K2ST and Jukes-Cantor models.

- *M.D. Hendy is with the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11222, Palmerston North 4410, New Zealand. E-mail: m.hendy@massey.ac.nz.*
- *S. Snir is with the Mathematics Department, University of California, Berkeley, Berkeley, CA 94720. E-mail: ssagi@math.berkeley.edu.*

Manuscript received 25 Oct. 2006; revised 30 Mar. 2007; accepted 13 May 2007; published online 27 June 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0196-1006.
Digital Object Identifier no. 10.1109/TCBB.2007.70227.

A path-set in a phylogenetic tree T is a generalization of the concept of a path. This approach allows the concept of pairwise distances between sequences to be extended to distances connecting larger sets of taxa. It provides properties that can be related to other evolutionary phenomena such as the molecular clock hypothesis. This has, for example, proved pivotal in allowing a simpler analytic expression of the likelihood function, as developed in [5], leading to an algebraic solution for the maximum likelihood points. We demonstrate this use, as well as the relation to the molecular clock property in our last section describing the application of the Hadamard conjugation, as was used in [5]. It has also proved useful in identifying phylogenetic invariants [15], [27], [4] and introducing the projected spectra [30], which reduces both the variance in the parameter estimates and the computational complexity of the Closest Tree algorithm [11]. All the above examples rely on some relationships between the phylogenetic tree and the probabilities of obtaining sequences evolved under that tree. These relationships were proved in the past by algebraic tools on more general model of evolution. However, on the K3ST, these relationship can be expressed as identities between expressions in the tree parameters and expressions in the sequence probabilities. These relationship were outlined in [26]. Here, we provide a self contained more rigorous proof that bears some resemblance (Section 8) to the sketch in [26]. However, that outline lacks the details of the combinatorial properties of the intermediate variables, which we find to be of interest. Therefore, our proof serves as a more intuitive alternative to the presentations in [16] and [26].

We model the relationship of the differences of n sequences labeled by elements of $[n] = \{1, 2, \cdots, n\}$, from a reference sequence labeled 0 (note that $0 \notin [n]$). Because the models are reversible, the choice of reference sequence is arbitrary. The topology of T and the model parameters are presented in a sparse matrix Q_T of 2^n rows and columns, called the edge-length spectrum. The probabilities of each site pattern are presented in a similar sized matrix S_T called the sequence probability

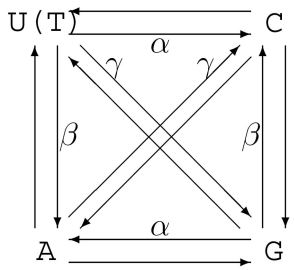


Fig. 1. K3ST, showing the three substitution types, α, β, and γ, applied either to the RNA nucleotides A, C, G, and U or to the DNA nucleotides A, C, G, and T.

spectrum. We also define a Hadamard matrix H_n of 2^n rows and columns and show that the matrix products:

$$H_n Q_T H_n, \quad H_n S_T H_n,$$

both relate to properties of path sets. We prove the major result by interpreting corresponding components of each entry of these matrices. In particular, we show that the (E, F) entry in both matrices corresponds to certain "evolutionary distances" defined by path sets E and F.

We note that we were motivated to provide this new proof as these variables served as the defining parameters in the likelihood equation in [5] while their biological interpretation has not been elaborated sufficiently. We start by describing the K3ST model over a single edge and then generalizing it to a set of edges in a tree. Next, we introduced the notion of a site pattern and the matrix S_T. In Sections 5 and 6, we introduce the Hadamard matrices and path sets and the relationship between them. Section 8 is the main part of the proof, where we show the relationship between corresponding entries of the matrices $H_n Q_T H_n$ and $H_n S_T H_n$. We end by describing the derivation of the equations used in [5] leading to an analytical solution of the ML problem.

We believe this is an important contribution that can serve in the burgeoning area of algebraic statistics in biology and phylogenetics, in particular (see, e.g., [1], [2], [3], [23], [24], and [27]).

2 Kimura's 3ST Model

In this section, we first describe the K3ST model. We then derive identities relating the substitution matrix M, and the matrix of expected numbers of substitution along each edge. Finally, we encode these relationship by means of two simpler matrices P and Q, and the Hadamard matrix H_1.

K3ST [21] specified independent rates for each of the substitutions between pairs of RNA or DNA nucleotides. Here, we will refer to Kimura's three substitution rates as λ_α, λ_β and λ_γ, and use α, β, and γ to refer to the substitution types, as illustrated in Fig. 1. These are defined formally as

- α. The substitutions $\mathtt{A} \leftrightarrow \mathtt{G}$, $\mathtt{U(T)} \leftrightarrow \mathtt{C}$ (transitions).
- β. The substitutions $\mathtt{A} \leftrightarrow \mathtt{U(T)}$, $\mathtt{G} \leftrightarrow \mathtt{C}$ (transversions type β).
- γ. The substitutions $\mathtt{A} \leftrightarrow \mathtt{C}$, $\mathtt{U(T)} \leftrightarrow \mathtt{G}$ (transversions type γ).

By including the identity transformation ϵ, we find that the set of substitution types:

$$\mathcal{T} = \{\epsilon, \alpha, \beta, \gamma\}$$

is a group under composition, acting on the nucleotide set $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U(T)}\}$. Thus, for example, $\alpha(\beta(\mathtt{C})) = \alpha(\mathtt{G}) = \mathtt{A} = \gamma(\mathtt{C})$, so $\alpha \circ \beta = \gamma$.

Consider the maps g_1, $g_2 : \mathcal{T} \to C_2 = \{1, -1\}$ defined by

$$\begin{array}{lllll} g_1 & : & \epsilon \mapsto 1, & \alpha \mapsto 1, & \beta \mapsto -1, \quad \gamma \mapsto -1; \\ g_2 & : & \epsilon \mapsto 1, & \alpha \mapsto -1, & \beta \mapsto 1, \quad \gamma \mapsto -1. \end{array} \tag{1}$$

We find that g_1 and g_2 are both homomorphisms from $(\mathcal{T}, \circ)$ onto the 2-group $(C_2, \times)$, and the map

$$g : \theta \mapsto (g_1(\theta), g_2(\theta)), \qquad \theta \in \mathcal{T},$$

is an isomorphism onto the group $(C_2 \times C_2, \times)$.

Observation 1. *$(\mathcal{T}, \circ)$ is isomorphic to the Klein 4-group, $(C_2 \times C_2, \times)$.*

In contrast, the set of substitutions of the K2ST model and of the Jukes-Cantor model do not form groups, as products are not well defined (for example, a product of two transversions in K2ST could either be a transition or the identity).

A related, however, different aspect is the property of generalization/specialization between models. For example, we can specialize from K3ST down to each of these models by imposing restrictions on parameters (for example, if the expected numbers of transitions and of transversions of each type are equated, then K3ST specializes to the Jukes-Cantor model). A different restriction on the values of the model parameters is imposed by the Molecular Clock constraint, however, this is beyond the scope of this work (see, e.g., [17] and [6]).

Kimura modeled the expected differences between two sequences separated by time t. With the three specified rates, the expected numbers of substitutions of each type are therefore

$$q(\alpha) = \lambda_\alpha t, \quad q(\beta) = \lambda_\beta t, \quad q(\gamma) = \lambda_\gamma t.$$

By setting $\lambda_\beta = \lambda_\gamma$, this model projects to K2ST, Kimura's better known two substitution type model [20]. Setting $\lambda_\alpha = \lambda_\beta = \lambda_\gamma$ gives the simple Jukes-Cantor model [19].

The probabilities $p(\alpha)$, $p(\beta)$, and $p(\gamma)$ of observing differences of each type over the time period t underestimate $q(\alpha)$, $q(\beta)$, and $q(\gamma)$, as multiple changes are not directly observed. Observed frequencies of differences can be taken as estimates of $p(\theta)$ for $\theta \in \{\alpha, \beta, \gamma\}$.

In [21], Kimura derived expressions for the expected numbers $q(\theta)$ as functions of the probabilities $p(\theta)$. These are equivalent to the standard expression of the stochastic matrix M, derived from the rate matrix

$$R = \begin{bmatrix} -(\lambda_\alpha+\lambda_\beta+\lambda_\gamma) & \lambda_\alpha & \lambda_\beta & \lambda_\gamma \\ \lambda_\alpha & -(\lambda_\alpha+\lambda_\beta+\lambda_\gamma) & \lambda_\gamma & \lambda_\beta \\ \lambda_\beta & \lambda_\gamma & -(\lambda_\alpha+\lambda_\beta+\lambda_\gamma) & \lambda_\alpha \\ \lambda_\gamma & \lambda_\beta & \lambda_\alpha & -(\lambda_\alpha+\lambda_\beta+\lambda_\gamma) \end{bmatrix},$$

over time t, so that (e.g., see [26])

$$M = \exp(Rt), \tag{2}$$

where

$$M = \begin{bmatrix} p(\epsilon) & p(\alpha) & p(\beta) & p(\gamma) \\ p(\alpha) & p(\epsilon) & p(\gamma) & p(\beta) \\ p(\beta) & p(\gamma) & p(\epsilon) & p(\alpha) \\ p(\gamma) & p(\beta) & p(\alpha) & p(\epsilon) \end{bmatrix},$$

$$Rt = \begin{bmatrix} -K & q(\alpha) & q(\beta) & q(\gamma) \\ q(\alpha) & -K & q(\gamma) & q(\beta) \\ q(\beta) & q(\gamma) & -K & q(\alpha) \\ q(\gamma) & q(\beta) & q(\alpha) & -K \end{bmatrix}.$$

$K = q(\alpha) + q(\beta) + q(\gamma)$ is the total number of substitutions, and exp is the standard exponential function for square matrices. We note that as R and t are fixed, so too are M, p, q, and K as they are defined by them.

Let H_2 be the 4×4 Hadamard matrix:

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Observation 2. *H_2 diagonalizes both M and Rt. In particular,*

$$H_2^{-1} M H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2p(\alpha)-2p(\gamma) & 0 & 0 \\ 0 & 0 & 1-2p(\beta)-2p(\gamma) & 0 \\ 0 & 0 & 0 & 1-2p(\alpha)-2p(\beta) \end{bmatrix},$$

and

$$H_2^{-1} Rt H_2 = -2 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & q(\alpha)+q(\gamma) & 0 & 0 \\ 0 & 0 & q(\beta)+q(\gamma) & 0 \\ 0 & 0 & 0 & q(\alpha)+q(\beta) \end{bmatrix}.$$

Recall the exponential of a matrix is a power series, so

$$\exp(H_2^{-1} Rt H_2) = \sum_{n \geq 0} \frac{(H_2^{-1} Rt H_2)^n}{n!} = \sum_{n \geq 0} H_2^{-1} \frac{(Rt)^n}{n!} H_2 = H_2^{-1} \exp(Rt) H_2.$$

As $\exp(H_2^{-1} Rt H_2)$ is diagonal, so too is $H_2^{-1} \exp(Rt) H_2$, with entries

$$1 = e^0, \quad e^{-2(q(\alpha)+q(\gamma))}, \quad e^{-2(q(\beta)+q(\gamma))}, \quad e^{-2(q(\alpha)+q(\beta))}.$$

Now, using (2), we observe

$$H_2^{-1} M H_2 = H_2^{-1} (\exp(Rt)) H_2.$$

Equating the diagonal entries shows that the eigenvalues of M and $\exp(Rt)$ are

$$\begin{aligned} 1 = p(\epsilon)+p(\alpha)+p(\beta)+p(\gamma) &= e^0 &&= e^{-K+q(\alpha)+q(\beta)+q(\gamma)}, \\ 1-2(p(\alpha)+p(\gamma)) = p(\epsilon)-p(\alpha)+p(\beta)-p(\gamma) &= e^{-2(q(\alpha)+q(\gamma))} &&= e^{-K-q(\alpha)+q(\beta)-q(\gamma)}, \\ 1-2(p(\beta)+p(\gamma)) = p(\epsilon)+p(\alpha)-p(\beta)-p(\gamma) &= e^{-2(q(\beta)+q(\gamma))} &&= e^{-K+q(\alpha)-q(\beta)-q(\gamma)}, \\ 1-2(p(\alpha)+p(\beta)) = p(\epsilon)-p(\alpha)-p(\beta)+p(\gamma) &= e^{-2(q(\alpha)+q(\beta))} &&= e^{-K-q(\alpha)-q(\beta)+q(\gamma)}. \end{aligned}$$

These equations can be succinctly expressed (see [12]) as

$$H_1 P H_1 = \mathrm{Exp}(H_1 Q H_1), \tag{3}$$

where

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad P = \begin{bmatrix} p(\epsilon) & p(\alpha) \\ p(\beta) & p(\gamma) \end{bmatrix}, \quad Q = \begin{bmatrix} -K & q(\alpha) \\ q(\beta) & q(\gamma) \end{bmatrix},$$

and Exp is the exponential function applied to each entry of the matrix. Equation (3) can be inverted (as the arguments of ln are all positive) to give

$$H_1 Q H_1 = \mathrm{Ln}(H_1 P H_1), \tag{4}$$

where Ln is the natural logarithm applied to each entry of a matrix.

The invertibility of (3) and (4) means that provided the parameters are in valid ranges, the model could be specified either by the three probabilities $p(\alpha)$, $p(\beta)$, and $p(\gamma)$, or by the three parameters $q(\alpha)$, $q(\beta)$, and $q(\gamma)$. This inversion does not rely on a rate/time specification and a Poisson process of substitution. Hence, we are able to test the validity of a constant rate model in analyzing observed data.

3 Substitutions across the Edges of a Tree

We now extend the model of the previous section to handle sets of edges. We derive the probability of a substitution of type θ along a set of edges W and record it in a stochastic matrix M_W. We also define the path length matrix Q_W and by a similar fashion to that in the previous section, obtain the relationship between M_W and Q_W. Finally, we define the edge length spectrum that records all the tree parameters.

Let $[n] = \{1, 2, \ldots, n\}$ and $[n]_0 = [n] \cup \{0\}$. Let T be a tree (phylogeny) with leaf set $L(T) = [n]_0$ and edge set $e(T)$. For each edge $e \in e(T)$, we can postulate three independent Kimura probability parameters $p_e(\alpha)$, $p_e(\beta)$, and $p_e(\gamma)$. These are collected in a stochastic matrix:

$$M_e = \begin{bmatrix} p_e(\epsilon) & p_e(\alpha) & p_e(\beta) & p_e(\gamma) \\ p_e(\alpha) & p_e(\epsilon) & p_e(\gamma) & p_e(\beta) \\ p_e(\beta) & p_e(\gamma) & p_e(\epsilon) & p_e(\alpha) \\ p_e(\gamma) & p_e(\beta) & p_e(\alpha) & p_e(\epsilon) \end{bmatrix},$$

with eigenvalues 1, $\exp(-2(q_e(\alpha) + q_e(\gamma)))$, $\exp(-2(q_e(\beta) + q_e(\gamma)))$, and $\exp(-2(q_e(\alpha) + q_e(\beta)))$. Let

$$P_e = \begin{bmatrix} p_e(\epsilon) & p_e(\alpha) \\ p_e(\beta) & p_e(\gamma) \end{bmatrix}, \quad Q_e = \begin{bmatrix} -K_e & q_e(\alpha) \\ q_e(\beta) & q_e(\gamma) \end{bmatrix},$$

then by (3), we see that the probabilities $p_e(\alpha)$, $p_e(\beta)$, and $p_e(\gamma)$ are related to the three parameters $q_e(\alpha)$, $q_e(\beta)$, and $q_e(\gamma)$ as

$$H_1 P_e H_1 = \mathrm{Exp}(H_1 Q_e H_1).$$

As the matrices M_e, for $e \in e(T)$, are each diagonalized by H_2, they commute. Hence, for any subset of edges $W \subseteq e(T)$, we can formally define the product:

$$M_W = \prod_{e \in W} M_e = \begin{bmatrix} p_W(\epsilon) & p_W(\alpha) & p_W(\beta) & p_W(\gamma) \\ p_W(\alpha) & p_W(\epsilon) & p_W(\gamma) & p_W(\beta) \\ p_W(\beta) & p_W(\gamma) & p_W(\epsilon) & p_W(\alpha) \\ p_W(\gamma) & p_W(\beta) & p_W(\alpha) & p_W(\epsilon) \end{bmatrix}. \quad (5)$$

We note that the term $p_W(\theta)$ is the probability that the product of the substitutions of all the edges of W is θ. In particular, if W is a path in T, then $p_W(\theta)$ is the probability that the states at the endpoints of the path differ by the substitution θ. In addition, when $W = \{e\}$, we see $M_{\{e\}} = M_e$ and $p_{\{e\}}(\theta) = p_e(\theta)$.

We see that M_W is diagonalized by H_2:

$$H_2^{-1} M_W H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2(p_W(\alpha)+p_W(\gamma)) & 0 & 0 \\ 0 & 0 & 1-2(p_W(\beta)+p_W(\gamma)) & 0 \\ 0 & 0 & 0 & 1-2(p_W(\alpha)+p_W(\beta)) \end{bmatrix},$$

as is each factor in (5), so

$$H_2^{-1} M_W H_2 = H_2^{-1} \left(\prod_{e \in W} M_e \right) H_2 = \prod_{e \in W} H_2^{-1} M_e H_2 = \prod_{e \in W} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \exp(-2(q_e(\alpha)+q_e(\gamma))) & 0 & 0 \\ 0 & 0 & \exp(-2(q_e(\beta)+q_e(\gamma))) & 0 \\ 0 & 0 & 0 & \exp(-2(q_e(\alpha)+q_e(\beta))) \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \exp(-2((q_W(\alpha)+q_W(\gamma))) & 0 & 0 \\ 0 & 0 & \exp(-2(q_W(\beta)+q_W(\gamma))) & 0 \\ 0 & 0 & 0 & \exp(-2(q_W(\alpha)+q_W(\beta))) \end{bmatrix},$$

where

$$q_W(\alpha) = \sum_{e \in W} q_e(\alpha), \quad q_W(\beta) = \sum_{e \in W} q_e(\beta), \quad q_W(\gamma) = \sum_{e \in W} q_e(\gamma).$$

Thus, equating the diagonals, we have

Observation 3.

$$\begin{aligned} 1 &= p_W(\epsilon)+p_W(\alpha)+p_W(\beta)+p_W(\gamma) = \exp(0), \\ 1-2(p_W(\alpha)+p_W(\gamma)) &= p_W(\epsilon)-p_W(\alpha)+p_W(\beta)-p_W(\gamma) = \exp(-2(q_W(\alpha)+q_W(\gamma))), \\ 1-2(p_W(\beta)+p_W(\gamma)) &= p_W(\epsilon)+p_W(\alpha)-p_W(\beta)-p_W(\gamma) = \exp(-2(q_W(\beta)+q_W(\gamma))), \\ 1-2(p_W(\alpha)+p_W(\beta)) &= p_W(\epsilon)-p_W(\alpha)-p_W(\beta)+p_W(\gamma) = \exp(-2(q_W(\alpha)+q_W(\beta))). \end{aligned}$$

We now define the corresponding P and Q matrices for the edge set W:

$$P_W = \begin{bmatrix} p_W(\epsilon) & p_W(\alpha) \\ p_W(\beta) & p_W(\gamma) \end{bmatrix},$$
$$Q_W = \begin{bmatrix} -K_W & q_W(\alpha) \\ q_W(\beta) & q_W(\gamma) \end{bmatrix} = \sum_{e \in W} Q_e,$$

where $K_W = q_W(\alpha) + q_W(\beta) + q_W(\gamma)$. The relationships of Observation 3, similar to (3), can now be expressed as

$$H_1 P_W H_1 = \mathrm{Exp}(H_1 Q_W H_1). \quad (6)$$

As the Q_e matrices are additive over edge sets of T, we refer to the expected numbers $q_e(\alpha)$, $q_e(\beta)$, and $q_e(\gamma)$ as the three **edge-length** parameters, for each edge $e \in e(T)$. We can thus specify our model by the set of $3|e(T)|$ independent edge-length parameters

$$\{q_e(\theta) : \quad \theta \in \{\alpha, \beta, \gamma\}; e \in e(T)\}.$$

Given T and the $3|e(T)|$ edge length parameters, we can model sequence evolution on T under the K3ST model, if we specify a sequence of nucleotides at one leaf and generate corresponding nucleotides at every other vertex according to the probabilities $p_e(\theta)$. We comment that the K3ST model induces uniform base distribution under equilibrium. However, since our work deals with the probabilities along the edges, our derivation is indifferent to the base distribution.

Edge indexing. The deletion of an edge $e \in e(T)$ induces two subtrees, whose leaf label sets A, A' (with $0 \in A'$) partition $L(T) = [n]_0$. Thus, A is the set of leaves of T separated from reference leaf 0 by the edge e. We choose the subset $A \subseteq [n]$ (the subset not containing 0) to index e as e_A. Thus, for $e = e_A \in e(T)$:

$$A = \{i \in [n] : e \in \Pi_{0i}\},$$

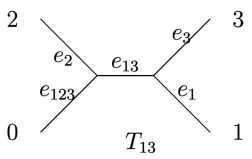
where Π_{0i} is the path (in T) connecting leaves 0 and i.

A partition of a set X into two subsets $\{A, A'\}$ (so $A \cap A' = \emptyset$ and $A \cup A' = X$) is called a *split* of X. When $X = [n]_0$, we will identify each split $\{A, A'\}$ by the subset A, which does not contain 0 and, hence, we see the set of splits of $[n]_0 = \{0, 1, \ldots, n\}$ is bijective with the set of subsets of $[n] = \{1, 2, \ldots, n\}$.

Now, for each $A \subseteq [n]$, we define the three values, $(q_\alpha)_A$, $(q_\beta)_A$, and $(q_\gamma)_A$, by

$$(q_\theta)_A = \begin{cases} q_{e_A}(\theta) & \text{if } e_A \in e(T), \\ -K_\theta = -\sum_{e_B \in e(T)} q_{e_B}(\theta) & \text{if } A = \emptyset, \\ 0 & \text{else,} \end{cases}$$

for $\theta = \alpha, \beta, \gamma$. We incorporate these values into three vectors $\mathbf{q}_\alpha$, $\mathbf{q}_\beta$, and $\mathbf{q}_\gamma$, each of 2^n entries. We order the



$$\mathbf{q}_\alpha = \begin{bmatrix} -K(\alpha) \\ q_1(\alpha) \\ q_2(\alpha) \\ 0 \\ q_3(\alpha) \\ q_{13}(\alpha) \\ 0 \\ q_{123}(\alpha) \end{bmatrix}, \mathbf{q}_\beta = \begin{bmatrix} -K(\beta) \\ q_1(\beta) \\ q_2(\beta) \\ 0 \\ q_3(\beta) \\ q_{13}(\beta) \\ 0 \\ q_{123}(\beta) \end{bmatrix}, \mathbf{q}_\gamma = \begin{bmatrix} -K(\gamma) \\ q_1(\gamma) \\ q_2(\gamma) \\ 0 \\ q_3(\gamma) \\ q_{13}(\gamma) \\ 0 \\ q_{123}(\gamma) \end{bmatrix},$$

$$Q_T = \begin{bmatrix} -K & q_1(\alpha) & q_2(\alpha) & 0 & q_3(\alpha) & q_{13}(\alpha) & 0 & q_{123}(\alpha) \\ q_1(\beta) & q_1(\gamma) & . & . & . & . & . & . \\ q_2(\beta) & . & q_2(\gamma) & . & . & . & . & . \\ 0 & . & . & 0 & . & . & . & . \\ q_3(\beta) & . & . & . & q_3(\gamma) & . & . & . \\ q_{13}(\beta) & . & . & . & . & q_{13}(\gamma) & . & . \\ 0 & . & . & . & . & . & 0 & . \\ q_{123}(\beta) & . & . & . & . & . & . & q_{123}(\gamma) \end{bmatrix},$$

Fig. 2. Example: The edge-length spectrum of the tree $T = T_{13}$.

components of the vectors by the subsets of $[n]$ as follows: $\emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \{4\}, \cdots, [n]$, etc. As $(q_\theta)_\emptyset = -K_\theta$, the sum of the components in each vector is 0.

We will also find it convenient to incorporate the vectors into a $2^n \times 2^n$ matrix:

$$Q_T = \left[q_{A,B}\right]_{A,B\subseteq[n]},$$

where

$$q_{A,B} = \begin{cases} q_{e_A}(\beta) & \text{if } e_A \in e(T), B = \emptyset, \\ q_{e_B}(\alpha) & \text{if } A = \emptyset, e_B \in e(T), \\ q_{e_A}(\gamma) & \text{if } A = B, e_A \in e(T), \\ -K_T & \text{if } A = B = \emptyset, \\ 0 & else, \end{cases} \tag{7}$$

where

$$K_T = \sum_{e\in e(T)} (q_e(\alpha) + q_e(\beta) + q_e(\gamma)) = \sum_{e\in e(T)} K_e.$$

Thus, the leading row of Q_T is $\mathbf{q}_\alpha$, the leading column is $\mathbf{q}_\beta$, and the leading diagonal is $\mathbf{q}_\gamma$, all other entries are 0, apart from $Q_{\emptyset,\emptyset} = -K_T$ (hence, the sum of all entries of Q_T is 0). Q_T is referred to as the **edge length spectrum** for T. The positive entries of this spectrum identify the edges of T.

Fig. 2 shows an example of the tree $T = T_{13}$ on $n+1 = 4$ taxa, and its edge-length spectrum as three vectors, and incorporated in the 8×8 matrix Q_T. Corresponding coordinates of the vectors $\mathbf{q}_\alpha$, $\mathbf{q}_\beta$, and $\mathbf{q}_\gamma$ give the three edge length parameters for the corresponding edge. The "0" value indicates that there is no corresponding edge in T. These vectors are placed in the leading row, column, and main diagonal of the matrix Q_T. This means that for $A,\ B \subseteq \{1,2,3\}$, $q_{\emptyset,B} = q_B(\alpha)$, $q_{A,\emptyset} = q_A(\beta)$, $q_{A,A} = q_A(\gamma)$, and for all other entries, $q_{A,B} = 0$, except the first entry $q_{\emptyset,\emptyset} = -K$, where $K = K(\alpha) + K(\beta) + K(\gamma)$. The entries indicated by "." are all zero and are zero for every tree. The entries indicated by "0" are zero for the topology of T, signifying that the splits represented by them are not part of T. Different topologies can have positive values for these entries. The nonzero entries (in the leading row, column, and main diagonal) should each be in the same coordinates as they identify the edge splits of T. For general trees on $n+1$ taxa, the edge length spectra are vectors and square matrices of order 2^n.

4 Site Patterns

In this section, we introduce the notion of a character χ. We also define the notion of site pattern and show that each site pattern is identified by an ordered pair of splits, (C, D), $(C, D \subseteq [n])$, and that every character χ can be recovered from the site pattern and the state at taxon 0. This leads to the definition of the sequence probability spectrum that records the probability of obtaining every site pattern.

When we propose a sequence of nucleotides[1] at leaf 0, and an edge-length spectrum Q_T on a phylogeny T with leaf set $L = [n]_0$, we can generate homologous sequences at each of the other leaves of T under this stochastic model. A common position in each of these sequences is called a *site.* An assignment of nucleotides at a given site is called a *character* χ. Specifically,

$$\chi : L \to \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$$

assigns a nucleotide to each leaf, with $\chi(i)$ the *character state* at leaf i. This assignment partitions L into subsets L_A, L_C, L_G, and L_T, where for $\mathtt{X} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

$$L_\mathtt{X} = \{i \in L : \chi(i) = \mathtt{X}\}.$$

Given the character χ, we define the *character substitution map* $\theta : L \to \mathcal{T}$ such that

$$\theta(i)(\chi(0)) = \chi(i),$$

and a pair of sets $C(\chi)$, $D(\chi) \subseteq [n]$, where

$$C(\chi) = \{i \in [n] : \theta(i) \in \{\beta, \gamma\}\},$$
$$D(\chi) = \{j \in [n] : \theta(j) \in \{\alpha, \gamma\}\}.$$

The pair of subsets $(C, D) = (C(\chi), D(\chi))$ is called the **site pattern** for χ. Given the site pattern (C, D) and the character state $\chi(0)$ at the reference leaf 0, we can recover χ. For example, if $i \in D - C$ and $\chi(0) = \mathtt{G}$, then $\theta(i) = \alpha$, so $\chi(i) = \alpha(\mathtt{G}) = \mathtt{A}$.

There are four characters χ (depending on the state of $\chi(0)$) that correspond to the same site pattern (C, D). Under equilibrium and by the symmetries of K3ST model, each has the same probability of being generated on T by this model. However, the transition matrices at the tree edges are not dependent on this. Let $s_{C,D}$ be the probability of obtaining the site pattern (C, D) (recall the site pattern (C, D) is obtained from four characters χ, as $\chi(0)$ takes each character value). We now define the $2^n \times 2^n$ matrix S_T, the **sequence probability spectrum**, with rows and columns indexed by the subsets of $[n]$, where

$$S_T = [s_{C,D}]_{C,D\subseteq[n]}.$$

1. The assumption about nonuniform base frequency holds here as well.

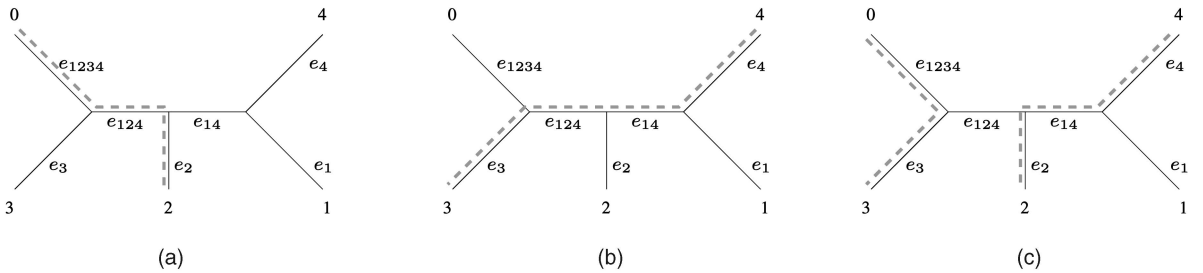


Fig. 3. (a) The path $\Pi_{02} = \{e_{1234}, e_{124}, e_2\}$ (dashed line). (b) The path $\Pi_{34} = \{e_3, e_{124}, e_{14}, e_4\}$ (dashed line). (c) The path set $\Pi_{0234} = \{e_{1234}, e_3, e_2, e_{14}, e_4\}$ (dashed lines). Note that, in each case, $\Pi_E = \{e_A \in e(T) : |A \cap E| \text{ is odd}\}$. We note $\Pi_{02} \cup \Pi_{34}$ can be partitioned into $W = \Pi_{02} \cap \Pi_{34} = \{e_{124}\}$, $U = \Pi_{02} - W = \{e_{1234}, e_2\}$, and $V = \Pi_{34} - W = \{e_3, e_{14}, e_4\}$, as in (21).

The main theorem of this paper (Theorem 10) links between the probability $s_{C,D}$, for each C, $D \subseteq [n]$, and the edge length parameters $q_e(\theta) : e \in e(T),\ \theta \in \mathcal{T}$. We will derive explicit formulas for $s_{C,D}$ as a function of edge length parameters.

5 Hadamard Matrices

We define recursively the family $\{H_n : n \in \mathbb{Z}^+\}$, (known as Sylvester matrices), where for $n \geq 2$

$$H_n = H_1 \otimes H_{n-1} = \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix},$$

is a symmetric Hadamard matrix of order 2^n, with H_1 and H_2 as previously defined. It is easily seen that $H_n^{-1} = 2^{-n} H_n$.

It is known [10] that if we index the rows and columns of H_n by the subsets of $[n]$, then, for A, $B \subseteq [n]$, we have the following observation.

Observation 4.

$$[H_n]_{A,B} = h(A,B) = (-1)^{|A \cap B|} = h(B,A).$$

Further, for B, $C \subseteq [n]$, we write their symmetric difference as $B \Delta C (= (B \cup C) - (B \cap C))$, and we see

$$\begin{aligned}(-1)^{|A \cap (B \Delta C)|} &= (-1)^{(|A \cap B| + |A \cap C| - 2|A \cap (B \cap C)|)} \\ &= (-1)^{|A \cap B|} (-1)^{|A \cap C|}.\end{aligned}$$

Hence, we have the following.

Observation 5.

$$h(A, (B \Delta C)) = h(A,B) h(A,C).$$

6 Path Sets

In this section, we show how to decompose a set of paths connecting an even number of leaves into a set of edge disjoint paths. We denote the latter as a path set. We then generalize the edge length into path-set distances with respect to each substitution $\theta \in \mathcal{T}$.

For any i, $j \in L(T) = [n]_0$, we define the **path** Π_{ij} to be the set of edges in T connecting leaves i and j. In particular, we note

$$\Pi_{0i} = \{e_A \in e(T) : i \in A\}. \tag{8}$$

Π_{ij} is obtained by deleting the common edges of Π_{0i} and Π_{0j} from their union, so

$$\begin{aligned}\Pi_{ij} = \Pi_{0i} \Delta \Pi_{0j} &= \{e_A \in e(T) : |A \cap \{i,j\}| = 1\} \\ &= \{e_A \in e(T) : h(A, \{i,j\}) = -1\}.\end{aligned}$$

For any $E \subseteq [n]$, let

$$\Pi_E = \{e_A \in e(T) : h(A,E) = -1\},$$

so, in particular, for i, $j \in [n]$, we see

$$\Pi_{\{i,j\}} = \Pi_{ij}, \quad \Pi_{\{i\}} = \Pi_{0i}, \quad \Pi_\emptyset = \emptyset.$$

Observation 6. *In [14], it is shown that Π_E is a collection of edge disjoint paths, with end-point set E or $E \cup \{0\}$.*

Π_E is called a **path set**. Figs. 3a and 3b show the two paths $\Pi_{0,2}$ and $\Pi_{3,4}$, respectively, while (Fig. 3c) shows the path set induced by the set {0, 2, 3, 4}.

By similar arguments to the discussion above, we find the following.

Observation 7. *The set of path sets is a group (under symmetric difference) isomorphic to C_2^n. In particular, $\Pi_{E \Delta F} = \Pi_E \Delta \Pi_F$.*

The sum of edge lengths on a path connecting two leaves can naturally be thought of as the **distance** between the leaves. We extend this distance concept, for each substitution type $\theta \in \{\alpha, \beta, \gamma\}$, to path sets. We define the **path-set distance** of path set Π_E to be the sum of the corresponding edge lengths of each edge of the path set, that is,

$$d_{\Pi_E}(\theta) = \sum_{e_A \in \Pi_E} q_A(\theta).$$

7 Site Patterns and the Hadamard Matrix

Here, for the sake of the explanation, we extend the notion of a character to conceptually assign values to the internal vertices of T. This allows us to extend the notion of the substitution function θ to the context of edges and, subsequently, to path sets.

Suppose we are given $\chi(0)$, the character state at leaf 0 and assign a transformation $\theta(v) \in \mathcal{T}$ to each vertex v of T, such that the character state at v is $\chi(v) = \theta(v)(\chi(0))$. (In particular, note that $\theta(0) = \epsilon$, the identity.) If we restrict

ourselves to the set of leaves $[n]$, then we find that the consequent site pattern is $(C, D) = (C(\chi), D(\chi))$, where

$$C(\chi) = \{i \in [n] : \theta(i) \in \{\beta, \gamma\}\},$$
$$D(\chi) = \{j \in [n] : \theta(j) \in \{\alpha, \gamma\}\},$$

are subsets of $[n]$.

Further, for each edge $e = (u, v)$, the transformation across e is

$$\theta(e) = \theta(u)^{-1}\theta(v) = \theta(u)\theta(v),$$

(as $(\mathcal{T}, \circ)$ is a Boolean group). For the path $\Pi_{i,j}$ connecting leaves i and j, we find

$$\prod_{e\in\Pi_{i,j}} \theta(e) = \prod_{e=(u,v)\in\Pi_{i,j}} \theta(u)^{-1}\theta(v) = \theta(i)^{-1}\theta(j) = \theta(i)\theta(j), \quad (9)$$

(as the products at each internal vertex cancel, and $\theta^{-1} = \theta$). We extend this to any path set Π_E (where $E \subseteq [n]$)) and define

$$\theta(E) = \prod_{e\in\Pi_e} \theta(e).$$

Hence, as in (9), the products at all internal vertices cancel, so we find the following.

Observation 8.

$$\theta(E) = \prod_{i\in E} \theta(i). \quad (10)$$

Consider now a character χ, inducing a site pattern $(C, D) = (C(\chi), D(\chi))$. Recall (1) as the homomorphisms g_1, $g_2 : \mathcal{T} \to \{1, -1\}$, where

$$g_1(\theta) = -1 \iff \theta \in \{\beta, \gamma\} \text{ and } g_2(\theta) = -1 \iff \theta \in \{\alpha, \gamma\}.$$

Then, for $j = 1, 2$

$$g_j(\theta(E)) = g_j\left(\prod_{i\in E} \theta(i)\right) = \prod_{i\in E} g_j(\theta(i)). \quad (11)$$

Then, $g_j(\theta(E)) = 1$ exactly when the number of factors $g_j(\theta(i)) = -1$ in the product 11 is even. Now, for $i \in E$

$$g_1(\theta(i)) = -1 \iff \theta(i) \in \{\beta, \gamma\} \iff i \in C,$$

so

$$g_1(\theta(E)) = 1 \iff |C \cap E| \equiv 0(\text{mod } 2) \iff h(C, E) = 1.$$

However, as $g_1(\theta(E)) = 1 \iff \theta(E) \in \{\epsilon, \alpha\}$, we find

$$\theta(E) \in \{\epsilon, \alpha\} \iff h(C, E) = 1. \quad (12)$$

Similarly, we find

$$\theta(E) \in \{\epsilon, \beta\} \iff g_2(\theta(E)) = 1 \iff |D \cap E| \equiv 0(\text{mod } 2) \iff h(D, E) = 1. \quad (13)$$

Hence, we have shown the following.

Observation 9. *Given a character χ inducing a site pattern*

$$(C, D) = (C(\chi), D(\chi)),$$

and path set Π_E with $\theta(E) = \prod_{i\in E} \theta(i)$,

$$h(C, E) = g_1(\theta(E)), \qquad h(D, E) = g_2(\theta(E)). \quad (14)$$

8 Hadamard Conjugation

This is the final section in the derivation in which we prove the main theorem of this paper. We start with the right-hand side $H_nQ_TH_n$ and show that an entry in that matrix, corresponds to a path-set distance. We decompose these possibly overlapping distances into three disjoint edge sets using previous identities to derive probabilities of substitutions along each of these edge sets and then recombine them to the original path sets.

Q_T is the matrix containing the edge-length parameters across T. S_T is the matrix of probabilities of patterns at the leaves of T. The link between these are the rotations $H_nS_TH_n$ and $H_nQ_TH_n$. These both relate to path-set properties and enable us to state our major result.

Theorem 10.

$$S_T = H_n^{-1}(\text{Exp}(H_nQ_TH_n))H_n^{-1}, \quad (15)$$

which, provided the arguments of the logarithm are positive, is invertible and gives

$$Q_T = H_n^{-1}(\text{Ln}(H_nS_TH_n))H_n^{-1}. \quad (16)$$

Proof. The proof of this theorem is based on interpreting the corresponding components, for E, $F \subseteq [n]$,

$$[H_nS_TH_n]_{E,F} \quad \text{and} \quad [H_nQ_TH_n]_{E,F}.$$

Expanding the second term, we find

$$\begin{aligned}[H_nQ_TH_n]_{E,F} &= \sum_{A,B\subseteq[n]} h(A, E)h(B, F)q_{A,B} \\ &= q_{\emptyset,\emptyset} + \sum_{e_A\in e(T)} (h(A, E)q_{A,\emptyset} + h(A, F)q_{\emptyset,A} \\ &\quad + h(A, E)h(A, F)q_{A,A}),\end{aligned} \quad (17)$$

as the only nonzero entries in Q_T are $q_{A,\emptyset}$, $q_{\emptyset,A}$, $q_{A,A}$ for $e_A \in e(T)$, and $q_{\emptyset,\emptyset}$.

Recall (Observation 5)

$$q_{\emptyset\emptyset} = -\sum_{e_A\in e(T)} (q_{A,\emptyset} + q_{\emptyset,A} + q_{A,A}), \text{ and}$$
$$h(A, E)h(A, F) = h(A, E\Delta F),$$

hence, the RHS of (17) can be written as

$$\begin{aligned}\sum_{e_A\in e(T)} &((h(A, E) - 1)q_{A,\emptyset} + (h(A, F) - 1)q_{\emptyset,A} \\ &+ (h(A, E\Delta F) - 1)q_{A,A}).\end{aligned} \quad (18)$$

Now, as the terms with $h(A, E) = 1$ cancel, and by the definition of Π_E, we can write (18) as

$$[H_n Q_T H_n]_{E,F} = -2 \sum_{e_A \in \Pi_E} q_{A,\emptyset} - 2 \sum_{e_A \in \Pi_F} q_{\emptyset,A} - 2 \sum_{e_A \in \Pi_{E\Delta F}} q_{A,A}. \tag{19}$$

By Observation 7

$$\Pi_{E\Delta F} = \Pi_E \Delta \Pi_F,$$

and by the definition of the matrix $Q = Q_T$

$$q_{A,\emptyset} = q_{e_A}(\beta), \quad q_{\emptyset,A} = q_{e_A}(\alpha), \quad q_{A,A} = q_{e_A}(\gamma),$$

so

$$\sum_{e_A \in \Pi_E} q_{A,\emptyset} = d_{\Pi_E}(\beta), \quad \sum_{e_A \in \Pi_F} q_{\emptyset,A} = d_{\Pi_F}(\alpha),$$
$$\sum_{e_A \in \Pi_{E\Delta F}} q_{A,A} = d_{\Pi_E \Delta \Pi_F}(\gamma).$$

Hence,

$$[H_n Q_T H_n]_{E,F} = -2(d_{\Pi_E}(\beta) + d_{\Pi_F}(\alpha) + d_{\Pi_E \Delta \Pi_F}(\gamma)). \tag{20}$$

We can partition $\Pi_E \cup \Pi_F$ into three parts

$$U = \Pi_E - \Pi_F, \quad V = \Pi_F - \Pi_E, \quad W = \Pi_E \cap \Pi_F, \tag{21}$$

as illustrated in Fig. 3c, with the path sets partitioned as

$$\Pi_E = U \Delta W, \quad \Pi_F = V \Delta W, \quad \Pi_{E\Delta F} = \Pi_E \Delta \Pi_F = U \Delta V.$$

The path-set distances split into corresponding summands, as $d_U(\theta) = \sum_{e \in U} q_e(\theta)$, etc., so that

$$d_{\Pi_E}(\beta) = d_U(\beta) + d_W(\beta), \quad d_{\Pi_F}(\alpha) = d_V(\alpha) + d_W(\alpha),$$
$$d_{\Pi_{E\Delta F}}(\gamma) = d_U(\gamma) + d_V(\gamma).$$

Thus,

$$\begin{aligned}[H_n Q_T H_n]_{E,F} &= -2[d_U(\beta) + d_W(\beta) + d_V(\alpha) + d_W(\alpha) + d_U(\gamma) \\ &\quad + d_V(\gamma)] \\ &= -2[d_U(\beta) + d_U(\gamma)] - 2[d_V(\alpha) + d_V(\gamma)] \\ &\quad - 2[d_W(\alpha) + d_W(\beta)],\end{aligned}$$

and

$$[\mathrm{Exp}(H_n Q_T H_n)]_{E,F} = \mathrm{e}^{-2[d_U(\beta)+d_U(\gamma)]} \mathrm{e}^{-2[d_V(\alpha)+d_V(\gamma)]} \mathrm{e}^{-2[d_W(\alpha)+d_W(\beta)]}. \tag{22}$$

Now, by Observation 3

$$\begin{aligned}\mathrm{e}^{-2[d_U(\beta)+d_U(\gamma)]} &= p_U(\epsilon) + p_U(\alpha) - p_U(\beta) - p_U(\gamma) \\ &= \sum_{\theta \in \mathcal{T}} g_1(\theta) p_U(\theta), \\ \mathrm{e}^{-2[d_V(\alpha)+d_V(\gamma)]} &= p_V(\epsilon) - p_V(\alpha) + p_V(\beta) - p_V(\gamma) \\ &= \sum_{\theta \in \mathcal{T}} g_2(\theta) p_V(\theta), \\ \mathrm{e}^{-2[d_W(\alpha)+d_W(\beta)]} &= p_W(\epsilon) - p_W(\alpha) - p_W(\beta) + p_W(\gamma) \\ &= \sum_{\theta \in \mathcal{T}} g_1(\theta) g_2(\theta) p_W(\theta).\end{aligned}$$

Hence, (22) becomes

$$\begin{aligned}[\mathrm{Exp}(H_n Q_T H_n)]_{E,F} &= [p_U(\epsilon) + p_U(\alpha) - p_U(\beta) - p_U(\gamma)] \\ &\quad \times [p_V(\epsilon) - p_V(\alpha) + p_V(\beta) - p_V(\gamma)] \\ &\quad \times [p_W(\epsilon) - p_W(\alpha) - p_W(\beta) + p_W(\gamma)] \\ &= \left[\sum_{\theta \in \mathcal{T}} g_1(\theta) p_U(\theta)\right] \left[\sum_{\phi \in \mathcal{T}} g_2(\phi) p_V(\phi)\right] \\ &\quad \left[\sum_{\psi \in \mathcal{T}} g_1(\psi) g_2(\psi) p_W(\psi)\right] \\ &= \sum_{\theta,\phi,\psi \in \mathcal{T}} g_1(\theta\psi) g_2(\phi\psi) p_U(\theta) p_V(\phi) p_W(\psi) \\ &\quad \text{(expanding the products)} \\ &= \sum_{\xi,\eta,\psi \in \mathcal{T}} g_1(\xi) g_2(\eta) p_U(\xi\psi) p_V(\eta\psi) p_W(\psi), \\ &\quad \text{(with } \xi = \theta\psi \text{ and } \eta = \phi\psi) \\ &= \sum_{\xi,\eta, \in \mathcal{T}} g_1(\xi) g_2(\eta) \\ &\quad \left[\sum_{\psi, \in \mathcal{T}} p_U(\xi\psi) p_V(\eta\psi) p_W(\psi)\right].\end{aligned} \tag{23}$$

Now, as E and F are partitioned as $U\Delta W$, and $V\Delta W$, respectively, then $\theta(E) = \theta(U)\theta(W)$ and $\theta(F) = \theta(V)\theta(W)$. Hence,

$$\sum_{\psi \in \mathcal{T}} p_U(\xi\psi) p_V(\eta\psi) p_W(\psi) = \Pr[\theta(E) = \xi \wedge \theta(F) = \eta],$$

which is the joint probability that the product of substitutions across the edges of $\Pi(E)$ is ξ, and the product across the edges of $\Pi(F)$ is η. Thus, (23) implies

$$[\mathrm{Exp}(H_n Q_T H_n)]_{E,F} = \sum_{\xi,\eta \in \mathcal{T}} g_1(\xi) g_2(\eta) \Pr[\theta(E) = \xi \wedge \theta(F) = \eta]. \tag{24}$$

By Observation 9, given a character χ, inducing a site pattern $(C, D) = (C(\chi), D(\chi))$, we have

$$g_1(\theta(E)) = h(C, E), \qquad g_2(\theta(F)) = h(D, F), \tag{25}$$

that is, under χ

$$\theta(E) \in \{\epsilon, \alpha\} \iff h(C, E) = 1, \text{ and}$$
$$\theta(F) \in \{\epsilon, \beta\} \iff h(D, F) = 1.$$

Summing the probabilities $s_{C,D}$, where both $h(C, E) = 1$ and $h(D, F) = 1$, we find

$$\Pr[\theta(E) \in \{\epsilon, \alpha\}, \theta(F) \in \{\epsilon, \beta\}] = \sum_{C,D \subseteq [n]: h(C,E)=1, h(D,F)=1} s_{C,D}.$$

Similarly, we see

$$\begin{aligned}
&\Pr[\theta(E) \in \{\epsilon, \alpha\} \wedge \theta(F) \in \{\alpha, \gamma\}] \\
&= \sum_{C,D \subseteq [n]: h(C,E)=1, h(D,F)=-1} ; s_{C,D}, \\
&\Pr[\theta(E) \in \{\beta, \gamma\} \wedge \theta(F) \in \{\epsilon, \beta\}] \\
&= \sum_{C,D \subseteq [n]: h(C,E)=-1, h(D,F)=1} s_{C,D}, \\
&\Pr[\theta(E) \in \{\beta, \gamma\} \wedge \theta(F) \in \{\alpha, \gamma\}] \\
&= \sum_{C,D \subseteq [n]: h(C,E)=-1, h(D,F)=-1} s_{C,D}.
\end{aligned}$$

Substituting these into (24), we obtain by Observation 9

$$\begin{aligned}
[\mathrm{Exp}(H_n Q_T H_n)]_{E,F} &= \sum_{C,D \subseteq [n]} h(C,E) h(D,F) s_{C,D} \\
&= [H_n S_T H_n]_{E,F},
\end{aligned}$$

giving

$$\mathrm{Exp}(H_n Q_T H_n) = H_n S_T H_n, \tag{26}$$

from which (15) and (16) follow. □

9 Applying the Hadamard Conjugation

In this section, we provide an example application of using the Hadamard conjugation in real biological problems. We use the application reported in [5] of obtaining an analytical solution for the maximum likelihood problem of a phylogenetic reconstruction. As was shown above, a tree T uniquely determines the sequence spectrum $S = S_T$. In real life, however, we do not find such a "perfect" S. Given a set of input aligned sequences, every column induces a site pattern. The matrix $\hat{S} = [\hat{s}]_{C,D}$, denoted as the *observed sequence spectrum* records the frequency of each site pattern (C, D). For a tree T, the likelihood function is defined as

$$L(T) = \prod_{C,D \subseteq [n]} s_{C,D}^{\hat{s}_{C,D}}, \tag{27}$$

where $s_{C,D}$ comes from (15). That is, (27) expresses the probability of seeing $\hat{S}$ given T. The maximum likelihood problem is to find a tree such that the probability of obtaining the given data $\hat{S}$ is maximized.

In [5], the ML problem of a triplet tree under the Jukes-Cantor model and the molecular clock hypothesis was studied (see Fig. 4). The Jukes-Cantor model of evolution [19] is the simplest model for four states DNA evolution. The assumption in this model is that when a base changes, it has equal probabilities to change to each of the other three bases. This model can be derived from the more general K3ST model by setting, for each edge of T, each of the three edge length parameters equal to a common value, namely, setting $q_e(\alpha) = q_e(\beta) = q_e(\gamma) = q_e$. We now look at a general tree T on three taxa {0, 1, 2} before determining where the root is. The molecular clock hypothesis and determination of the root location are done at a later stage. T has just one topology, the star with the three edges $e_{\{1\}}$, $e_{\{2\}}$, and $e_{\{1,2\}}$

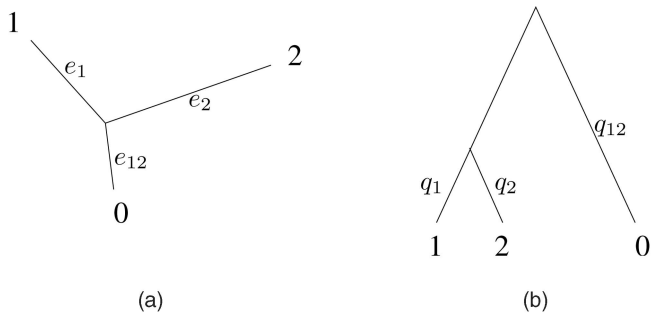


Fig. 4. (a) A general triplet tree over the species {0, 1, 2}. (b) A rooted triplet tree under Jukes-Cantor model and the molecular clock hypothesis.

(see Fig. 4a). For simplicity we denote them as e_1, e_2 and $e_{1,2}$ respectively.

The edge-length spectrum of an arbitrary 3-tree can be expressed as

$$Q = \begin{bmatrix} -3(q_1+q_2+q_{12}) & q_1 & q_2 & q_{12} \\ q_1 & q_1 & 0 & 0 \\ q_2 & 0 & q_2 & 0 \\ q_{12} & 0 & 0 & q_{12} \end{bmatrix}.$$

Now, we see that

$$HQH = -4 \begin{bmatrix} 0 & q_1+q_{12} & q_2+q_{12} & q_1+q_2 \\ q_1+q_{12} & q_1+q_{12} & q_1+q_2+q_{12} & q_1+q_2+q_{12} \\ q_2+q_{12} & q_1+q_2+q_{12} & q_2+q_{12} & q_1+q_2+q_{12} \\ q_1+q_2 & q_1+q_2+q_{12} & q_1+q_2+q_{12} & q_1+q_2 \end{bmatrix},$$

and by (20), these are minus twice the sum of distances induced by every two path sets E, F for every entry $[HQH]_{E,F}$. For example,

$$\begin{aligned}
[H_n Q_T H_n]_{1,12} &= -2(d_{\Pi_1}(\beta) + d_{\Pi_{12}}(\alpha) + d_{\Pi_1 \Delta \Pi_{12}}(\gamma)) \\
&= -2(d_{\Pi_1}(\beta) + d_{\Pi_{12}}(\alpha) + d_{\Pi_2}(\gamma)) \\
&= -4(q_1 + q_2 + q_{12}).
\end{aligned}$$

When applying the exponential function to each element of the matrix HQH, we obtain the so called path-set spectrum, R:

$$R = \exp(HQH) = \begin{bmatrix} 1 & x_1x_{12} & x_2x_{12} & x_1x_2 \\ x_1x_{12} & x_1x_{12} & x_1x_2x_{12} & x_1x_2x_{12} \\ x_2x_{12} & x_1x_2x_{12} & x_2x_{12} & x_1x_2x_{12} \\ x_1x_2 & x_1x_2x_{12} & x_1x_2x_{12} & x_1x_2 \end{bmatrix}, \tag{28}$$

where

$$x_i = e^{-4q_i}. \tag{29}$$

The x_i values can replace the $s_{C,D}$ values as the defining parameters in the likelihood function (27). The entries of R relate to the joint probabilities of differences between the end-points of the corresponding path sets in T, as implied by (24).

By using our main Theorem 10, (15), the sequence probability spectrum equals

$$S = H^{-1}RH^{-1}, \tag{30}$$

$$= \frac{1}{16}\begin{bmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_1 & a_4 & a_4 \\ a_2 & a_4 & a_2 & a_4 \\ a_3 & a_4 & a_4 & a_3 \end{bmatrix}, \tag{31}$$

where

$$\begin{aligned} a_0 &= (1 + 3x_1x_2 + 3x_1x_{12} + 3x_2x_{12} + 6x_1x_2x_{12}), \\ a_1 &= (1 - x_1x_2 - x_1x_{12} + 3x_2x_{12} - 2x_1x_2x_{12}), \\ a_2 &= (1 - x_1x_2 + 3x_1x_{12} - x_2x_{12} - 2x_1x_2x_{12}), \\ a_3 &= (1 + 3x_1x_2 - x_1x_{12} - x_2x_{12} - 2x_1x_2x_{12}), \\ a_4 &= (1 - x_1x_2 - x_1x_{12} - x_2x_{12} + 2x_1x_2x_{12}). \end{aligned} \tag{32}$$

Thus, we see that each expected sequence frequency takes one of the above values, which are the functions of the three parameters x_1, x_2, and x_{12}. We now apply the molecular clock constraint that asserts $q_1 = q_2 \Rightarrow x_1 = x_2$ (see Fig. 4b). From (32), it can be seen that under this constraint, $a_1 = a_2$, so the number of free variables in the likelihood equation reduces to 4, leading to a further simplification. The rest of the ML solution is orthogonal to the material discussed in this paper and can be found in [5].

We remark here that the choice of these parameters was proved crucial in the derivation of the analytical solution. In former works [6], [7], [8], the defining parameters were the sequence probability variables themselves, and additional constraints were required to guarantee that they reside on a tree surface. This approach failed in this case, and as can be seen using the path-set variables, these constraints are removed.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referee II who provided very helpful comments and suggestions.

REFERENCES

[1] E.S. Allman and J.A. Rhodes, "Phylogenetic Invariants for the General Markov Model of Sequence Mutation," *Math. Biosciences,* vol. 186, pp. 113-144, 2003.

[2] E.S. Allman and J.A. Rhodes, "Quartets and Parameter Recovery for the General Markov Model of Sequence Mutation," *Applied Math. Research eXpress,* vol. 4, pp. 107-131, 2004.

[3] E.S. Allman and J.A. Rhodes, "Phylogenetic Ideals and Varieties for the General Markov Model," *Advances in Applied Math.,* vol. 40, no. 2, pp. 127-148, 2008.

[4] M. Casanellas and J. Fernández-Sánchez, "Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees," *Molecular Biology and Evolution,* vol. 24, no. 1, pp. 288-293, 2007.

[5] B. Chor, M.D. Hendy, and S. Snir, "Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions," *Molecular Biology and Evolution,* vol. 23, no. 3, pp. 626-632, 2006.

[6] B. Chor, A. Khetan, and S. Snir, "Maximum Likelihood on Four Taxa Phylogenetic Trees: Analytic Solutions," *Proc. Seventh Ann. Int'l Conf. Computational Molecular Biology (RECOMB '03),* pp. 76-83, 2003.

[7] B. Chor and S. Snir, "Molecular Clock Fork Phylogenies: Closed Form Analytic Maximum Likelihood Solutions," *Systematic Biology,* vol. 53, no. 6, pp. 963-967, 2004.

[8] B. Chor, M. Hendy, B. Holland, and D. Penny, "Multiple Maxima of Likelihood in Phylogenetic Trees: An Analytic Approach," *Molecular Biology and Evolution,* vol. 17, pp. 1529-1541, 2000.

[9] S.N. Evans and T.P. Speed, "Invariants of Some Probability Models Used in Phylogenetic Inference," *Annals of Statistics,* vol. 21, pp. 355-377, 1993.

[10] M.D. Hendy, "The Relationship between Simple Evolutionary Tree Models and Observable Sequence Data," *Systematic Zoology,* vol. 38, pp. 310-321, 1989.

[11] M.D. Hendy, "A Combinatorial Description of the Closest Tree Algorithm for Finding Evolutionary Trees," *Discrete Math.,* vol. 96, pp. 51-58, 1991.

[12] M.D. Hendy, "Hadamard Conjugation: An Analytic Tool for Phylogenetics," *Math. of Evolution and Phylogeny,* chapter 6, first ed., O. Gascuel, ed., pp. 143-177, Oxford Univ. Press, 2005.

[13] M.D. Hendy and D. Penny, *A Framework for the Quantitative Study of Evolutionary Trees,* Systematic Zoology, vol. 38, pp. 297-309, 1989.

[14] M.D. Hendy and D. Penny, "Spectral Analysis of Phylogenetic Data," *J. Classification,* vol. 10, pp. 5-24, 1993.

[15] M.D. Hendy and D. Penny, "Complete Families of Linear Invariants for Some Stochastic Models of Sequence Evolution with and without the Molecular Clock Assumption," *J. Computational Biology,* vol. 3, pp. 19-31, 1996.

[16] M.D. Hendy, D. Penny, and M.A. Steel, "A Discrete Fourier Analysis for Evolutionary Trees," *Proc. Nat'l Academy of Sciences,* vol. 91, pp. 3339-3343, 1994.

[17] B. Holland, D. Penny, and M. Hendy, "Outgroup Misplacement and Phylogenetic Inaccuracy under a Molecular Clock—A Simulation Study," *Systematic Biology,* vol. 52, pp. 229-238, 2003.

[18] K.T. Huber, M. Langton, D. Penny, V. Moulton, and M. Hendy, "Spectronet: A Package for Computing Spectra and Median Networks," *Applied Bioinformatics,* vol. 1, pp. 159-161, 2002.

[19] T.H. Jukes and C.R. Cantor, "Evolution of Protein Molecules," *Mammalian Protein Metabolism III,* H.N. Munro, ed., Academic Press, 1969.

[20] M. Kimura, "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences," *J. Molecular Evolution,* vol. 16, pp. 111-120, 1980.

[21] M. Kimura, "Estimation of Evolutionary Distances between Homologous Nucleotide Sequences," *Proc. Nat'l Academy of Sciences,* vol. 78, pp. 454-458, 1981.

[22] J.L. Neyman, "Molecular Studies of Evolution: A Source of Novel Statistical Problems," *Statistical Decision Theory and Related Topics,* S.S. Gupta and J. Yackel, eds., Academic Press, 1971.

[23] L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology.* Cambridge Univ. Press, 2005.

[24] L. Pachter and B. Sturmfels, "The Mathematics of Phylogenomics," submitted for publication.

[25] M.A. Steel, M.D. Hendy, L.A. Székely, and P.L. Erdös, "Spectral Analysis and a Closest Tree Method for Genetic Sequences," *Applied Math. Letters,* vol. 5, pp. 63-67, 1992.

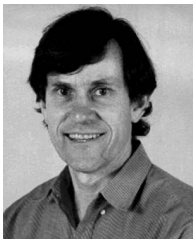
[26] M.A. Steel, M.D. Hendy, and D. Penny, "Reconstructing Phylogenies from Nucleotide Pattern Probabilities: A Survey and Some New Results," *Discrete Applied Math.,* vol. 88, pp. 367-396, 1998.

[27] B. Sturmfels and S. Sullivant, "Toric Ideals of Phylogenetic Invariants," *J. Computational Biology,* vol. 12, pp. 204-228, 2005.

[28] L. Székely, P.L. Erdös, M.A. Steel, and D. Penny, "A Fourier Inversion Formula for Evolutionary Trees," *Applied Math. Letters,* vol. 6, pp. 13-17, 1993.

[29] L. Székely, M.A. Steel, and P.L. Erdös, "Fourier Calculus on Evolutionary Trees," *Advances in Applied Math.,* vol. 14, pp. 200-216, 1993.

[30] P.J. Waddell and M.D. Hendy, "Using Phylogenetic Invariants to Enhance Spectral Analysis of Nucleotide Sequence Data," Information and Math. Sciences Report Series B, Massey Univ., 1997.



Michael D. Hendy received the PhD degree in algebraic number theory from the University of New England, NSW, Australia, in 1973. He was with Massey University, where he began research into phylogenetics in collaboration with molecular biologist, David Penny. Since 1996, he has been a personal chair in mathematical biology at Massey University, Palmerston North, New Zealand. He is also the executive director of the Allan Wilson Centre for Molecular Ecology and Evolution, one of the seven Centres of Research Excellence, New Zealand—the Allan Wilson Centre is a group of nearly 100 researchers in biology and mathematics across five New Zealand universities. In 2000-2001, he held a 10-month mercator professorship in biomathematics at the University of Greifswald, Germany. He has published more than100 research papers in the fields of algebraic number theory and phylogenetics.

Sagi Snir received the BA degree in computer science and economics from Bar Ilan University, Israel, and the MSc and PhD degrees in computer science from the Technion, Israel. He was with various information technologies companies, including IBM Haifa Research Lab. He is currently a postdoctoral researcher at the University of California, Berkeley. His main research interest is computational biology and in particular phylogenetics.