AUSTRALASIAN JOURNAL OF COMBINATORICS
Volume **66(2)** (2016), Pages 177–191

Remarks on Hadamard conjugation and combinatorial phylogenetics

Cayla D. McBee

Department of Mathematics and Computer Science
Providence College
1 Cunningham Square, Providence, RI 02918
U.S.A.
cmcbee@providence.edu

Tim Penttila

Department of Mathematics
Colorado State University
841 Oval Drive, Fort Collins, CO 80523
U.S.A.

Abstract

Hadamard conjugation can be used in the reconstruction of evolutionary trees and analysis of molecular sequence evolution. Despite the number of advantages Hadamard conjugation provides the number of evolutionary substitution models that can be used with this technique is limited. In this paper, we consider the question of whether Hadamard conjugation is limited to group-based evolutionary models, both for nucleotide substitution models and for codon substitution models. We expand the number of nucleotide substitution models that can be used with Hadamard conjugation and suggest new connections between phylogenetics and algebraic combinatorics especially with (commutative) association schemes.

1 Introduction

Phylogenetics is the study of evolutionary relatedness among various groups of organisms. Initially, techniques such as examining the morphological characteristics of species were used to determine historical relationships, however the explosion of data over the past 40 years resulting from advances in technology, the availability of

DNA sequences, and the completion of the human genome project have led to statistical, computational, and algorithmic work on determining evolutionary relatedness between organisms. The increase in data has also allowed for a majority of statisticians to reach a consensus of the statistical foundations in the area. As a result of this consensus, mathematical analysis of the statistical models is now appropriate.

An important topic of interest in combinatorial phylogenetics is the reconstruction of evolutionary trees. All statistical models used to reconstruct evolutionary trees, also known as phylogenetic trees, use either quantitative character data or genetic data available for presently extant species to determine historical relationships between groups of organisms or *taxa*.

Many techniques exist for inferring phylogenetic relationships from molecular data. One such technique involves making assumptions about the evolutionary process and incorporating these assumptions into a Markov model. The Markov model relates the rates at which substitutions in the genetic data take place to the probabilities of different substitutions taking place using the equation $\mathbf{P} = \exp(\mathbf{Q}t)$ where $\mathbf{P}$ and $\mathbf{Q}$ are probability and rate matrices, respectively. Including the conjugation of each side of this equation by a Hadamard matrix allows the derivation of invertible analytic formulas relating relative frequencies of observed patterns from the genetic data to an estimation of the phylogenetic tree that corresponds to the data.

This conjugation of rates and probabilities by a Hadamard matrix is often referred to as Hadamard conjugation. It is also referred to as the Hadamard transform, spectral analysis and various combinations of the names Hadamard, Rademacher, Walsh and Sylvester reflecting the work done by J. Sylvester in 1867, Hadamard in 1893, Rademacher in 1922 and Walsh in 1923. Hadamard conjugation can also be thought of as a Fourier transform over a finite abelian group.[1] A current disadvantage of Hadamard conjugation is its limitation to group-based substitution models.

While association schemes are a thriving area of algebraic combinatorics, they are not known to many biologists. The gentlest introduction to the topic is the text by Rosemary Bailey [2]. The history of the subject and its relationship with experimental design in statistics is covered in Bailey's last chapter. In this book, you learn that when Bose and Shimamoto introduced association schemes in R. C. Bose et al. [4] they made a very sound judgement that enabled a relationship between concurrence and variance to be deduced. (See especially Section 5.3 of Bailey.) Association schemes were independently reintroduced (and generalized) by Weisfeiler and Leman in 1968 (as cellular algebras) by Nair in 1964 (what are now called homogeneous coherent configurations) and by D.G.Higman in 1971 (as coherent configuarations), building on work of Frobenius, Schur and Wielandt concerning centralizer rings of finite permutation groups.

The other standard reference is the text by Bannai and Ito [3], which takes the perspective of Higman, but is also strongly influenced by Delsarte's 1973 Ph. D. thesis. As Bannai and Ito put things, in their preface (page i): It is possible to

[1]See [14] for details of this approach.

describe Algebraic Combinatorics as "a character theoretical study of combinatorial objects", or "a group theory without groups"!

Since the central concern of this paper is whether or not the beautiful techniques of Hadamard conjugation require the existence of an underlying group in the model, this remark of Bannai and Ito's shows what our lodestar will be: we will attempt to replace the necessity for a group by combinatorial regularity conditions.

The details of the idea of generalizing group theory using coherent configurations are less informally presented in the introduction to Higman [9]. The substitute for the group is really a substitute for the centralizer ring, namely the Bose-Mesner algebra.

Other work suggests that Lie algebras may be of importance in this context. The paper by Sumner et al. [17] shows that group based models form Lie Markov models and says that the success of Hadamard conjugation on the Kimura three-substitution type model results from the fact that the Lie algebras of these models are abelian.

We too see that, for the results to have the impact we desire, an algebra should be abelian, namely the Bose-Mesner algebra. But we don't seem to need the Lie structure. The mathematical details of all of this are suppressed in the paper that follows this introduction, as we feel that they would get in the way of understanding for those interested in the biological aspects. Instead, we use the most elementary arguments we are able to discover. But behind the scenes there is a vast literature on algebraic combinatorics.

The beginning of this paper will discuss the use of Hadamard conjugation with group-based evolutionary models. Specifically we will examine the use of the Kimura three-substitution type model commonly written as the K3ST model. Following the discussion we will look at the possibility of using Hadamard conjugation with models other than the K3ST model and submodels. Finally we will consider the use of association schemes with evolutionary models.

2 Preliminaries

A phylogenetic tree is a connected graph with no cycles such that each leaf, or vertex of degree one, is labeled with a different taxon. Given genetic data the objective is to produce a phylogenetic tree that best represents the historical relationships between different taxa. Trees can either be rooted or unrooted. In an unrooted tree the earliest ancestor is not identified whereas a rooted tree is a directed tree that illustrates ancestry of the given taxa. Since the models we are considering in this paper are time-reversible models the likelihood of a specific tree does not depend on the choice of the root vertex [18].

The evolutionary models we will examine use character sequences, which are sequences of fixed size that provide information about a specific taxon. Although evolutionary processes consist of mutations other than substitutions the models considered here make the simplifying assumption that there are no insertions or deletions,

and the sequences being compared are properly aligned. For the majority of this paper we will let each character come from the set of nucleotides that make up DNA. These nucleotides are the purine bases A (adenine) and G (guanine), and the pyrimidine bases T (thymine) and C (cytosine) and therefore each taxon is represented by a sequence of A, C, G, and Ts.[2]

Assumptions regarding the processes by which nucleotide substitutions are made define a substitution model. In this paper, the types of models considered are restricted to Markov models. Markov models have the property that the probability a site changes from base i to base j is independent from a site's earlier values. This implies the evolution of a nucleotide at a given site only depends on its immediate ancestral state.

Markov models can be defined by specifying an instantaneous rate matrix $\mathbf{Q}$, such that the entries of $\mathbf{Q}$, q_{ij}, give the rate at which state i will change to state j. For nucleotide substitution models $\mathbf{Q}$ is a 4×4 matrix that is indexed by {A, C, G, T}. We will use the order A C G T for our indices, but warn the reader that some papers use a different ordering. For amino acid or codon models the dimensions and indices would be adjusted accordingly.

There are a few constraints on the $\mathbf{Q}$ matrix to ensure that the probabilities of starting in a given state and either ending in a different state or ending in the same state add up to one. All entries q_{ij} such that $i \neq j$ must be nonnegative and diagonal entries $q_{ii} = -\sum_{j \neq i} q_{ij}$ so that row sums of $\mathbf{Q}$ are equal to zero.

For processes where the state space is finite the transition probabilities can also be represented by a matrix. The transition probability matrix $\mathbf{P}(t)$ is a matrix whose rows and columns are indexed by the states and whose $(i,j)^{th}$ entry is equal to $p_{ij} = Pr(Y_{n+1} = j | Y_n = i)$. [3]

By using Kolmogorov's backward equation $\frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t)$ with initial condition $\mathbf{P}(0) = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix it is possible to relate $\mathbf{P}(t)$ to $\mathbf{Q}$ in a single equation. The unique solution to this differential equation, subject to this initial condition, is $\mathbf{P}(t) = \exp(\mathbf{Q}t)$.

Most nucleotide substitution models used are time-reversible, which implies the probability of sampling nucleotide i from the stationary distribution and going to nucleotide j is the same as the probability of sampling nucleotide j from the stationary distribution and going to nucleotide i. It is important to observe that the rate matrices of time-reversible nucleotide substitution models are real symmetric matrices and real symmetric matrices are diagonalizable.

A well known nucleotide substitution model is the K3ST model. The K3ST model was introduced in 1981 by Motoo Kimura [12]. The K3ST model assumes there are three independent substitution rates; α is the rate of transitions and β and γ are the rates of the two types of tranversions. We will use α, β and γ to represent the

[2]For more information on character data see [18].

[3]For additional information on Markov models on trees see [16].

substitution rates and tr_α, tr_β and tr_γ to represent the substitution types. The K3ST model is typically given by the rate matrix $\mathbf{Q}$ indexed by the set $\{A, C, G, T\}$

$$\mathbf{Q} = \begin{bmatrix} -K & \gamma & \alpha & \beta \\ \gamma & -K & \beta & \alpha \\ \alpha & \beta & -K & \gamma \\ \beta & \alpha & \gamma & -K \end{bmatrix}$$

where $K = \alpha + \beta + \gamma$.

The Kimura two-parameter model, K2ST, is a submodel of the K3ST model and was introduced in 1980 [13]. The K2ST model can be obtained from the K3ST model by setting $\beta = \gamma$.

Another submodel of the K3ST model is the Jukes and Cantor model which was published in 1969 by T.H. Jukes and C.R. Cantor [11]. The Jukes and Cantor model is the most basic nucleotide substitution model. It can be obtained by letting $\alpha = \beta = \gamma$. A good reference summarizing the relationships between special cases of the general time-reversible model is figure 11 in [18].

In hopes of obtaining more biologically realistic models some researchers have looked at using models that require a larger number of states than the four state nucleotide substitution models. For instance, in 1994 two papers, one by Nick Goldman and Ziheng Yang [7] and another by Spencer Muse and Brandon Gaut [15], started discussion on the use of codon models of evolution. Using codon models of evolution would lead to evolutionary models containing between sixty-one and sixty-four states. Support for these models is provided in the 2007 paper [1]. In addition to considering codon substitution models it may also be of interest to consider amino acid substitution models. Disagreements in the number of amino acids that should be considered imply it may be of interest to develop amino acid substitution models with twenty, twenty-one or twenty-two states.

3 Hadamard conjugation: group-based models

One of the major limitations of Hadamard conjugation is its restriction to group-based substitution models [5]. David Bryant states that if the substitution model is assumed to be time-reversible then the only three-parameter group-based nucleotide substitution model is the K3ST model and the only nucleotide substitution models that are available to use with Hadamard conjugation are special cases of the K3ST model [5]. Therefore although there are 203 different time-reversible nucleotide substitution models only a handful are currently used with Hadamard conjugation [10].

The relationship between the Klein four group and the K3ST model was first published in the early 1990s. One of the first papers recognizing the group structure of the K3ST model was a 1993 paper by Steven Evans and T.P. Speed [6]. The paper

describes the relationship between the evolutionary model and the group by creating a correspondence between the bases $\{A, C, G, T\}$ and the elements of an abelian group with the group operation defined by the following addition table:

$+$	A	C	G	T
A	A	C	G	T
C	C	A	T	G
G	G	T	A	C
T	T	G	C	A

.

This group is isomorphic to the Klein four group, $\mathbb{Z}_2 \times \mathbb{Z}_2$, with one possible isomorphism given by $A \leftrightarrow (0,0)$, $C \leftrightarrow (0,1)$, $G \leftrightarrow (1,0)$ and $T \leftrightarrow (1,1)$. Each substitution type can also be associated to a group element by assigning to each substitution type the difference between the group element associated to the starting nucleotide and the group element associated to the ending nucleotide. For example if tr_γ is the substitution which takes A to C, C to A, G to T and T to G, then using the isomorphism above, tr_γ corresponds to the group element $(0,0) - (0,1) = (1,0) - (1,1) = (0,1)$. The substitution types tr_α, tr_β and tr_γ along with the identity substitution, tr_ϵ, produce a group under composition which acts on the nucleotide set $\{A, C, G, T\}$. From this perspective the fact that the K3ST model is abelian group-based means there exists a permutation group acting regularly on the four bases.

Given a group-based substitution model it is possible to apply Hadamard conjugation. Although there have been several different derivations all leading to the invertible formulas known as Hadamard conjugation, the utility of all of these formulas comes from the fact that they provide an analytic formula relating observed pattern frequencies from DNA data to a vector containing information about the structure of the phylogenetic tree [8], [5]. In general, analytic formulas which relate this information do not exist. The remainder of this paper examines the necessity of using group-based models with Hadamard conjugation.

4 Moving beyond group-based models

An existing theme of algebraic combinatorics has been the removal of hypotheses about having groups acting, with their successful replacement by regularity conditions that still guarantee the presence of an algebra. Instances of this include the move from distance-transitive graphs to distance-regular graphs and Don Higman's program of replacing permutation groups by coherent configurations, with the centralizer algebra being replaced by the Bose-Mesner algebra. It appears that a similar move may also be beneficial in the analysis of nucleotide substitution models. This is suggested by the fact that the set of possible instantaneous rate matrices, when expanded to allow the addition of linear combinations of the identity matrix, is isomorphic to the real group algebra of the Klein four group, V, that seems to be necessary for the application of Hadamard conjugation. This implies it may be useful

to move attention away from the group V and shift it to the group algebra $\mathbb{R}V$.

We may interpret a probability distribution on a group G as an element of the group algebra $\mathbb{R}$G, with coefficients of group elements between 0 and 1 and with the sum of all the coefficients being 1. Moreover, the distribution on a leaf is the product (in the group algebra $\mathbb{R}$G) of the distributions on each edge of the unique path from the root to the leaf. Thus, we may re-interpret these models in terms of algebras rather than groups, which, we will see, expands the number of models to which Hadamard conjugation applies.

We will refer to the set of possible instantaneous rate matrices expanded to allow the addition of linear combinations of the identity matrix of a given nucleotide substitution model as a Q space. For example, given the K3ST model whose rate matrix is given below

$$\mathbf{Q}^{(K3ST)} = \begin{bmatrix} -K & \gamma & \alpha & \beta \\ \gamma & -K & \beta & \alpha \\ \alpha & \beta & -K & \gamma \\ \beta & \alpha & \gamma & -K \end{bmatrix}$$

where $K = \alpha+\beta+\gamma$, we have a $Q^{(K3ST)}$ space equal to $\{\mathbf{Q}^{(K3ST)}+\delta\mathbf{I}|\alpha,\beta,\gamma,\delta \in \mathbb{R}\}$.

Theorem 4.1. *If there exists a real invertible four by four matrix X that simultaneously diagonalizes A, a three parameter Q algebra, then $A \cong \mathbb{R}V$ where V is the Klein four group.*

Proof. Let D be the set of all four by four diagonal matrices and A be a three parameter Q algebra. $XAX^{-1} \subset D$, and by comparing dimensions, equality occurs so that $XAX^{-1} = D$. Let Q' and Q'' belong to A. Then $Q'Q'' \in A$ since $XQ'X^{-1} \in D$ and $XQ''X^{-1} \in D$ which implies $XQ'X^{-1}XQ''X^{-1} = XQ'Q''X^{-1} \in D$. Therefore $Q'Q'' \in X^{-1}DX = A$. Therefore A is an algebra.

A is isomorphic to D via X and since $\mathbb{R}$V is a four dimensional algebra which is diagonalizable by $\mathbf{H}$, a Hadamard matrix, $\mathbb{R}V \cong D$. Therefore $A \cong \mathbb{R}V$. Notice that an algebra F is isomorphic to the group algebra FG if and only if there is a basis that under the algebra multiplication forms a group isomorphic to G. □

Given this result it is interesting to consider the M_{37} nucleotide substitution model which was published in [10]. The M_{37} model has rate matrix

$$\mathbf{Q}^{(M37)} = \begin{bmatrix} - & \alpha & \beta & \beta \\ \alpha & - & \beta & \beta \\ \beta & \beta & - & \eta \\ \beta & \beta & \eta & - \end{bmatrix}$$

with diagonal entries chosen so that the row sums are equal to zero. The three parameter $Q^{(M37)}$ space is isomorphic to $\mathbb{R}V$ where V is the Klein four group. Notice also that the M_{37} model is not a submodel of the K3ST model.

Just as K2ST is a submodel of the K3ST model and can be used with Hadamard conjugation, submodels of M_{37} can also be used with Hadamard conjugation.

Corollary 4.2. *Every subspace of a nucleotide substitution model in which all* $\mathbf{Q}$ *matrices are simultaneously diagonalizable is a submodel of a group algebra where the group is the Klein four group.*

Proof. This follows from theorem 4.1. To see this consider a three parameter Q algebra so that theorem 4.1 holds. Set parameters equal to each other. A is still simultaneously diagonalizable. □

Since the Q algebra must be simultaneously diagonalizable in order to produce the useful analytic formula Hadamard conjugation provides, the above result shows it is not possible to completely move away from using the abelian group. On the other hand the result only proves isomorphism, not equality, so it does not show that Hadamard conjugation is restricted to the K3ST model and submodels. As the following example will show, it allows the use of additional models with Hadamard conjugation.

5 A worked example

The following example illustrates the isomorphism between the K3ST model and the M_{37} model. To keep the size of the example manageable the tree assumed will be the 3-claw tree, which is isomorphic to $K_{1,3}$. This tree has three leaves and one internal vertex. We will assume the tree is rooted at one of the leaves. Recall that because the models we are considering are time reversible the placement of the root is independent of the likelihood of the tree.

To begin, consider the M_{37} substitution model whose rate matrix was given above and that has tr_α type substitutions occurring with probability 0.01, tr_β type substitutions occurring with probability 0.02, and tr_η type substitutions occurring with probability 0.03.

Given a transition probability matrix of the form

$$\mathbf{P} = \begin{bmatrix} p_0 & p_1 & p_2 & p_2 \\ p_1 & p_0 & p_2 & p_2 \\ p_2 & p_2 & p_3 & p_4 \\ p_2 & p_2 & p_4 & p_3 \end{bmatrix}$$

it is possible to use $\mathbf{P} = \exp(\mathbf{Q}t)$ to generate equations for p_i, $i \in \{0, 1, 2, 3, 4, 5\}$. From here it can be determined that $\alpha\ t \approx 0.0100922996$, $\beta\ t \approx 0.0208454022$ and $\eta\ t \approx 0.0318348556$.

It is generally not possible to estimate t; therefore, in order to separate t from the substitution rate an assumption is made. The average rate of substitution at equilibrium is set equal to 1. By doing this the length of a branch corresponds to the expected number of substitutions per site along that branch, rather than corresponding to the evolutionary time it represents [19]. Consequently, to solve for α, β, and γ the average of the off diagonal row sums of $\mathbf{Q}$ are set equal to 1. This implies that $(\alpha + 2\beta + \eta + 2\beta)/2 = 1$ and so $t \approx 0.06265438$ and $\alpha \approx 0.161078911$, $\beta \approx 0.332704617$, and $\eta \approx 0.508102619$.

Next the M_{37} model must be converted into the form of the K3ST model. To do this consider the M_{37} Q algebra, $M = \{Q^{(M37)} + \delta I | \alpha, \beta, \eta, \delta \in \mathbb{R}\}$, and $\mathbf{Y}$ which simultaneously diagonalizes the M_{37} algebra.

$$\mathbf{Y} = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & 0 & -\frac{1}{2} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$V = \{I, A, B, C\}$ is the Klein four group, where

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad C = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

and V spans the M algebra[4]. Therefore there is an isomorphism between the M algebra and the group algebra $\mathbb{R}V$. Notice that

$$\mathbf{Y}^{-1}(M)\mathbf{Y} = \{\text{set of diagonal matrices}\} = D$$

$$\mathbf{H}^{-1}(D)\mathbf{H} = K3ST\ Q \text{ algebra} = \{Q^{(K3ST)} + \delta I | \alpha, \beta, \gamma, \delta \in \mathbb{R}\}.$$

Conjugating by the matrix $\mathbf{YH}$ yields a matrix with the form of the K3ST model rate matrix,

$$\mathbf{H}^{-1}\mathbf{Y}^{-1}\mathbf{Q}^{(M37)}\mathbf{YH} = \begin{bmatrix} -\frac{1}{2}\alpha - \frac{1}{2}\eta - 2\beta & \frac{1}{2}\eta + \frac{1}{2}\alpha & -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta & \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta \\ \frac{1}{2}\eta + \frac{1}{2}\alpha & -\frac{1}{2}\alpha - \frac{1}{2}\eta - 2\beta & \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta & -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta \\ -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta & \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta & -\frac{1}{2}\alpha - \frac{1}{2}\eta - 2\beta & \frac{1}{2}\eta + \frac{1}{2}\alpha \\ \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta & -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta & \frac{1}{2}\eta + \frac{1}{2}\alpha & -\frac{1}{2}\alpha - \frac{1}{2}\eta - 2\beta \end{bmatrix}$$

[4]The presence of negative entries in the matrices in the Klein four group look troubling, however the isomorphism ensures that after the transformation back the results will be biologically meaningful.

with

$$\begin{array}{lcl} \frac{1}{2}\eta + \frac{1}{2}\alpha & \approx & 0.020963 \\ -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta & \approx & 0.017391 \\ \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta & \approx & 0.009974. \end{array}$$

Thus **YH** is an isomorphism between the M_{37} Q algebra and the $K3ST$ Q algebra. Also since $\mathbf{P}^{(K3ST)} = \exp(\mathbf{Q}t)$,

$$\mathbf{P}^{(K3ST)} = \begin{bmatrix} 0.94 & 0.02 & 0.03 & 0.01 \\ 0.02 & 0.94 & 0.01 & 0.03 \\ 0.03 & 0.01 & 0.94 & 0.02 \\ 0.01 & 0.03 & 0.02 & 0.94 \end{bmatrix}.$$

At this point it is possible to determine the leaf coloration probabilities for the K3ST model with the above probabilities. The probabilities are contained in the following vector.

$$\begin{array}{cc} AA & \begin{bmatrix} 0.83062 \end{bmatrix} \\ AC & 0.01806 \\ AG & 0.02736 \\ AT & 0.00896 \\ CA & 0.01806 \\ CC & 0.01806 \\ CG & 0.00104 \\ CT & 0.00104 \\ GA & 0.02736 \\ GC & 0.00104 \\ GG & 0.02736 \\ GT & 0.00104 \\ TA & 0.00896 \\ TC & 0.00104 \\ TG & 0.00104 \\ TT & \begin{bmatrix} 0.00896 \end{bmatrix} \end{array}$$

In the Kimura three-substitution type setting the group elements are:

$$g_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, g_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$g_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, g_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

while in the M_{37} setting the group elements are:

$$\hat{g}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \hat{g}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\hat{g}_3 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \hat{g}_4 = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Calculating $\sum_{i,j} p_{i,j}(\hat{g}_i, \hat{g}_j) = p_{11}(\hat{g}_1, \hat{g}_1) + p_{12}(\hat{g}_1, \hat{g}_2) + \ldots + p_{44}(\hat{g}_4, \hat{g}_4)$ yields

$$\left(\begin{bmatrix} 0.9034 & 0.0198 & 0.0384 & 0.0384 \\ 0.0198 & 0.9034 & 0.0384 & 0.0384 \\ 0.0384 & 0.0384 & 0.8666 & 0.0566 \\ 0.0384 & 0.0384 & 0.0566 & 0.8666 \end{bmatrix}, \begin{bmatrix} 0.9034 & 0.0198 & 0.0384 & 0.0384 \\ 0.0198 & 0.9034 & 0.0384 & 0.0384 \\ 0.0384 & 0.0384 & 0.8666 & 0.0566 \\ 0.0384 & 0.0384 & 0.0566 & 0.8666 \end{bmatrix} \right).$$

Notice that the $(ij)^{th}$ entry of these matrices give the probability of starting at state i going through an intermediate state and ending at state j assuming the M_{37} model. For example the probability of the root of a path of length two being in state T and the leaf in state C is equal to $(0.02)(0.01)+(0.02)(0.95)+(0.03)(0.02)+(0.93)(0.02) = 0.0384$ given the M_{37} probabilities assumed at the beginning of this example. This value is equal to the T, C entry in the matrices above.

Theorem 4.1 and corollary 4.2 along with the previous example show that a Klein four group must be present in order for a nucleotide substitution model to allow simultaneous diagonalization of the Q-matrices and therefore allow the full force of Hadamard conjugation to apply. The necessary distinction that needs to be made is that although the Klein four group must be present, it need not be present as a group of automorphisms of the model. The automorphism group is the centralizer in the symmetric group on the states of the set of Q-matrices. The Klein four group for the M_{37} model is found instead in the centralizer in the general linear group of degree 4 of the set of Q-matrices, and this is sufficient to be able to use Hadamard conjugation. This results in an expansion of the number of time-reversible nucleotide substitution models that can be used with Hadamard conjugation.

6 Prospects

Recall that the rate matrices for general time-reversible models are real symmetric matrices and therefore are diagonalizable. Recall also that a set of square diagonalizable matrices commute if and only if they are simultaneously diagonalizable and consequently, if the set of rate matrices of an evolutionary model correspond to a commutative algebra they must be simultaneously diagonalizable. The ability

to simultaneously diagonalize a set of matrices makes it possible to apply Hadamard conjugation. The question remaining however, is how to find an appropriate commutative algebra. One way is to use an association scheme.

If $\mathcal{A}$ is the linear span over the reals of the adjacency matrices $A_0, \ldots, A_s$ of an association scheme then $\mathcal{A}$ forms an algebra known as the Bose-Mesner algebra of the scheme. This implies that if the set of rate matrices form an association scheme there exists a commutative algebra.

It is possible to look at the form of $\mathbf{Q}$ and determine if the model corresponds to an association scheme. The following result shows the correspondence between certain nucleotide substitution models and association schemes on four points

Theorem 6.1. *A time-reversible s-parameter nucleotide substitution model with rate matrix* $\mathbf{Q}$ *such that all entries* Q_{ii} *are equal for* $1 \leqslant i \leqslant 4$ *and* $Q_{ii} \neq Q_{ij}$ *for* $i \neq j$*, corresponds to an association scheme with s associate classes on a set* $\mathfrak{X} = \{A, C, G, T\}$.

The above result implies that the K2ST and K3ST models must correspond to association schemes on four points.

Other types of time-reversible evolutionary models with a larger number of states also exist. Due to the larger number of states in amino acid and codon models the theory of when Hadamard conjugation applies becomes slightly different. For example, complex numbers are required if the number of states is not a power of two. The increased number of states also raises the question of how many parameters should be included in a model in order to accurately model the biological process without including extraneous parameters. Attempting to use a group-based codon model could result in a model with as many as sixty-three parameters, if a group of order 64 is used. In general the number of parameters is equal to one less than the order of the group. Using other techniques to construct the evolutionary models, such as obtaining models from association schemes with few associate classes, provides a method of obtaining models with the number of parameters equal to the number of associate classes. Such models currently lack biological realism, but provide a source of potential models. The desire is to find balance between biologically motivated models and models that are mathematically tractable.

Theorem 6.2. *If there exists an invertible* $n \times n$ *matrix* X *that simultaneously diagonalizes* $\bar{A} = A \otimes \mathbb{C}$*, an* $n-1$ *parameter* Q *space, then* $\bar{A} \cong \mathbb{C}V$*, where* V *is a cyclic group of order* n.

Proof. Let D be the set of all n by n diagonal matrices and $\bar{A}$ be an $n-1$ parameter Q algebra. $X\bar{A}X^{-1} \subset D$, and by comparing dimensions, equality occurs so that $X\bar{A}X^{-1} = D$. Let Q' and Q'' belong to $\bar{A}$. Then $Q'Q'' \in \bar{A}$ since $XQ'X^{-1} \in D$ and $XQ''X^{-1} \in D$ which implies $XQ'X^{-1}XQ''X^{-1} = XQ'Q''X^{-1} \in D$. Therefore $Q'Q'' \in X^{-1}DX = \bar{A}$. Therefore $\bar{A}$ is an algebra.

$\bar{A}$ is isomorphic to D via X and since $\mathbb{C}V$ is a n dimensional algebra which is diagonalizable, $\mathbb{C}V \cong D$. Therefore $\bar{A} \cong \mathbb{C}V$. □

Corollary 6.3. *Every n state evolutionary model in which all Q matrices are simultaneously diagonalizable is a submodel of a group algebra where the group is cyclic of order n.*

Proof. The set of matrices $X\bar{A}X^{-1}$ is contained in D the set of n by n diagonal matrices. $D \cong \mathbb{C}V$, where V is a cyclic group of order n. Since $X\bar{A}X^{-1} \subseteq D$, $\bar{A} \subseteq X^{-1}DX \cong \mathbb{C}V$. Therefore $\bar{A}$ is a submodel of $\mathbb{C}V$. □

The above theorems show that even for larger evolutionary models there is still an abelian group present. That does not mean however, that there is an abelian permutation group acting regularly on the bases. Currently Hadamard conjugation is only applied to evolutionary models in which an abelian permutation group acting regularly on the bases exists. It appears however, that it is possible to use the structure provided by an association scheme, or more generally a commutative algebra in order to apply Hadamard conjugation.

Association schemes can be used with other types of evolutionary models to produce models that do not rely on an abelian permutation group acting regularly. For instance, finding an association scheme on twenty points can lead to an amino acid substitution model that does not rely on a group, yet has simultaneously diagonalizable rate matrices for which Hadamard conjugation would apply.

Each association scheme on a given number of points corresponds to a class of evolutionary models. Association schemes on twenty to twenty-two points will correspond to amino acid models while association schemes on sixty-one to sixty-four points correspond to codon models of evolution. To see the correspondence between an association scheme and an evolutionary model consider an association scheme on n points. Each of the n vertices of the graph corresponding to the association scheme can be labeled with an amino acid or codon. Different labelings will produce biologically distinct models, which is why given one association scheme we end up with a class of models.

The instantaneous rate matrix is produced from an association scheme by introducing parameters α_k for each associate class R_k and setting $Q_{ij} = \alpha_k$ if and only if $(i, j) \in R_k$. The entries Q_{ii} are chosen so that row sums of $\mathbf{Q}$ are zero. Given this construction it is clear that choosing an association scheme with a small number of associate classes will lead to a model with a small number of parameters.

In conclusion, there is a rich source of codon-based models of evolution arising from algebraic combinatorics and these should be examined to see if one or more can be found that are biologically reasonable.

7 Future Work

The work presented above seeks to expand the number of nucleotide substitution models that can be used with Hadamard conjugation. To demonstrate how this can

be done we presented an example that illustrates the isomorphism between K3ST and the M_{37} model. At this point it would be of interest to expand the example and examine the mapping from the expected sequence spectrum to the edge length spectrum. The biological realism of this model and other models isomorphic to K3ST should also be considered.

Additionally, given the connection between association schemes and evolutionary models it would be interesting to consider the question of whether codon models corresponding to a association scheme could be developed so that they are biologically meaningful.

It would also be of interest to determine whether the approach of Sumner et al. can be usefully generalised in the light of the work in this paper. They suggest complex models similar to the General Markov Model using their approach, and their main issue is multiplicative closure of the matrix group that is a model of evolution, and the association scheme approach addresses this issue from a different angle.

References

[1] A. Ambrogelly, S. Palioura and D. Söll, Natural expansion of the genetic code, *Nat. Chem. Biol.* **3 (1)** (2007), 29–35.

[2] R.A. Bailey, *Association Schemes, Designed Experiments, Algebra and Combinatorics*, Cambridge University Press, 2004.

[3] E. Bannai and T. Ito, *Algebraic Combinatorics I: Association Schemes*, Benjamin-Cummings, 1984.

[4] R.C. Bose and T. Shimamoto, Classification and Analysis of Partially Balanced Incomplete Block Designs with Two Associate Classes, *JASA* **47 (258)** (1952), 151–184.

[5] D. Bryant, Hadamard Phylogenetic Methods and the n-taxon Process, *Bull. Math. Biol.* **71** (2009), 339–351.

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

[7] N. Goldman and Z. Yang, A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences, *Mol. Biol. Evol.* **11 (5)** (1994), 725–736.

[8] M.D. Hendy and D. Penny, A framework for the study of evolutionary trees, *Syst. Biol.* **38** (1989), 297–309.

[9] D.G. Higman, Coherent configurations, *Geometriae Dedicata* **4 (1)** (1975), 1–32.

[10] J.P. Huelsenbeck, B. Larget and M.E. Alfaro, Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo, *Mol. Biol. Evol.* **21 (6)** (2004), 1123–1133.

[11] T.H. Jukes and C.R. Cantor, *Evolution of Protein Molecules*, Academy Press, 1969.

[12] M. Kimura, Estimation of evolutionary distances between homologous nucleotide sequences, *Proc. Natl. Acad. Sci.* **78 (1)** (1981), 454–458.

[13] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.* **16 (2)** (1980), 111–120.

[14] T.W. Körner, *Fourier Analysis*, Cambridge University Press, 1988.

[15] S.V. Muse and B.S. Gaut, A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome, *Mol. Biol. Evol.* **11 (5)** (1994), 715–724.

[16] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2005.

[17] J.G. Sumner, J. Fernández-Sánchez and P.D. Jarvis, Lie Markov models, *J. Theor. Biol.* **298** (2012), 16–31.

[18] D.L. Swofford, G.J. Olsen, P.J. Waddell and D.M. Hillis, Phylogenetic Inference, In: David M. Hillis, Craig Moritz and Barbara K. Mable, eds., *Molecular Systematics*, Sinauer Associates, Inc., 1996.

[19] Z. Yang, Estimating the pattern of nucleotide substitution, *J. Mol. Evol.* **39 (1)** (1994), 105–111.

(Received 22 Apr 2015; revised 27 May 2016)