

DISSERTATION

SOME TOPICS IN COMBINATORIAL PHYLOGENETICS

Submitted by

Cayla D. McBee

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado


Summer 2010

COLORADO STATE UNIVERSITY

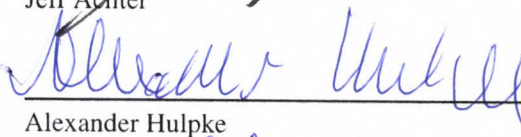
May 13, 2010

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY CAYLA D. MCBEE ENTITLED SOME TOPICS IN COMBINATORIAL PHYLOGENETICS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

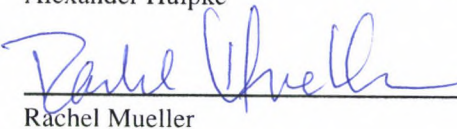
Committee on Graduate Work



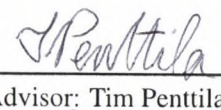
Jeff Achter



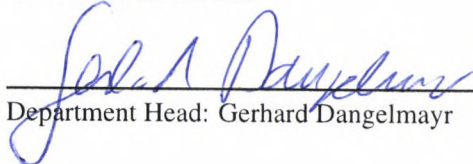
Alexander Hulpke



Rachel Mueller



Advisor: Tim Penttila



Department Head: Gerhard Dangelmayr

ABSTRACT OF DISSERTATION

SOME TOPICS IN COMBINATORIAL PHYLOGENETICS

This thesis is in combinatorial phylogenetics and is focused on a study of Hadamard conjugation. It examines the question of whether the presence of an abelian permutation group acting regularly on the states is necessary for the application of this technique. New connections between phylogenetics and algebraic combinatorics are suggested, especially with (commutative) association schemes.

Cayla D. McBee
Department of Mathematics
Colorado State University
Fort Collins, CO 80523
Summer 2010

TABLE OF CONTENTS

1	Introduction	1
1.1	Mathematical Introduction	4
1.1.1	Graphs and Phylogenetic Trees	4
1.1.2	Representing Phylogenetic Trees	6
1.1.3	Splits Networks	8
1.1.4	Hadamard Matrices	13
1.1.5	Hadamard Transform	14
1.1.6	Continuous-time Markov Processes	15
1.1.7	Matrix Exponential	16
1.1.8	Coherent Configurations and Association Schemes	17
	Coherent Configurations	18
	Association Schemes	19
1.1.9	Strongly regular and distance regular graphs	22
1.1.10	Johnson Scheme	23
1.1.11	Bose-Mesner Algebra	25
1.1.12	Results from Linear Algebra	25
1.2	Biological Introduction	26
1.2.1	Evolution	26
1.2.2	DNA and RNA	29
2	Models of Evolution	33
2.1	An introduction to Nucleotide Substitution Models	34
2.1.1	Assumptions of many nucleotide substitution models	34
2.1.2	General nucleotide substitution model	35
2.1.3	General time-reversible model	36
2.1.4	Tamura and Nei Model	37
2.1.5	Hasegawa-Kishino-Yano Model	38
2.1.6	Kimura's Three-Substitution type Model	39
2.1.7	Kimura's Two-parameter Model	40
2.1.8	Jukes and Cantor Model	40
2.1.9	Strand Symmetric Models	41
2.2	Codon Models of Evolution	42

3	Relating DNA sequence data to Phylogenetic Trees	44
3.1	The observed sequence spectrum vector $f(D)$	44
3.1.1	Two state substitution models and bipartitions	44
3.1.2	Sequence Spectrum for a model with three substitution types	47
	Site Patterns	47
3.1.3	Relating the observed sequence spectrum to a tree	50
3.2	The edge length spectrum as it relates to the expected sequence spectrum	51
3.2.1	Preliminary Notation	52
3.2.2	Notation related to phylogenetic trees	53
	The A Matrix	54
	Leaf colorations	55
	Equations	55
3.2.3	Results	56
3.2.4	An example assuming the Jukes-Cantor model	59
4	Hadamard Conjugation and Splits Networks: An Extension to Splits Networks	64
4.1	Notation	64
4.1.1	A Matrix	65
4.1.2	Leaf Colorations	66
4.1.3	Preliminary Result	67
4.2	An illustrated example	68
4.3	A Result	72
5	A Different Perspective on Hadamard Conjugation	81
5.1	Bryant's approach	81
5.1.1	Models of evolution along one branch	82
5.1.2	The n -taxon process	82
	Terminology and Notation	83
	Branch Rate Matrices	83
	Diagonalizing a branch rate matrix	85
	Transition probabilities over multiple lineages	87
5.1.3	Example	88
	Jukes-Cantor model using Bryant 2009	89
6	Extending Hadamard Conjugation	91
6.1	Group-based substitution models	91
6.2	Simultaneous Diagonalization	95
6.3	Substitution Models and Groups	96
6.3.1	An example with model M_{37}	98
6.4	Algebras and Association Schemes	103
6.5	Larger evolutionary models	108
	Association Scheme on 20 points	111
	Association Scheme on 21 points	111
	Association Scheme on 22 points	111

	Association Scheme on 61 points	112
	Association Schemes on 62, 63, and 64 points	112
6.6	Networks and Hadamard conjugation	112
6.7	Strand Symmetric Models and Hadamard Conjugation	113
6.7.1	Conclusion	115
7	Problems Raised	117
A		118
A.1	Hierarchy	118
A.2	An indexing error	119
A.3	A counterexample to Bryant's Lemma	119

Chapter 1

INTRODUCTION

Biologists have been interested in Phylogenetics, the study of evolutionary relatedness among various groups of organisms, for more than 140 years. It has only been in the last 40 years that advances in technology and the availability of DNA sequences have led to statistical, computational and algorithmic work in determining evolutionary relatedness between species. Previously morphology, the outward appearance (shape, structure, color, and pattern) of an organism, was the only way to determine the historical relationships between groups of organisms or taxa ¹. Morphology relies on the fact that typically, closely related taxa differ less than distantly related ones, however there are exceptions to this generalization. For example the Earthworm and Flatworm have similar morphology but are contained in different taxa. Phylum is the taxon below kingdom and above class and the Earthworm is contained in the Phylum Annelida while the Flatworm is contained in the Phylum Platyhelminthes. Partly, as a result of these exceptions scientists are looking to DNA analysis to classify related organisms.

A 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 genetic data available for presently extant species to determine historical relationships. Not surprisingly, phylogenetic trees are common in evolutionary biology where biologist use the trees to classify new species and determine relationships between known species. Applications

¹For a list of the major taxonomic ranks see page 118

of phylogenetic trees are not restricted to evolutionary biology, instead the construction of phylogenetic trees helps researchers answer questions in a variety of fields. For example, evolutionary analysis of how a virus, such as the influenza virus, evolves helps immunologists develop new vaccines. Phylogenetic analysis has also been used to gain insight into the Human Immunodeficiency Virus (HIV) [50]. Techniques used to create phylogenetic trees have even been used by linguists and anthropologists to determine how various cultures have spread across the globe [45].

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 $P = \exp(Qt)$ where P and 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 formulae relating relative frequencies of observed patterns from the genetic data to an estimation of the phylogenetic tree that corresponds to the data.

In addition to providing the analytic formulae Hadamard conjugation also corrects for the number of observed differences between genetic sequences. Since the succession of two or more substitutions between sequences will not be observed directly, the number of observed differences between sequences will be underestimated. In 1981 Motoo Kimura [44] derived a correction to estimate the number of substitutions taking place in a three-substitution nucleotide model. It can be shown that this correction can be obtained using Hadamard conjugation².

The usefulness of Hadamard conjugation also stems from the fact that it provides a method to determine how likely an edge in a phylogenetic tree is to exist. Due to assumptions made in the evolutionary models, errors in the data and events such as hybridization and horizontal gene transfer it is unlikely that exactly one phylogenetic tree will fit the genetic data perfectly.

²For more information on this see [31]

The use of Hadamard conjugation allows for examination of the data to determine how tree-like it is. At times complex evolutionary scenarios are poorly described by models assuming a tree and situations motivate the use of phylogenetic networks rather than phylogenetic trees. Phylogenetic networks are similar to phylogenetic trees in that vertices represent taxa while edges represent evolutionary relationships, however networks differ from trees in that they allow cycles. There are several types of networks used to display phylogenetic data. One type of phylogenetic network is a splits network. Splits networks serve several purposes such as to represent incompatibilities within and between data sets, to summarize a large collection of trees resulting from multiple gene analysis, and to help determine if an incorrect evolutionary model was used. Due to the possible benefits of using splits networks I extended the analysis of Székely et al [60] from phylogenetic trees to phylogenetic splits networks.

Despite the usefulness of Hadamard conjugation it is limited by the number of evolutionary models to which it can be applied. One of the objectives of this paper is to examine when Hadamard conjugation can be used. Currently in the literature Hadamard conjugation has been used with group-based evolutionary models or submodels of group-based models. I claim that depending on the type of genetic data being considered it is not necessary for the evolutionary model to satisfy the definition of a group-based model as given in the literature.

The first chapter of this paper contains the mathematical and biological background useful to understanding Hadamard conjugation and the discussion of phylogenetic inference. Chapters 2 and 3 discuss evolutionary models as well as how information obtained from genetic sequences can be used to determine the most likely phylogenetic tree given the data. In chapter 4 a result from Székely et al [60] is extended to apply to phylogenetic networks and chapter 5 considers a different perspective on Hadamard conjugation. Chapter 6 examines whether the presence of an abelian permutation group acting regularly is necessary to apply Hadamard conjugation. It also considers new connections between phylogenetics and combinatorics. Finally, chapter 7 raises questions for future work.

1.1 Mathematical Introduction

The following sections include a mathematical introduction and references as well as vocabulary necessary to discuss topics which arise later in the paper.

1.1.1 Graphs and Phylogenetic Trees

A *taxon* is a group of organisms which are inferred to be phylogenetically related and have characteristics which set them apart from other organisms. Historical relationships between taxa are often displayed on graphs with specific properties. The following section will introduce terminology to describe these graphs. It will also include a discussion of splits which have played a large role in the mathematical development of combinatorial phylogenetics.

Definition 1. [51] A **graph** G is an ordered pair (V, E) consisting of a non-empty set V of vertices and a set E of edges each of which is an element of $\{\{x, y\} : x, y \in V\}$. The set of edges in a given graph G is denoted by $E(G)$.

The biological meaning behind the graphs used in combinatorial phylogenetics implies that the graphs which arise are *simple*. Simple graphs are undirected graphs with the properties that no edge joins a single vertex to itself and no more than one edge joins any pair of vertices. In addition to considering simple graphs, the assumption that there are no cycles in the graph is often made. A sequence of vertices joined by edges such that there exists an edge connecting each vertex in the sequence to the next vertex in the sequence is known as a *path*. A path in a graph that begins and ends with the same vertex is known as a *cycle*.

If two vertices v_1 and v_2 in a graph G are joined by an edge $e \in E(G)$, v_1 and v_2 are said to be *adjacent* or *neighbors* and edge e is said to be *incident* with v_1 and v_2 .

Definition 2. [25] The **degree** (or **valence**) of a vertex v in a graph G , denoted $deg(v)$, is the number of edges incident with v .

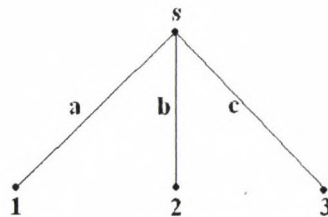
A vertex with degree one is known as a *pendant* vertex. Pendant vertices play an important role in the graphs which arise in combinatorial phylogenetics. *Trees*, connected graphs with no cycles, are also important in combinatorial phylogenetic. Trees contain pendant vertices and these pendant vertices are often referred to as *leaves*. Any vertex in a tree that is not a leaf is known as an *internal vertex*.

It is usually assumed in phylogenetics that the trees discussed are binary trees. Binary trees have the property that every internal vertex is of degree three. Scientists use binary trees to organize related organisms by assigning taxa to the leaves of the tree and letting the edges in the tree represent evolutionary relationships between the taxa. Binary trees labeled in this fashion are known as *phylogenetic trees*. The taxa are generally represented by DNA sequences for a particular gene. These sequences are made up of nucleotides, the structural units of DNA. The hope is that information about the phylogenetic tree can be gained by looking at homologous nucleotide sequences, or nucleotide sequences which descend from a common ancestral sequence.

A *rooted* phylogenetic tree is a directed tree with a distinguished vertex r , called the root, such that for every other vertex v , there exists a unique directed path from r to v . The taxon assigned to the root is the taxon from which all other taxa on the tree evolved. Phylogenetic trees that do not have this property are known as unrooted trees.

The following example demonstrates some of the terminology introduced above.

Example 1. Consider the binary tree provided below.



Vertices 1,2 and 3 are the leaves of the tree while vertex s is an internal vertex of degree 3.

1.1.2 Representing Phylogenetic Trees

Phylogenetic trees can be uniquely represented by a specific collection of subsets of leaves. These subsets are known as *splits* and are formed by deleting an edge of the tree T . Since T is a tree and contains no cycles deleting an edge creates two disjoint subgraphs such that each leaf of T will be contained in exactly one subgraph. The partition of the set of leaves into two disjoint subsets is the split corresponding to the removed edge. It is important to notice that each edge removed from T creates a unique split.

Suppose the leaves of a tree T are labeled $1, 2, \dots, n$ where leaf n is the root vertex. When an edge is removed from T exactly one subset of leaves, or split half, will contain the root vertex n . Let the split half not containing the root n be labeled A . Let the set $\sigma(T)$ be composed of all sets A corresponding to the edges in T . The following properties hold for the set σ :

- (i) $\{1, 2, \dots, n-1\} \in \sigma$ and $\{i\} \in \sigma$ for all $i \in \{1, 2, \dots, n-1\}$.
- (ii) If $\beta, \beta' \in \sigma$ then $\beta \cap \beta' \in \{\beta, \beta', \emptyset\}$. [54]

In addition to obtaining a set of splits from a tree it is desirable to obtain a tree from a given set of splits. In 1971 Peter Buneman³ [10] stated a theorem which provides further information regarding the relationship between sets of splits and phylogenetic trees.

Theorem 1. [54] *Any collection σ , of non-empty subsets of $\{1, 2, \dots, n-1\}$ which satisfy (i) and (ii) corresponds to $\sigma(T)$ for a unique unrooted phylogenetic tree T on $\{1, \dots, n\}$. Furthermore, this tree can be recovered from σ in polynomial time.*

This theorem implies that any tree can be described by its set of splits and that given a set of splits satisfying the above conditions a unique tree can be produced. If a set of splits satisfies the above conditions the set of splits is *compatible*. An equivalent definition of a compatible split system is given below.

³Peter Buneman is a computer scientist who works in the areas of database systems and database theory.

Definition 3. Let the edge e of the tree T define a split $S = \{A, A'\}$ where A and A' are two disjoint subsets of leaves such that $A \cup A' = \{1, 2, \dots, n\}$. A system σ of such splits is called **compatible**, if for any two splits $S_1 = \{A_1, A'_1\}$ and $S_2 = \{A_2, A'_2\}$ in σ one of the four intersections

$$A_1 \cap A_2, \quad A_1 \cap A'_2, \quad A'_1 \cap A_2, \quad A'_1 \cap A'_2$$

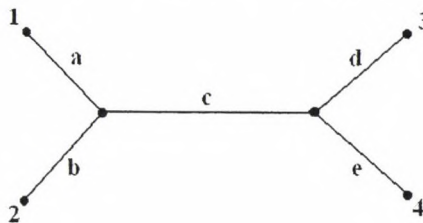
is empty.

Many tree reconstruction methods use splits because of the relationship between phylogenetic trees and compatible split systems. One such method which uses this relationship is Hadamard conjugation. Hadamard conjugation, which will be discussed in more detail later in the paper, takes a vector containing information regarding the genetic data representing the taxa being considered and transforms it into a vector containing values that indicate which splits are most likely to exist in the phylogenetic tree. Ideally the splits resulting from Hadamard conjugation are compatible and therefore correspond to a phylogenetic tree. In the case where the splits are not compatible a fitting algorithm is often used to determine which compatible set of splits best fits the data. An alternative approach to fitting data to a tree is to represent the data with a splits network. Splits networks will be discussed in section 1.1.3.

The following examples illustrate the relationship between phylogenetic trees and splits.

Example 2. The splits from the tree given in example 1 on page 5 are produced by removing edges a , b , and c . The splits produced are $\{1\}|\{2, 3\}$, $\{1, 3\}|\{2\}$, and $\{1, 2\}|\{3\}$, respectively. In general only the split halves not containing the root vertex are recorded. If leaf 3 is the root the splits are $\{1\}$, $\{2\}$ and $\{1, 2\}$.

Example 3. Below is a tree with leaves 1, 2, 3 and 4 and edges a, b, c, d , and e .



Let vertex 4 be the root vertex. The splits produced by removing edges a, b, c, d and e are $\{1\}$, $\{2\}$, $\{1, 2\}$, $\{3\}$, and $\{1, 2, 3\}$ respectively. Notice that the split $\{2, 3\}$ is not a split of this tree because there is no edge, which when removed partitions the graph into subgraphs containing leaves $\{1, 4\}$ and $\{2, 3\}$. Out of the possible eight subsets containing 1, 2 and 3 only five correspond to edges in the graph.

The ability to uniquely describe a phylogenetic tree by a set of splits has made splits extremely important in phylogenetic tree reconstruction methods. The connection between a set of splits and a phylogenetic tree has allowed splits to be used to describe the structure of a phylogenetic tree as well as used to index vectors containing information about the nucleotide sequences being considered.

Given the importance of splits in the setting of phylogenetic trees it is natural to think about splits in phylogenetic networks. Namely how would they be defined and are they useful in phylogenetic network reconstruction methods? The next section discusses splits in the context of splits networks, a specific type of phylogenetic network.

1.1.3 Splits Networks

Complex evolutionary scenarios are often poorly described by evolutionary models assuming a tree. Although many situations are well described using phylogenetic trees there are both biological and statistical motivations for studying phylogenetic networks. Phylogenetic networks are similar to phylogenetic trees in that vertices represent taxa while edges represent evolutionary relationships, however networks differ from trees in that they allow cycles. Several types of phylogenetic networks exist and the different networks attempt to explain different types of events. For example reticulate networks, which tend to look like phylogenetic trees with extra edges added, are used to explain events such as hybridization, genetic recombination and horizontal gene transfer. Hybridization is the interbreeding between two animals or plants of different taxa while genetic recombination is the process by which a strand of genetic material

is broken and joined to a different DNA molecule. Horizontal gene transfer occurs when an organism incorporates genetic material from another organism without being the offspring of that organism. A phylogenetic tree cannot appropriately describe any of these biological processes and therefore phylogenetic networks are necessary.

A second type of phylogenetic network that will be examined in this paper is a splits network. Splits networks serve several purposes such as to represent incompatibilities within and between data sets, to summarize a large collection of trees resulting from multiple gene analysis, and to help determine if an incorrect evolutionary model was used.

In order to define a splits network a few preliminary definitions are necessary. The first is that of a splits graph. There are a few characterizations of splits graphs. The one used here is attributed to Wetzel [62]. Given a graph G with vertices u and v let $d_G(u, v)$ denote the minimal number of edges in any path from vertex u to vertex v . Next, let κ be an edge coloring of G . For each pair of vertices u, v let $C_\kappa(u, v)$ denote the set of colors that appear on every shortest path from u to v .

Definition 4. [8] κ is an *isometric coloring* of the graph G if $d_G(u, v) = |C_\kappa(u, v)|$ for all pairs $u, v \in V(G)$.

This implies that if there exists an isometric coloring then any two shortest paths between the same pair of vertices have the same set of edge colors and that the colors along any shortest path between vertices are distinct.

Definition 5. A connected graph is a *splits graph* if and only if it has an isometric coloring [62].

At this point it is possible to define a splits network. Suppose G is a splits graph and there exists a map $\phi: X \rightarrow V(G)$ where X is a finite set.

Definition 6. [8] A *splits network* is a pair $\mathcal{N} = (G, \phi)$ such that

1. G is a splits graph,

2. each color class induces a distinct split of X .

In biological applications X is the set of taxa being considered. Vertices of degree one in a splits network are labeled by taxa just as the leaves of a phylogenetic tree are labeled by taxa. A splits network has the appearance of a phylogenetic tree with parallel edges. The following is an example of a splits network.

Example 4. Consider the set of set of four taxa $\{1, 2, 3, 4\}$ labeled on the splits network below.



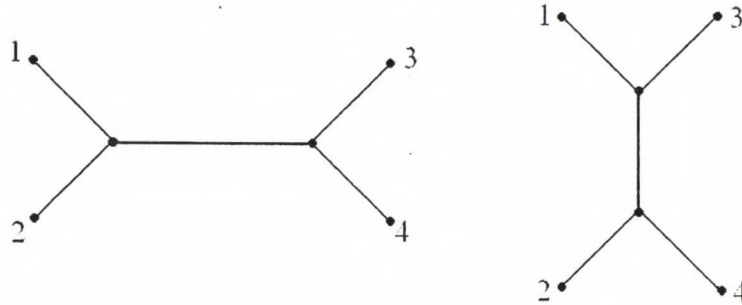
The different shades of gray represent an isometric coloring of the graph. The colors of the edges along any shortest path between vertices are distinct and any two shortest paths between two distinct vertices contain the same set of edge colors. Although there can be more than one shortest path between vertices in a splits network, the paths are essentially indistinguishable. Notice that the graph resulting from removing all edges of a particular shade of gray consists of precisely two connected components and therefore each color class induces a distinct split of the vertices.

In a phylogenetic tree each split corresponds to exactly one edge. In the case of a splits network, splits are formed by removing edges of the same color class. For example removing the lightest gray edges from the splits network above results in the split $\{1, 2\}|\{3, 4\}$. There are 6 color classes in the splits network above corresponding to six splits. The splits are:

$$\{\{1\}|\{2, 3, 4\}, \{1, 2\}|\{3, 4\}, \{1, 3\}|\{2, 4\}, \{1, 3, 4\}|\{2\}, \{1, 2, 4\}|\{3\}, \{1, 2, 3\}|\{4\}\}$$

Notice that this is not a compatible set of splits. For example, consider the pair of splits $\{1,2\}|\{3,4\}$ and $\{1,3\}|\{2,4\}$. By definition 3 on page 7 in order to be compatible one of the four intersections $\{1,2\} \cap \{1,3\}$, $\{1,2\} \cap \{2,4\}$, $\{3,4\} \cap \{1,3\}$, $\{3,4\} \cap \{2,4\}$ must be empty, however this is not the case. Since the splits given above are not compatible they do not correspond to a unique phylogenetic tree, instead the subsets of the splits generated from the splits network in example 4 represent two different trees. The two trees are given in the next example.

Example 5. *The two trees below can each be represented by a subset of splits from the splits network in example 4.*



The tree on the left side can be described by the splits

$$\{\{1\}|\{2,3,4\}, \{1,2\}|\{3,4\}, \{1,3,4\}|\{2\}, \{1,2,4\}|\{3\}, \{1,2,3\}|\{4\}\}$$

while the tree on the right can be described by

$$\{\{1\}|\{2,3,4\}, \{1,3\}|\{2,4\}, \{1,3,4\}|\{2\}, \{1,2,4\}|\{3\}, \{1,2,3\}|\{4\}\}.$$

The ability to represent multiple phylogenetic trees in one splits network has several advantages. In determining the historical relationships between a set of taxa it is likely that multiple genes will be analyzed. It is also likely that looking at different genes will produce different phylogenetic trees. When multiple trees arise they can be summarized by a single splits network.

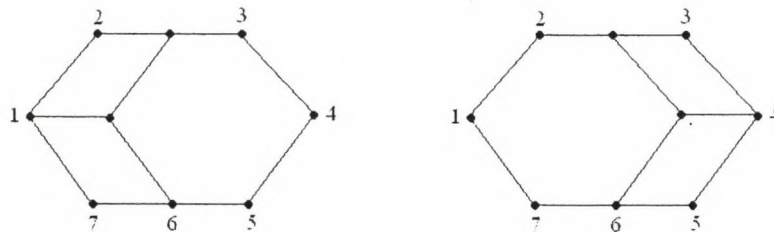
In addition to representing multiple trees, splits networks can also be used to determine if a systematic error, or mistake in the assumptions of a model, has occurred. The two most common types of errors are systematic and sampling errors. Sampling error is the random error resulting from a small sample size. With improvements in technology and the possibility of analyzing larger nucleotide sequences sampling error is becoming less of a concern than it was in the past. The hope is that if an appropriate evolutionary model is chosen the resulting splits network will be close to treelike. If the resulting splits network has a high number of parallel edges relative to the number of taxa being considered it is possible a systematic error occurred.

A third benefit of using splits networks is the ability to extract phylogenetic signals missed by tree-based methods. If the set of splits produced using a tree-based method is not compatible a fitting algorithm is used to choose the most likely compatible set of splits. This requires that at least one split must be ignored. In the case of splits networks all splits can be displayed at once even when they are not compatible.

Unfortunately there is also a disadvantage to using splits networks instead of phylogenetic trees. A compatible split system uniquely describes a phylogenetic tree, however this is not the case for splits networks. Instead, one split system can describe more than one splits network as shown in the next example.

Example 6. [8] Consider the set of splits

$$\mathcal{S} = \{ \{1, 2, 3\} | \{4, 5, 6, 7\}, \{2, 3, 4\} | \{1, 5, 6, 7\}, \\ \{1, 2, 7\} | \{3, 4, 5, 6\}, \{1, 2, 6, 7\} | \{3, 4, 5\} \}$$



Despite the fact that a set of non-compatible splits may correspond to more than one splits network, splits and splits networks are still useful. Splits can be defined for splits networks and can be used in the construction of splits networks representing the historical relationships of a set of nucleotide sequences.

The focus in the next several subsections will provide a basic introduction to some of the mathematics used in creating phylogenetic trees and networks.

1.1.4 Hadamard Matrices

Several nucleotide substitution models utilize a technique known as Hadamard conjugation. This technique will be introduced in section 1.1.5 and examined more closely throughout the paper. In order to use Hadamard conjugation the definition of a Hadamard matrix is necessary.

Definition 7. A *Hadamard matrix*⁴ of order n is an $n \times n$ matrix H with entries $+1$ and -1 , such that $HH^T = nI$.

The set of Hadamard matrices of order 2^n defined below are known as Sylvester matrices⁵. These matrices can be constructed recursively given

$$H_0 = [1], \quad H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

$H_{n+1} = H_1 \otimes H_n$, where the Kronecker product $A \otimes B$ is the matrix such that each entry $a_{i,j}$ of A is replaced by $a_{i,j}B$. In other words, $H_{n+1} = H_1 \otimes H_n = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$. These matrices are invertible and the inverse of a Sylvester matrix is $H_n^{-1} = 2^{-n}H_n$.

Example 7. The Hadamard matrix H_2 of order 2^2 .

⁴Named after the mathematician Jacques Hadamard, who introduced them in 1893.

⁵They were first constructed by the mathematician J. J. Sylvester in 1867.

$$H_2 = H_1 \otimes H_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

1.1.5 Hadamard Transform

This section will provide a very brief introduction to the Hadamard transform as well as Hadamard conjugation. The Hadamard transform is known by many names, but is usually referred to by some combination 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. The Hadamard transform \mathbf{H}_n is given by a $2^n \times 2^n$ Hadamard matrix. The matrix transforms a vector of 2^n real numbers x_1, \dots, x_{2^n} into a vector of real numbers with entries X_1, \dots, X_{2^n} . A very small example of the Hadamard transform is the 2-point Hadamard transformation.

Example 8. Let $\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$. Then $\mathbf{y} = H_1 \mathbf{x}$ implies that the entries of \mathbf{y} are the sum and differences of the two entries in \mathbf{x} , namely, $\mathbf{y} = \begin{bmatrix} x_0 + x_1 \\ x_0 - x_1 \end{bmatrix}$. This transform is the same as the 2-point discrete Fourier transform [26].

Larger examples of the Hadamard transform are possible using larger Hadamard matrices. For an $N = 2^n$ dimensional vector \mathbf{x} the transform of the vector is given by $\mathbf{y} = H_n \mathbf{x}$ and because the Hadamard matrix consists of entries of +1 and -1 the entries in the vector \mathbf{y} are found by adding and subtracting the components of \mathbf{x} .

The Hadamard transform has applications in a number of areas including signal processing, data compression algorithms, video compression and combinatorial phylogenetics. In combinatorial phylogenetics a technique known as Hadamard conjugation is used. Although there are several formulations of Hadamard conjugation, it is essentially the multiplication of rate and probability matrices by Hadamard matrices and their inverses. One benefit of using Hadamard conjugation is that it corrects for the number of substitutions which have actually taken place.

Since the succession of two or more substitutions between sequences will not be observed directly the number of substitutions will be underestimated without some correction. Hadamard conjugation provides this correction.

The problem of inferring a phylogenetic tree from observed sequences can be quite difficult for many nucleotide substitution models⁶. For some models however, Hadamard conjugation allows the derivation of invertible analytic formula from the observed sequences which can be used to determine the appropriate phylogenetic tree.

The invertible formula relating observed sequences to a vector which encodes the tree T known as Hadamard conjugation are:

$$\mathbf{p} = H^{-1} \exp(H\mathbf{q})$$

$$\mathbf{q} = H^{-1} \ln(H\mathbf{p}).$$

The vector \mathbf{p} gives the probabilities of each of the 4^n patterns of nucleotide differences at a site, the \mathbf{q} vector encodes T and model parameters on the edges of T , and H is a Hadamard matrix. Later sections of this paper will describe in more detail a few of the equivalent formulations of Hadamard conjugation.

1.1.6 Continuous-time Markov Processes

A Markov processes can be in either discrete or continuous time and in either discrete or continuous space. The continuous-time Markov processes considered here will be in continuous time and take values from a discrete space. It turns out that these types of Markov processes are useful in evolutionary modeling.

A continuous-time Markov process is a random process that satisfies the Markov property and takes values from a set called a state space. For the discussion here let Y be a random variable that takes on values in some discrete space, but whose values change in continuous

⁶Information on nucleotide substitution models can be found in section 2.1 on page 34

time. The Markov property is sometimes referred to as the memoryless property and is defined in the following way.

Definition 8. A continuous-time Markov process satisfies the **Markov property** if given $Y = i$ at time u , then the probability that $Y = j$ at time $u + t$ does not depend on the values of Y before time u .

In other words, the probability of future states occurring in the process depends only on the present state and not on past states.

Continuous-time Markov processes are most commonly defined by specifying a set of transition rates, q_{ij} , which give the rate at which state i will change to state j . The rate q_{ij} is contained in the $(i, j)^{th}$ entry of a *transition rate matrix* Q . In order to ensure that the probabilities of starting in a given state and either ending up in a different state or ending in the same state add up to one there are a few constraints on the Q matrix. 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 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 $P(t)$ is a matrix whose rows and columns are indexed by the states and whose $(i, j)^{th}$ entry is equal to $p_{ij} = Pr(Y_{n+1} = j | Y_n = i)$.

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

$$P(t) = e^{tQ}.$$

For details on this derivation see [56].

1.1.7 Matrix Exponential

When considering continuous time Markov processes it is necessary to understand the matrix exponential. This section provides a brief introduction to the matrix exponential as well as a few standard results. The definition and results can be found in [5].

Definition 9. [5] Let $A \in \mathbb{F}^{n \times n}$. Then the **matrix exponential** $e^A \in \mathbb{F}^{n \times n}$ or $\exp(A) \in \mathbb{F}^{n \times n}$ is the matrix

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

This series converges for all $A \in \mathbb{F}^{n \times n}$.

Proposition 2. [5] Let $A \in \mathbb{F}^{n \times n}$. Then the following statements hold:

1. e^A is nonsingular and $(e^A)^{-1} = e^{-A}$.
2. If $A = \text{diag}(A_1, \dots, A_k)$, where $A_i \in \mathbb{F}^{n_i \times n_i}$ for all $i = 1, \dots, k$, then $e^A = \text{diag}(e^{A_1}, \dots, e^{A_k})$.
3. If $S \in \mathbb{F}^{n \times n}$ is nonsingular, then $e^{SAS^{-1}} = Se^AS^{-1}$.

Proposition 3. [5] Let $A, B \in \mathbb{F}^{n \times n}$, then $AB = BA$ if and only if for all $t \in [0, \infty)$,

$$e^{tA} e^{tB} = e^{t(A+B)}.$$

Corollary 1. [5] Let $A, B \in \mathbb{F}^{n \times n}$, and assume that $AB = BA$. Then

$$e^A e^B = e^B e^A = e^{A+B}.$$

The next several sections focus on combinatorial objects that will be used in chapter 6.

1.1.8 Coherent Configurations and Association Schemes

This section contains a mathematical introduction to association schemes as well as an introduction to a more general combinatorial object known as a coherent configuration. Association schemes were first used by R. C. Bose in experimental design in statistics. Since then they have been used to study various topics in combinatorics.

Both association schemes and coherent configurations can be defined in terms of a set \mathfrak{X} and binary relations on that set \mathfrak{X} .

Definition 10. A **binary relation** on a set \mathfrak{X} is a subset of the cartesian product $\mathfrak{X} \times \mathfrak{X}$. A binary relation is symmetric if for $x, y \in \mathfrak{X}$, $(x, y) \in R_i$ implies that $(y, x) \in R_i$.

Coherent Configurations

There are several ways to define both coherent configurations and association schemes. This section will provide two such ways to define a coherent configuration while a later section will discuss association schemes.

Definition 11. [12] Let \mathfrak{X} be a finite set. A **coherent configuration** on \mathfrak{X} is a set $\mathcal{P} = \{R_1, R_2, \dots, R_s\}$ of binary relations on \mathfrak{X} satisfying the following four conditions:

1. \mathcal{P} is a partition of \mathfrak{X}^2 ;
2. there is a subset \mathcal{P}_0 of \mathcal{P} which is a partition of the diagonal $\Delta = \{(\alpha, \alpha) : \alpha \in \mathfrak{X}\}$;
3. for every relation $R_i \in \mathcal{P}$, its converse $R_i^T = \{(\beta, \alpha) : (\alpha, \beta) \in R_i\}$ is in \mathcal{P} ; let $R_i^T = R_{i^*}$.
4. there exists integers p_{ij}^k , for $1 \leq i, j, k \leq s$ such that for any $(\alpha, \beta) \in R_k$, the number of points $\gamma \in \mathfrak{X}$ such that $(\alpha, \gamma) \in R_i$ and $(\gamma, \beta) \in R_j$ is equal to p_{ij}^k , and, in particular, is independent of the choice of $(\alpha, \beta) \in R_k$.

The integers p_{ij}^k are also known as *intersection numbers* since if $R(\alpha) = \{\beta \in \mathfrak{X} : (\alpha, \beta) \in R\}$ then $p_{ij}^k = |\{R_i(\alpha) \cap R_j^T(\beta) : (\alpha, \beta) \in R_k\}|$.

In addition to describing coherent configurations in terms of binary relations on \mathfrak{X} it is also possible to describe coherent configurations in terms of a certain partition, or coloring of the edges of a complete graph. This is done by identifying the relations R_i to adjacency relations of a graph G_i on the vertex set \mathfrak{X} . For instance if $\alpha, \beta \in \mathfrak{X}$, then the relation $R_i = (\alpha, \beta)$ represents an edge from vertex α to vertex β in the graph G_i . Notice that since the relations are not required to be symmetric in a coherent configuration the edges in the graph G_i are directed. Using this interpretation the binary relations on \mathfrak{X} are represented by the adjacency matrices of the graphs G_i for $1 \leq i \leq s$. The rows and columns of the adjacency matrices, A_i , are indexed by \mathfrak{X} and the (α, β) entry is 1 if $(\alpha, \beta) \in R_i$ and 0 otherwise. This leads to the following definition of a coherent configuration.

Definition 12. [11] Let I be the identity matrix and J be the all ones matrix. The set of matrices $\{A_1, \dots, A_s\}$ is called a *coherent configuration* if

1. $\sum_{i=1}^s A_i = J$;
2. there is a subset of $\{A_1, A_2, \dots, A_s\}$ whose sum is the identity matrix I ;
3. for any i , there exists j such that $A_i^T = A_j$;
4. for any i, j , the product $A_i A_j$ is a linear combination of $\{A_1, \dots, A_s\}$.

In fact, for each pair i, j ,

$$A_i A_j = \sum_{k=1}^s p_{ij}^k A_k.$$

Notice that for $\alpha, \beta \in \mathfrak{X}$ the (α, β) entry of $A_i A_j$, denoted $(A_i A_j)_{(\alpha, \beta)}$, is equal to

$$\sum_{\gamma \in \mathfrak{X}} (A_i)_{(\alpha, \gamma)} (A_j)_{(\gamma, \beta)} = |\{\gamma : (\alpha, \gamma) \in R_i \text{ and } (\gamma, \beta) \in R_j\}| = p_{ij}^k.$$

As a result of the biological assumptions that will be made later on it will be of interest to consider the situation when the matrices $\{A_1, \dots, A_s\}$ are symmetric. A coherent configuration in which all matrices are symmetric is an association scheme.

Association Schemes

As with coherent configurations there are a number of ways to define an association scheme. The first definition given here defines an association scheme in terms of binary relations on \mathfrak{X} .

Definition 13. [61] An *association scheme* on a set \mathfrak{X} of points consists of $r + 1$ nonempty symmetric binary relations R_0, R_1, \dots, R_r on \mathfrak{X} which partition $\mathfrak{X} \times \mathfrak{X}$, where $R_0 = \{(x, x) : x \in \mathfrak{X}\}$ is the identity relation, and such that for some nonnegative integers p_{ij}^k , $0 \leq i, j, k \leq r$ the following holds:

1. given any $(x, y) \in R_k$, there are exactly p_{ij}^k elements $z \in \mathfrak{X}$ such that $(x, z) \in R_i$ and $(z, y) \in R_j$.

If $(x, y) \in R_i$, then x and y are i^{th} associates and the numbers p_{ij}^k are called the parameters of the scheme.

As with coherent configurations association schemes can be thought of in terms of an edge coloring of a complete graph with vertex set \mathfrak{X} . Since the relations in an association scheme are symmetric, the complete graph being considered is undirected. The following two definitions are given in terms of this graph coloration.

Definition 14. [3] *An association scheme with s associate classes on a finite set \mathfrak{X} is a coloring of the edges of the complete undirected graph with vertex set \mathfrak{X} by s colors such that*

1. *for all i, j, k in $\{1, \dots, s\}$ there is an integer p_{ij}^k such that, whenever $\{\alpha, \beta\}$ is an edge of color k then*

$$|\{\gamma \in \mathfrak{X} : \{\alpha, \gamma\} \text{ has color } i \text{ and } \{\gamma, \beta\} \text{ has color } j\}| = p_{ij}^k;$$

2. *every color is used at least once;*
3. *there are integers a_i for i in $\{1, \dots, s\}$ such that each vertex is contained in exactly a_i edges of color i .*

Condition 1 states that if vertices α and β are fixed with the edge between them colored by k , then the number of triangles which consist of edge $\{\alpha, \beta\}$ with an i colored edge through α and a j colored edge through β is exactly p_{ij}^k .

The definition above is equivalent to following definition which is given in terms of adjacency relations of a graph G_i on the vertex set \mathfrak{X} .

Definition 15. [11] *A set $\{A_0, \dots, A_s\}$ of zero-one matrices is an association scheme if the following conditions hold:*

1. $\sum_{i=0}^s A_i = J$, the all 1 matrix;
2. $A_0 = I$;

3. for any i , $A_i^T = A_i$;

4. for any i, j the product $A_i A_j$ is a linear combination of A_0, \dots, A_s .

As with coherent configurations for each pair i, j ,

$$A_i A_j = \sum_{k=1}^s p_{ij}^k A_k.$$

It is straight forward to show that a set of zero-one matrices commute if they satisfy the conditions to be an association scheme.

Lemma 1. [3] If A_0, A_1, \dots, A_s are the adjacency matrices of an association scheme then $A_i A_j = A_j A_i$ for all i, j in $\{0, 1, \dots, s\}$.

Proof.

$$\begin{aligned} A_j A_i &= A_j^T A_i^T \\ &= (A_i A_j)^T \\ &= \left(\sum_k p_{ij}^k A_k \right)^T \\ &= \sum_k p_{ij}^k A_k^T \\ &= \sum_k p_{ij}^k A_k \\ &= A_i A_j \end{aligned}$$

□

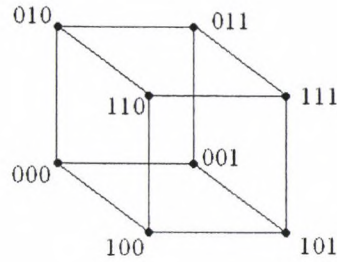
The following is an example of an association scheme. Chapter 6 will discuss the connection between association schemes and evolutionary models.

Example 9. Let Γ be an alphabet of size n and \mathfrak{X} be the set of words of length m over the alphabet Γ . Two m -tuples x, y are i^{th} associates if they disagree in exactly i coordinates, where $0 \leq i \leq m$. Notice that the parameters p_{ij}^k exist by symmetry. This is an association scheme known as the Hamming scheme and is denoted by $H(m, n)$.

The Hamming scheme $H(3,2)$ is an association scheme on the set of points

$$\mathfrak{X} = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}.$$

There are three associate classes which can be visualized by labeling the eight vertices of the cube with the eight points above in such a way that each edge of the cube connects first associates.



Next color the edges of the cube red, the main diagonals orange, and the face diagonals yellow. Each vertex is incident to three red edges, one orange edge, and three yellow edges. If 0^{th} associates are represented by green, then p_{ij}^k , for $k = \{\text{red, orange, yellow}\}$ are given by the following matrices where the rows and columns are indexed by $\{\text{green, red, orange, yellow}\}$.

$$p_{ij}^{\text{red}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix}, \quad p_{ij}^{\text{orange}} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \\ 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{bmatrix}, \quad p_{ij}^{\text{yellow}} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 \end{bmatrix}$$

For further information on association schemes see the texts by A.E. Brouwer, A.M. Cohen and A. Neumaier [7] and E. Bannai, and T. Ito [4].

1.1.9 Strongly regular and distance regular graphs

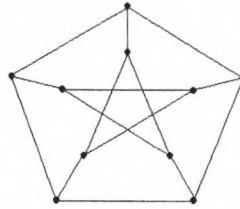
Strongly regular graphs are a type of graph which correspond to 2-class association schemes. A strongly regular graph is a *regular graph*, a graph where each vertex has the same degree, with additional structure.

Definition 16. [61] A **strongly regular graph** $srg(v, k, \lambda, \mu)$ is a graph with v vertices that is regular of degree k and that has the following properties:

1. For any two adjacent vertices x, y , there are exactly λ vertices adjacent to x and to y .
2. For any two nonadjacent vertices x, y , there are exactly μ vertices adjacent to x and to y .

A well known example of a strongly regular graph is the Peterson graph.

Example 10. The Peterson graph is an $srg(10, 3, 0, 1)$.



Letting two distinct vertices in a strongly regular graph be first associates if they are adjacent and second associates if they are not adjacent produces a 2-class association scheme.

A connected strongly regular graph can also be thought of as a distance regular graph with diameter two, where the *diameter* of a connected graph is the maximum distance between any pair of vertices.

Definition 17. [3] For any connected graph G let

$$G_i = \{(x, y) \in \Omega \times \Omega : \text{the distance between } x \text{ and } y \text{ is } i\}.$$

Then if G is a connected graph with diameter s and vertex set Ω , G is **distance regular** if G_0, G_1, \dots, G_s form an association scheme on Ω .

1.1.10 Johnson Scheme

The Johnson scheme denoted by $J(n, m)$ is an association scheme that is defined using the $\binom{n}{m}$ subsets of an n -set. Let Ω consist of all m -subsets of an n -set Γ where $1 \leq m \leq \frac{n}{2}$. For $i = 0, 1, \dots, m$ let α and β be i^{th} associates if $|\alpha \cap \beta| = m - i$. Thus 0^{th} associates are equal.

In [3] R. A. Bailey provides a proof that the Johnson scheme is in fact an association scheme by showing its correspondence to a distance regular graph. The following is an example of the Johnson Scheme $J(4,2)$.

Example 11. $J(4,2)$ yields an association scheme on $\binom{4}{2} = 6$ points.

$$\Gamma = \{1,2,3,4\}$$

$$\Omega = \{\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$$

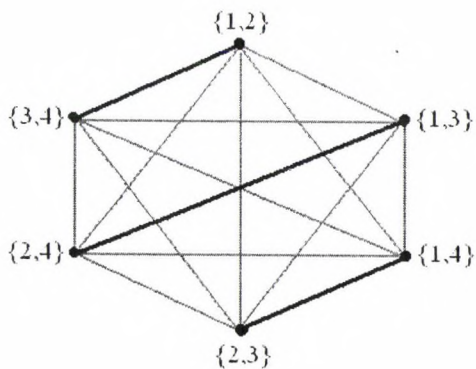
Two subsets are first associates if $|\alpha \cap \beta| = 2 - 1 = 1$. Therefore the list of first associates is

$$\begin{aligned} C_1 = & \{(\{1,2\}, \{1,3\}), (\{1,2\}, \{1,4\}), (\{1,2\}, \{2,3\}), (\{1,2\}, \{2,4\}), \\ & (\{1,3\}, \{1,4\}), (\{1,3\}, \{2,3\}), (\{1,3\}, \{3,4\}), (\{1,4\}, \{2,4\}), \\ & (\{1,4\}, \{3,4\}), (\{2,3\}, \{2,4\}), (\{2,3\}, \{3,4\}), (\{2,4\}, \{3,4\})\} \end{aligned}$$

and the list of second associates is made up of the subsets such that $|\alpha \cap \beta| = 2 - 2 = 0$.

$$C_2 = \{(\{1,2\}, \{3,4\}), (\{1,3\}, \{2,4\}), (\{1,4\}, \{2,3\})\}$$

The coloration of the complete graph that corresponds to the association scheme $J(4,2)$ is given below.



Example 12. The Johnson scheme $J(6,3)$ is an association scheme on twenty points with three associate classes.

1.1.11 Bose-Mesner Algebra

Given an association scheme on a set \mathfrak{X} with s associates and adjacency matrices A_0, A_1, \dots, A_s there is a corresponding associative commutative algebra, \mathcal{A} , of the association scheme. Notice that the matrices A_0, A_1, \dots, A_s must be linearly independent since $A_0 + A_1 + \dots + A_s = J$, the all ones matrix. Therefore the set $\mathcal{A} = \{\sum_{i=0}^s \mu_i A_i : \mu_0, \dots, \mu_s \in \mathbb{R}\}$ has dimension $s + 1$ as a vector space over \mathbb{R} . \mathcal{A} is also closed under multiplication since

$$A_i A_j = \sum_{k=0}^s p_{ij}^k A_k.$$

\mathcal{A} is a commutative and associative algebra known as the *Bose-Mesner algebra*. Since the matrices in \mathcal{A} are symmetric and commute they can be simultaneously diagonalized.

In addition to being closed under matrix multiplication the axioms for association schemes imply that the set \mathcal{A} is also closed under Schur multiplication.

Definition 18. *The Schur product also known as the Hadamard product of the $n \times m$ matrices A and B is an $n \times m$ matrix C such that $C_{ij} = A_{ij}B_{ij}$.*

Notice that the A_i are idempotents with respect to the Schur product and these A_i are known as the Schur idempotents of the scheme. In addition to association schemes producing associative commutative algebras it is also the case that a finite dimensional vector space of real symmetric matrices containing I and J , that is closed under both Schur multiplication and matrix multiplication has a unique basis formed of Schur idempotents and the matrices in this basis form an association scheme [23].

1.1.12 Results from Linear Algebra

The following standard results and definitions from linear algebra will be used later in the paper.

Definition 19. *Let $A \in \mathbb{C}^{n \times m}$. The complex conjugate transpose of A is $A^* = \bar{A}^T$.*

Definition 20. *For $A \in \mathbb{F}^{n \times n}$ define the following types of matrices:*

1. A is **unitary** if $A^*A = I$.
2. A is **orthogonal** if $A^T A = I$.
3. A is **normal** if $AA^* = A^*A$.

A set of matrices are said to be *simultaneously diagonalizable* if there exists a single invertible matrix P such that $P^{-1}AP$ is a diagonal matrix for every A in the set. The following theorems have been included since the ability to simultaneously diagonalize sets of matrices will be important later on.

Theorem 4. [5] Let $A \in \mathbb{F}^{n \times n}$. Then A is diagonalizable by a unitary matrix if and only if A is normal.

Corollary 2. Let A be an $n \times n$ real symmetric matrix, then A is diagonalizable by an orthogonal matrix.

Theorem 5. [34] Let A and B be $n \times n$ diagonalizable complex matrices. Then A and B commute if and only if they are simultaneously diagonalizable.

1.2 Biological Introduction

The following sections provide a brief introduction to some of the biological concepts related to the construction of phylogenetic trees which are used to display the evolutionary relationships between taxa.

1.2.1 Evolution

In an attempt to understand the world around us scientists have been interested in classifying and determining relationships between different species for hundreds of years. Maritime trade caused individuals around the world to become familiar with the wide range of organisms that inhabited the planet and in 1735 a Swedish botanist named Carl von Linné published the first universally accepted system of taxonomic ranks in *Systema naturae*. Since it was first published

the taxonomic ranks have been expanded upon and now include eight major taxonomic ranks as well as a number of minor rankings where the most general rank is life and the most specific is species.⁷

Unfortunately it is difficult to determine specific definitions for the various taxonomic rankings. For instance the definition of species is often thought to be a group of organisms that breeds internally, but is reproductively isolated from other such groups. Although this definition works well for most multi-celled organisms it does not deal with asexually reproducing single-celled organisms or the small number of parthenogenetic multi-celled organisms. Additionally it is often difficult to tell whether similar groups of organisms may be capable of interbreeding.

Despite the difficulty in defining taxonomic ranking and determining what organisms belong in a given taxon, there are benefits to determining historical relationships between groups of organisms. Determining historical relationships help evolutionary biologists understand how species evolved and evolutionary analysis of how a virus, such as the influenza virus, evolves helps immunologists develop new vaccines. Phylogenetic analysis has even been used to gain insight into the Human Immunodeficiency Virus (HIV) [50].

In order to understand the historical relationships between groups of organisms it is necessary to understand how organisms evolve. Evolution is the change in the inherited traits of a population from one generation to the next. These traits are expressions of genes and are passed on to future generations through reproduction. Mutations in genes can cause differences between organisms and these difference become more common or rare in a population as evolution takes place. As populations change they may speciate into different species, hybridize together again, or terminate by extinction.

Once new traits appear in a population there are three basic mechanisms that produce evolutionary change. Perhaps the most widely known mechanism is natural selection; an idea Charles Darwin developed during his voyage on the HMS *Beagle* from 1831 to 1836. During his time

⁷For a list of the major taxonomic ranks see page 118.

stopped on the Galapagos Islands he observed the distinct species of tortoise, mocking-thrush, finches and plants native to each of the islands. For twenty years following his voyage Darwin constructed the theory of natural selection. In 1856 he began to record his ideas, but before his work was completed he received a manuscript from Alfred Russel Wallace entitled *On the tendency of varieties to depart indefinitely from the original type*, which contained similar ideas to Darwin's theory of natural selection. Wallace's essay was presented to the Linnean Society of London on July 1, 1858, along with notes from Darwin. In November of 1859 Darwin published his theory in the text *On the origin of species*.

Natural selection is the process by which inherited traits that make it more likely for an organism to survive and reproduce become common in a population over time. This occurs because individuals of a given species have genetic variation. Since unlimited population growth of a species is not sustainable, some individuals will not survive to reproduce. If an individual has inherited a trait that makes it more likely for them to survive, they have the possibility of passing on this trait to their offspring. Individuals who are less fit are less likely to survive and therefore will not pass on their traits to future generations. Natural selection acts without regard to the future. As a result it is possible for changes to occur in a population that are initially beneficial, but over time can become less advantageous as the environment changes.

The second mechanism that produces evolutionary change is genetic drift. Unlike natural selection which is driven by environmental or adaptive pressures, genetic drift is a change in the relative frequency of traits in a population due to random sampling and chance. Genetic drift occurs because traits in offspring are a random sample of the traits in the parent generation. This process is most significant in small populations and less noticeable in larger populations.

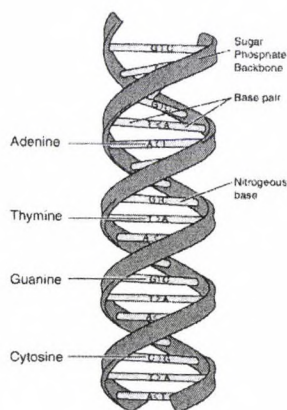
A third mechanism is gene flow. Gene flow is the exchange in genes between populations and often occurs as the result of migration. Consequently, greater mobility of a population results in a higher potential for gene transfer between populations. If gene flow is maintained between two populations the genetic variation between the two groups is greatly reduced making speciation less likely to occur.

In addition to the above mechanisms of evolutionary change, horizontal gene transfer is possible in bacteria. Horizontal gene transfer or lateral gene transfer is any process by which an organism transfers genetic material to another cell that is not its offspring. Although there is some evidence horizontal gene transfer exists in higher plants and animals its prevalence and importance is still being debated.

Since genetic data is only available for presently extant species phylogenetics attempts to use knowledge of evolution to reconstruct the evolutionary relationships between taxa. The relationships between taxa are displayed using a phylogenetic tree or phylogenetic network depending on what types of evolutionary events have occurred.

1.2.2 DNA and RNA

Deoxyribonucleic acid, also known as DNA, is a nucleic acid which contains genetic instructions used in the development and functioning of living organisms and some viruses. DNA has a double helix structure composed of one long strand of nucleic acid wrapped around another. The backbone of the strand is a string of alternating sugar molecules and phosphate groups. Attached to each sugar molecule is one of four bases: cytosine (C), guanine (G), adenine (A), and thymine (T). The bases A and G are purine bases while the bases C and T are pyrimidines. Each base on one strand pairs with exactly one base on the other strand. More specifically A bonds with T and C bonds with G. This is known as complementary base pairing and can be seen in the picture below.



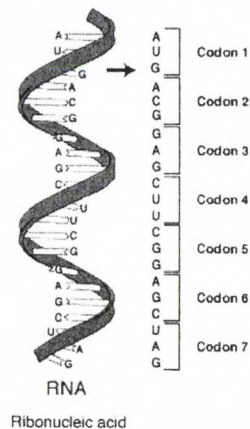
The evolutionary models described in the next section are nucleotide substitution models.

Definition 21. A *nucleotide* is composed of a nucleobase (cytosine, guanine, adenine, or thymine in the case of DNA), a five-carbon sugar (2'-deoxyribose), and one to three phosphate groups.

In the evolutionary models below the bases A, C, G, and T are used to abbreviate the four nucleotides and a strand of DNA is thought of as a string of A, C, G, and T's.

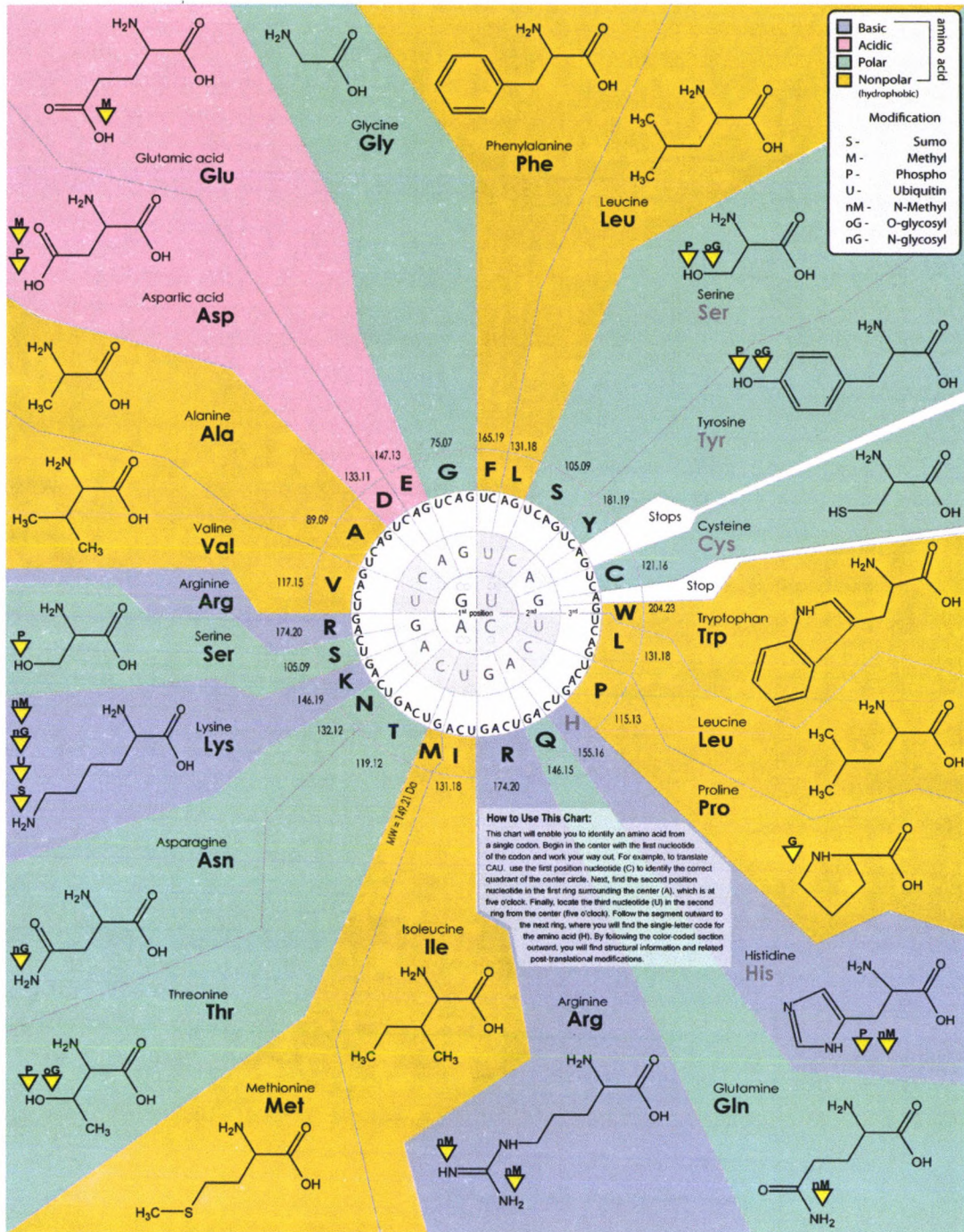
Ribonucleic acid, RNA, has a similar structure as DNA, however it is single stranded and the sugar-phosphate backbone contains ribose rather than deoxyribose. Also different are the four bases attached to the sugar-phosphate backbone. Instead of A, C, G, and T, the bases A, C, G, and uracil (U) are attached.

There are two main types of RNA. The first is messenger RNA also known as mRNA. mRNA is transcribed from DNA and carries coding information to the sites of protein synthesis. The second type of RNA is transfer RNA or tRNA. This is the nucleic acid that constructs the protein. Proteins are linear chains of amino acids and each of the amino acids is encoded by a triple of nucleotides. Although it was originally assumed that there are twenty amino acids, recent work has shown that, "there is also a natural expansion of the genetic code beyond the twenty-amino-acid repertoire that leads to the coding of modified amino acids [1]." Many scientists now believe there are twenty-two amino acids. The ordered triple of nucleotides that encode the amino acids form a codon. Since there are four nucleotides there are $4^3 = 64$ different codons.



Many scientists believe that with three exceptions each codon encodes for one of twenty amino acids used in the synthesis of proteins while the remaining three codons are stop codons that send a signal to stop building the protein. This belief has also come into question in recent years. Ambrogelly, Palioura, and Söll state in a 2007 paper entitled *Natural expansion of the genetic code* that in regard to the (stop) codons UAG, UAA, UGA, "... single occurrence per gene renders them particularly good candidates for reassignment given that such a change would cause minimal damage to the proteome[1]." They then go on to say that UGA, apart from functioning as a stop codon, also encodes selenocysteine, the twenty-first amino acid and that UAG encodes pyrrolysine, the twenty-second amino acid. The codon AUG corresponds to an amino acid and is a start codon; the first AUG in an mRNA's coding region is where translation into protein begins.

The following chart shows the correspondence between the sixty-four codons and the amino acids.



Chapter 2

MODELS OF EVOLUTION

Over the last several years there has been an increased emphasis on methods of phylogenetic inference that are based on models of evolutionary change [58]. The following chapter discusses types of data used along with some well know models of evolutionary change. Although it is not possible to provide an exhaustive review of the models which exist the chapter provides a brief description of some of the most commonly used models.

The models discussed here use unordered multistate character data, where the characters takes on two or more discrete values. The fact that the character data is unordered implies that a character at state i can transform into any other state j . To construct a phylogenetic tree sequences of character data are obtained from each taxon being considered. For example, in the case of a nucleotide substitution model each character in the sequence can take on one of four states, A , C , G , or T resulting in finite sequence made up of the nucleotides A , C , G , and T . When choosing sequence data for each taxon it is necessary that the sequences are aligned and that the states observed at a giving position in the taxa being considered should all trace their ancestry to a single position that occurred in a common ancestor of those taxa.

In addition to four state nucleotide sequence data there are also models of evolution that use other types of character data. One of the simplest evolutionary model is the Neyman model [48] which uses character sequences composed of two states, purines denoted by R and pyrimidines denoted Y . This model assumes symmetric substitutions so that the probabilities of a substitution from state R to Y and from state Y to R have the same value. There also exists evolutionary

models with a larger numbers of states. For instance, codon models are models with anywhere between sixty-one and sixty-four states. These models will be examined more closely in section 2.2.

Since nucleotide substitution models occupy a large portion of the literature regarding Hadamard conjugation the following section provides an introduction to nucleotide substitution models in general as well as a closer look at some specific named nucleotide models.

2.1 An introduction to Nucleotide Substitution Models

A substitution model describes the process by which a sequence of nucleotides of fixed size changes into another nucleotide sequence. Although evolutionary processes consist of mutations other than substitutions, most models of evolution make the simplifying assumption that only substitutions take place and that the sequences being compared are properly aligned. This is just one of a few common simplifying assumptions.

2.1.1 Assumptions of many nucleotide substitution models

There are a few assumptions made about the substitutions taking place at sites in the nucleotide sequence. The first of which is that changes in a single site of a nucleotide sequence do not affect the probability of changes taking place in another site. This assumption is known as independence. A second assumption results from the fact that the models discussed in this section are Markov models. This requires that a change at a single site from state i to j does not depend of the history of the site prior to state i and therefore knowledge of previous states is irrelevant. Finally, it is assumed that each site in the sequence may be changed multiple times.

In addition to the assumptions made about the substitutions taking place at different sites in the sequence, the majority of substitution models assume neutral evolution. Neutral evolution is the assumption that the molecular changes represented by substitutions do not influence the fitness of the individual organism. In other words, selection does not operate on the substitutions. The theory of neutral evolution was first formalized by Motoo Kimura in 1968 and is

now widely accepted, however the debate over relative percentages of neutral and non-neutral selection remains.

There are two main ways in which substitution models measure time. The first, which is used in nucleotide substitution models, is to measure time by keeping track of the number of substitutions which have taken place. One benefit of this method is that it avoids the issue of whether the rate of substitutions per unit of time has remained constant or not. The second way of dealing with time is to use the molecular clock assumption. This is the assumption that the rate of substitutions with respect to time remains constant. For this assumption to be used in a model a substitution rate must be known. Determining this rate is often very difficult and many scientists argue that the molecular clock assumption is unrealistic.

The above assumptions are common to the nucleotide substitution models discussed in this paper, however additional assumptions regarding what type of nucleotide substitutions take place can also be made. These additional assumptions define a specific nucleotide substitution model.

2.1.2 General nucleotide substitution model

Each nucleotide substitution model can be specified by a table of rates at which nucleotides are replaced by other nucleotides. The rates are contained in a four by four *rate matrix* \mathbf{Q} whose rows and columns are indexed by the nucleotides A, C, G and T, respectively. The entry Q_{ij} contains the rate of change of nucleotide i to nucleotide j during an infinitesimal time period. The most general nucleotide substitution model possible is given by a \mathbf{Q} matrix whose entries are made up of a mean instantaneous substitution rate along with a frequency parameter that describes the frequency of the given nucleotide. The matrix of the general nucleotide substitution model is given below.

$$\mathbf{Q} = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_C + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_C + j\pi_G + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_C + k\pi_G + l\pi_T) \end{bmatrix}$$

In the matrix above μz , for $z \in \{a, b, c, \dots, l\}$, is the mean instantaneous substitution rate for substitution z . So for example, given the indexing of the rows and columns, μa is the mean instantaneous substitution rate for the substitution from A to C. π_i for $i \in \{A, C, G, T\}$ are the frequency parameters providing the frequency of each of the four nucleotides. It is assumed that these frequencies remain constant over time. It is also the case that the diagonal entries of \mathbf{Q} are chosen such that the row sums are zero. If each μz and π_i are chosen to be distinct values the \mathbf{Q} matrix produced would represent the most general substitution model possible. In practice however, additional assumptions are made.

2.1.3 General time-reversible model

With the exception of strand symmetric models, all evolutionary models discussed in this paper will be assumed to be general time-reversible models. Most nucleotide substitution models used are time-reversible, which means the rate of change from nucleotide i to nucleotide j is the same as the rate of change from j to i .

Definition 22. *Nucleotide substitution models that assume the overall rate of change from base i to base j in a given length of time is the same as the rate of change from base j to base i are said to be **time-reversible**.*

This assumption is equivalent to setting $a = g$, $b = h$, $c = i$, $d = j$, $e = k$, and $f = l$. Setting these parameters equal produces the most general *time-reversible* model defined by the matrix below.

$$\mathbf{Q} = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_C + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_C + d\pi_G + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_C + e\pi_G + f\pi_T) \end{bmatrix}$$

A desirable result of choosing to use a time-reversible model, sometimes referred to as an REV model, is that the placement of the root need not be considered when choosing a tree that most closely fits the data. Although the general time-reversible model makes the simplifying

assumption that many of the mean instantaneous substitution rates are equal Ziheng Yang concluded in his 1994 paper “Estimating the pattern of nucleotide substitution” [64] that “...the use of the REV model in phylogenetic analysis can be recommended, especially for large data sets or for sequences with extreme substitution patterns...”. He also states that “the use of the unrestricted models does not appear to be worthwhile”.

Despite the fact that only a few general time-reversible models have been named there exist 203 different time-reversible models. Each time-reversible model can have anywhere between one and six substitution types. Therefore determining the number of time-reversible models is equivalent to finding the number of ways a set with six objects can be partitioned into disjoint and non-empty subsets. This is equal to the Bell number $B_6 = 203$.

The nucleotide substitution models mentioned in the following sections are specific examples of the 203 time-reversible models. They are produced by making assumptions about the rates of the different substitutions taking place, by making assumptions about the frequencies of the bases A, C, G and T, or by making assumptions about both the rates and the frequencies. A few of these models will be mentioned below.

The final observation regarding time-reversible models included in this section is given below.

Observation 1. *The rate matrices of time-reversible nucleotide substitution models are real symmetric matrices.*

Proof. It follows from definition of a time-reversible rate matrix. □

Real symmetric matrices have some nice properties that will be used in later sections to analyze time-reversible models.

2.1.4 Tamura and Nei Model

Tamura and Nei introduced the Tamura and Nei model which is also known as the TrN model in their 1993 paper “Estimation of the number of nucleotide substitutions in the control

region of mitochondrial DNA in humans and chimpanzees” [39]. The TrN model is a general time-reversible model that assumes there are three substitution types. It also assumes that the base frequencies are not necessarily equal. The nucleotides can be divided into two chemical classes; purines A and G and pyrimidines T and C. The substitutions taking place within the two chemical classes are known as transitions and substitutions between the classes are called transversions. The three substitution types in TrN are α_R , the probability that a purine will be replaced with a purine, α_Y , the probability a pyrimidine will be replaced with a pyrimidine, and β the probability that a transversion will occur.

To determine the rates, consider the following. As before let π_A , π_C , π_G , and π_T be the frequencies of A, C, G and T respectively. Only considering the purines, the ratio of purine A to purine G is $\pi_A : \pi_G$. The total frequency of the purines is $\pi_R = \pi_A + \pi_G$ so the frequencies of A and G considering purines are π_A/π_R and π_G/π_R . A similar computation can be done for the pyrimidines. If π_Y is the total frequency of pyrimidines, then π_C/π_Y and π_T/π_Y are the frequencies of C and T amongst the pyrimidines.

Given these probabilities and frequencies the rate matrix is given below. The entries on the diagonal are chosen such that the row sums are zero, however the terms have been omitted to make the matrix easier to read.

$$\mathbf{Q} = \begin{bmatrix} - & \beta\pi_C & \alpha_R\pi_G/\pi_R + \beta\pi_G & \beta\pi_T \\ \beta\pi_A & - & \beta\pi_G & \alpha_Y\pi_T/\pi_Y + \beta\pi_T \\ \alpha_R\pi_A/\pi_R + \beta\pi_A & \beta\pi_C & - & \beta\pi_T \\ \beta\pi_A & \alpha_Y\pi_C/\pi_Y + \beta\pi_C & \beta\pi_G & - \end{bmatrix}$$

The next model discussed is a specific case of the Tamura and Nei model.

2.1.5 Hasegawa-Kishino-Yano Model

The Hasegawa-Kishino-Yano model, also known as the HKY model, was introduced in the 1985 paper, “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”[27]. Although it was published after the Tamura and Nei model it is a specific case of the TrN model where $\alpha_R/\alpha_Y = \pi_R/\pi_Y$.

2.1.6 Kimura's Three-Substitution type Model

The Kimura three-substitution type model was introduced in Motoo Kimura's 1981 paper, "Estimation of evolutionary distances between homologous nucleotide sequences"[44] and is different from the TrN and HKY models above in that it requires the base frequencies of all four nucleotides to be equal. Like the HKY model however, it assumes there are three substitution types. The three substitution types are transitions, μ_α , and two types of transversions, μ_β and μ_γ .

$$\begin{aligned}\mu_\alpha: & A \leftrightarrow G, \quad T \leftrightarrow C \\ \mu_\beta: & A \leftrightarrow T, \quad G \leftrightarrow C \\ \mu_\gamma: & A \leftrightarrow C, \quad T \leftrightarrow G\end{aligned}$$

Using the notation from the rate matrices given above the \mathbf{Q} matrix for the three-substitution type model is

$$\mathbf{Q} = \begin{bmatrix} -.75(\mu_\alpha + \mu_\beta + \mu_\gamma) & .25\mu_\gamma & .25\mu_\alpha & .25\mu_\beta \\ .25\mu_\gamma & -.75(\mu_\alpha + \mu_\beta + \mu_\gamma) & .25\mu_\beta & .25\mu_\alpha \\ .25\mu_\alpha & .25\mu_\beta & -.75(\mu_\alpha + \mu_\beta + \mu_\gamma) & .25\mu_\gamma \\ .25\mu_\beta & .25\mu_\alpha & .25\mu_\gamma & -.75(\mu_\alpha + \mu_\beta + \mu_\gamma) \end{bmatrix}$$

It is common to combine the base frequency and substitution rate into a single term, therefore the rate matrix for the Kimura three-substitution type model is typically given as

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

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

The Kimura three-substitution type model will be discussed again later in the paper.

2.1.7 Kimura's Two-parameter Model

The Kimura two-parameter model was introduced in the 1980 paper "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences" [43] and is a specific case of the three-substitution type model. The model also known as the K2ST model again assumes equal base frequencies, but only allows two substitution types. The model has one substitution rate for transitions and one rate for transversions. Therefore the K2ST model can be obtained from the K3ST model by setting $\beta = \gamma$. The rate matrix is given by

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

where $K = \alpha + 2\beta$.

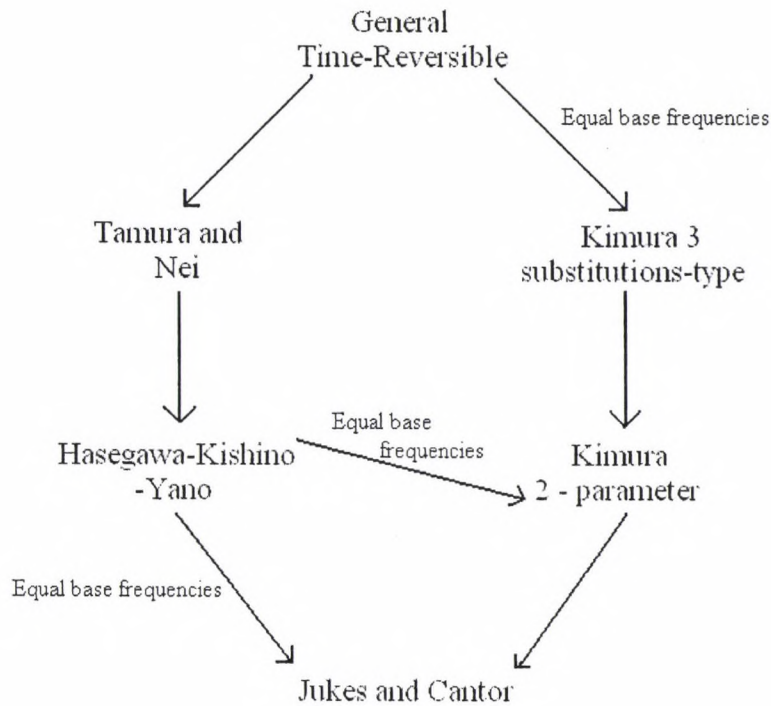
2.1.8 Jukes and Cantor Model

T.H. Jukes and C.R. Cantor's paper, "Evolution of protein molecules"[38], was published in 1969. The paper describes the Jukes and Cantor model which is the most basic nucleotide substitution model. The model assumes that all base frequencies are equal and that all substitutions occur at the same rate. Combining the base frequency and substitution rate into one term yields a rate matrix

$$\mathbf{Q} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}.$$

Several examples in this paper will assume the Jukes and Cantor model because of its simplicity.

The following diagram, modified from a diagram in Hillis [58], helps to demonstrate the relationships between the models that have been discussed.



2.1.9 Strand Symmetric Models

Strand symmetric models are a class of models which encompass three of the models mentioned above, the Jukes and Cantor model, the Kimura two-parameter model and the Kimura three-substitution type model. The idea of a strand symmetric model was first introduced in a 1995 paper, “Intrastrand Parity Rules of DNA Base Composition and Usage Biases of Synonymous Codons” by Noboru Sueoka [57] and resulted from recognizing the base-pairing rule of the Watson and Crick DNA structure and trying to incorporate that into an evolutionary model.

Strand symmetric substitution implies that both strands of the genome segment undergo any given type of substitution at the same rate, hence complementary substitution rates are equal. For example, the substitution rate, r_1 , for $A \rightarrow T$ on strand one is equivalent to the substitution rate,

s_2 , for $T \rightarrow A$ on strand two. Similarly the substitution rate, r_2 , for $T \rightarrow A$ on strand one is equivalent to the substitution rate, s_1 , for $A \rightarrow T$ on strand two and thus $r_1 = s_2$ and $r_2 = s_1$. Furthermore since $A \rightarrow T$ at the same rate in both strands, $r_1 = r_2$ and $s_1 = s_2$ under strand symmetric substitution.

The above equalities illustrate that when there are no biases in mutation and selection between the two strands of DNA, in the case of a strand symmetric model, the twelve possible substitution rates of the nucleotides reduce to six. The six rates correspond to the six substitution types

$$\begin{aligned}
 a : A &\leftrightarrow T \\
 b : A &\rightarrow G, & T &\rightarrow C \\
 c : G &\rightarrow A, & C &\rightarrow T \\
 d : G &\rightarrow T, & C &\rightarrow A \\
 e : A &\rightarrow C, & T &\rightarrow G \\
 f : G &\leftrightarrow C.
 \end{aligned}$$

Additionally strand symmetric models must have base frequencies that satisfy

$$\pi_A = \pi_T \quad \pi_C = \pi_G.$$

The models mentioned in this section are only a sampling of the existing nucleotide substitution models. Although all the models mentioned require assumptions to be made, they still help to provide insight into the evolutionary history of different taxa. Undoubtedly as more work is done in the area the models considered will continue to improve.

2.2 Codon Models of Evolution

Creating more realistic evolutionary models has been an important area of research in the area of phylogenetic inference. In 1994 two papers, one by Nick Goldman and Ziheng Yang [24] and another by Spencer Muse and Brandon Gaut [47], started discussion on the use of codon models of evolution.

Yang and Goldman describe in their paper how they "...devised a model of nucleotide substitution that uses simultaneously the nucleotide-level information in DNA sequences and knowledge of the genetic code and hence the amino acid-level information of the synonymous (silent) and nonsynonymous (replacement) nucleotide substitutions." [24] They did this by modeling at the codon level instead of modeling at the nucleotide or amino acid levels. Although they consider codon models of evolution much of the discussion contained in this paper regarding nucleotide substitution models applies to the codon models of evolution.

Both Yang and Goldman and Muse and Gaut consider Markov-Process Models of codon substitution. They assume that the data they are analyzing is a DNA sequence with no gaps. They also assume insertions and deletions cannot occur. Recall that a codon is a triple of nucleotides and that while there are $4^3 = 64$ different codons, Yang and Goldman assume that only sixty-one correspond to amino acids. The remaining three codons are stop codons and are often not considered. As a result the sixty-one sense codons make up the states of the continuous-time Markov process. A 61×61 rate matrix, Q , is indexed by the sixty-one sense codons so that Q_{ij} for $i \neq j$ contains the rate at which codon i changes to codon j . Q_{ii} is chosen so that row sums of the rate matrix equal zero. The equation

$$P(t) = \exp(Qt)$$

relates the Q matrix to the matrix P where P_{ij} gives the probability that codon i is replaced by codon j after time t .

It is also assumed that mutations occur in the three codon positions independently and that a mutation corresponds to a single nucleotide substitution. As with nucleotide substitution models the rate at which each substitution takes place in a codon model is proportional to the equilibrium frequency π_j of the codon j being changed to.

Given the fact that the number of codons coding for amino acids may actually be larger than sixty-one it may also be of interest to consider codon models of evolution with sixty-two, sixty-three or sixty-four states. Support for this is given in the 2007 paper [1] by Ambrogelly, Palioura, and Söll.

Chapter 3

RELATING DNA SEQUENCE DATA TO PHYLOGENETIC TREES

Hadamard conjugation allows invertible calculations between predictive data from DNA sequences and the phylogenetic tree for a given a nucleotide substitution model. The following chapter explains how information from DNA sequences can be organized into a vector known as the observed sequence spectrum. Hadamard conjugation can then applied to the observed sequence spectrum to obtain the conjugate spectrum which corresponds to a vector containing information about the corresponding tree. The vector associated to the tree will be discussed later in the paper, for now the discussion will focus on the observed sequence spectrum.

3.1 The observed sequence spectrum vector $f(D)$

The observed sequence spectrum vector is created by analyzing a set of aligned DNA sequences from n different taxa. The i^{th} component of the vector contains the expected relative frequency of the differences between the n taxa at a given site. In order to understand how this information is included in the vector some background on bipartitions and quadripartitions is necessary.

3.1.1 Two state substitution models and bipartitions

Let us first look at the 2-state model. The 2-state model allows for two substitutions. A purine (denoted R) can change to a pyrimidine (denoted Y) or a pyrimidine can change to a purine. Suppose we have the following four sequences from four different taxa:

root (4)	Y	R	Y	Y	Y	R	R
1	Y	Y	Y	R	Y	Y	R
2	R	R	Y	Y	Y	R	Y
3	R	Y	Y	R	Y	Y	Y

In this example the *character states* for these sequences are $\{R,Y\}$. The site partitions are determined by examining the character states at each site in the sequence.

Definition 23. *The character state at a site split the taxon set into subsets of taxa with a common character state. This partition is known as a **site partition**.*

For example the site partition of site 1 is $\{1,4\}, \{2,3\}$ since taxa 1 and 4 are pyrimidines while taxa 2 and 3 are purines. There are exactly 2^{n-1} bipartitions possible where n is the number of taxa being considered. It is important to be able to list the 2^{n-1} bipartitions in a logical order. The first step in doing this is to consider only the split half, or bipartition half, that does not contain the root. Without loss of generality I will consider the root to be taxa n unless otherwise noted. Next these subsets will be listed in a lexicographical order beginning with the set $N_0 = \{\emptyset\}$ and ending with $N_{2^{n-1}-1} = \{1, 2, \dots, n-1\}$ where taxa $i \in N_j$ if and only if the i^{th} binary digit of j counting from the right is 1. For example $N_{11} = \{1, 2, 4\}$ since $11 = 2^0 + 2^1 + 2^3$.

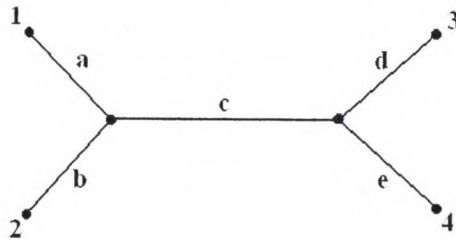
Example 13. *Below are the bipartitions and lexicographical ordering of four taxa $\{1, 2, 3, 4\}$.*

<i>Bipartitions</i>	<i>Splits</i>	<i>Lexicographical ordering</i>
B_0	$\{\emptyset\} \{1, 2, 3, 4\}$	$\{\emptyset\}$
B_1	$\{1\} \{2, 3, 4\}$	$\{1\}$
B_2	$\{2\} \{1, 3, 4\}$	$\{2\}$
B_3	$\{1, 2\} \{3, 4\}$	$\{1, 2\}$
B_4	$\{3\} \{1, 2, 4\}$	$\{3\}$
B_5	$\{1, 3\} \{2, 4\}$	$\{1, 3\}$
B_6	$\{2, 3\} \{1, 4\}$	$\{2, 3\}$
B_7	$\{1, 2, 3\} \{4\}$	$\{1, 2, 3\}$

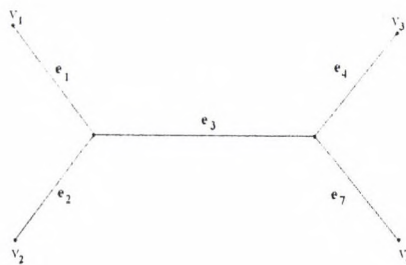
The lexicographical ordering provides a method of indexing vectors and matrices in a consistent way and provides a standard labeling of the edges in a tree. Before continuing with

the discussion of the observed sequence spectrum an example of the lexicographical ordering corresponding edge labeling is given.

Example 14. Recall the four leaf tree given in example 3.



Since removing edge a corresponds to the split $\{1\}$ which has bipartition number 1, edge a is labeled 1. Removing edge e corresponds to the split $\{1, 2, 3\}$ which has bipartition number 7, therefore edge e is labeled 7. Below is a graph with edge labels coming from the bipartitions.



The lexicographical ordering also provides the indexing for the observed sequence spectrum whose definition is given below.

Definition 24. The vector $f(D)$ is known as the **observed sequence spectrum**. Each component f_i of $f(D)$ is the expected frequency of the i^{th} bipartition as a site bipartition, where the bipartitions are numbered with the same lexicographical order.

Example 15. Recall the four 2-state sequences from above:

<i>root</i>	Y	R	Y	Y	Y	R	R
1	Y	Y	Y	R	Y	Y	R
2	R	R	Y	Y	Y	R	Y
3	R	Y	Y	R	Y	Y	Y

The site partitions for the seven sites are given below.

Site	Partition
1	{2,3} {1,4}
2	{1,3} {2,4}
3	{0} {1,2,3,4}
4	{1,3} {2,4}
5	{0} {1,2,3,4}
6	{1,3} {2,4}
7	{2,3} {1,4}

Therefore the sequence spectrum is:

$$f(D) = (2/7, 0/7, 0/7, 0/7, 0/7, 3/7, 2/7, 0/7)$$

3.1.2 Sequence Spectrum for a model with three substitution types

Site patterns are used to create an analogous description of the sequence spectrum for the three-parameter general time-reversible group-based model.

Site Patterns

Previously - when each site was represented by a purine or pyrimidine - the taxa, or leaves of the tree, were divided into a bipartition of leaves labeled R and leaves labeled Y. If sequences are made up of the nucleotides A, C, G, and T each site can now be divided into quadripartitions. Site patterns are used to keep track of which substitutions have taken place .

Suppose the DNA sequences of four taxa are given below.

root (4)	C	A	T	C	C	A	A
1	C	C	T	A	C	C	A
2	G	G	T	C	T	A	T
3	G	T	T	G	T	C	T

Since each site contains a nucleotide rather than just a purine or pyrimidine the substitutions taking place between the root and the leaves can be broken into three types. These substitution types are listed below.

$$\alpha : A \leftrightarrow G, C \leftrightarrow T$$

$$\beta : A \leftrightarrow T, C \leftrightarrow G$$

$$\gamma : A \leftrightarrow C, G \leftrightarrow T$$

If a group element from $\mathbb{Z}_2 \times \mathbb{Z}_2$ is associated to each of the nucleotides A , C , G , and T then each of the substitution types can also be associated to a group element. This is done by associating to each substitution type the difference between the group element associated to the starting nucleotide and the group element of the ending nucleotide. Let the nucleotides be assigned to the group elements of $\mathbb{Z}_2 \times \mathbb{Z}_2$ in the following way:

$$A \leftrightarrow (0,0)$$

$$C \leftrightarrow (1,1)$$

$$G \leftrightarrow (0,1)$$

$$T \leftrightarrow (1,0).$$

Therefore substitution type α corresponds to

$$(0,0) - (0,1) = (1,1) - (1,0) = (0,1) \rightarrow G.$$

Referring back to the table containing the four nucleotide sequences the substitution in site one, from the root to taxon 3, is of type β which can be represented as $(1,0)$. Determining the substitutions of each leaf for each site yields the following table.

site:	1	2	3	4	5	6	7
root to 1	(00)	(11)	(00)	(11)	(00)	(11)	(00)
root to 2	(10)	(01)	(00)	(00)	(01)	(00)	(10)
root to 3	(10)	(10)	(00)	(10)	(01)	(11)	(10)

Next notice that for any given site a 1 appearing in the first entry of an ordered pair implies that a substitution of type β or γ took place between the root and that leaf. Similarly if a 1 appears in the second entry of an ordered pair a substitution of type α or γ took place between the root and that leaf. Given this, let the set A be defined for a given site as the set of leaves with a 1 in first position of the ordered pair. Let the set B be defined for a given site as the set of leaves with a 1 in the second position of the ordered pair. The sets A and B for the example above are:

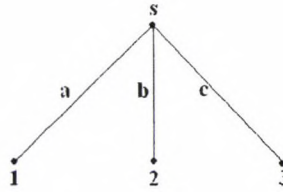
site:	1	2	3	4	5	6	7
A	{2,3}	{1,3}	{ \emptyset }	{1,3}	{ \emptyset }	{1,3}	{2,3}
B	{ \emptyset }	{1,2}	{ \emptyset }	{1}	{2,3}	{1,3}	{ \emptyset }

If a leaf belongs to the sets A and B then a γ substitution took place. If a leaf belongs to just A then a β substitution took place and if a leaf belongs to just B then an α substitution took place. Therefore (A, B) gives the substitutions that took place between the root and each leaf at a given site. (A, B) is known as a *site pattern*.

The number of possible site patterns (A, B) depends on the number of taxa being considered. Notice that each set A and B are possible bipartition halves which do not contain the root. Therefore there are 2^{n-1} possibilities for both sets A and B . Hence there are $2^{n-1} \times 2^{n-1} = 4^{n-1}$ possibilities for (A, B) . An ordering of the sets (A, B) that will be used later on is to fix B as the first bipartition half given in the lexicographical ordering above and let the set A range through the bipartition half ordering given above. Then let B be the next bipartition half in the ordering and let A range through the values again. Continue this until the final pair listed is (A, B) where A and B are the sets $\{1, 2, 3, \dots, n-1\}$. These sets (A, B) are also referred to as *quadripartitions*, a partition into four or fewer subsets of the taxa. This is because each set A and B corresponds to a split half not containing the root vertex. Therefore (A, B) refers to four subsets of the taxa.

In the case of a three-substitution type model the i^{th} component of the observed sequence spectrum contains the expected frequency of the i^{th} site pattern where the site patterns are listed in the order given above.

Example 16. Consider the three leaf tree below.



If leaf 3 is the root vertex then the possible bipartitions of the leaves not containing the root are

$$B_0 = \{\emptyset\}, \quad B_1 = \{1\}, \quad B_2 = \{2\}, \quad B_3 = \{1, 2\}.$$

All site patterns of the tree above are of the form (B_i, B_j) where $0 \leq i \leq 3$ and $0 \leq j \leq 3$. The ordering of the site patterns as described above is:

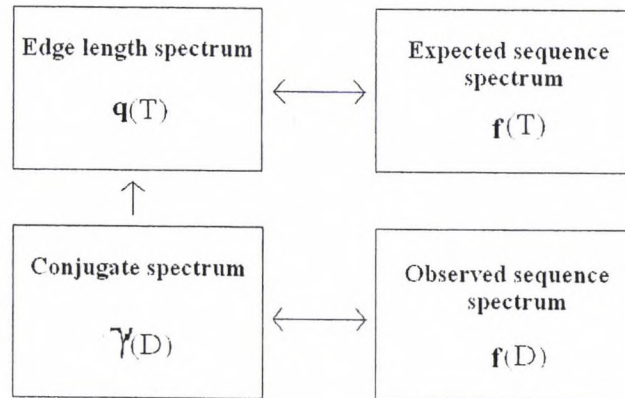
$$(B_0, B_0), (B_1, B_0), (B_2, B_0), (B_3, B_0), (B_0, B_1), (B_1, B_1), (B_2, B_1), (B_3, B_1), \\ (B_0, B_2), (B_1, B_2), (B_2, B_2), (B_3, B_2), (B_0, B_3), (B_1, B_3), (B_2, B_3), (B_3, B_3).$$

3.1.3 Relating the observed sequence spectrum to a tree

The observed sequence spectrum is obtained from aligned DNA sequences that may contain some error. There are also assumptions made in the evolutionary model that will produce error. As a result it is generally not the case that applying Hadamard conjugation to the expected sequence spectrum will produce a vector that exactly corresponds to a phylogenetic tree; rather it produces a vector known as the conjugate spectrum, $\gamma(D)$. A fitting algorithm such as least-squares best fit is then used to obtain the edge length spectrum, $\mathbf{q}(T)$, which corresponds exactly to a phylogenetic tree. In the case the that $\gamma(D) = \mathbf{q}(T)$ the data fits a model exactly and no fitting algorithm is necessary.

If a tree is assumed initially it is possible write down the expected sequence spectrum, $\mathbf{f}(T)$. This is the vector that the observed sequence spectrum, $\mathbf{f}(D)$, hopes to approximate. If Hadamard conjugation is applied to $\mathbf{f}(T)$ the edge length spectrum $\mathbf{q}(T)$ is obtained. This result is given in the following section along with further explanation of the expected sequence spectrum and the

edge length spectrum. Before moving to the next section a diagram is provided to make clear the relationships between the vectors $\mathbf{f}(D)$, $\gamma(D)$, $\mathbf{q}(T)$, and $\mathbf{f}(T)$.



The double arrows are invertible Hadamard conjugations and the single arrow represents a fitting algorithm [28].

3.2 The edge length spectrum as it relates to the expected sequence spectrum

Ideally it would be nice to use the observed sequence spectrum, $\mathbf{f}(D)$, to say something about the phylogenetic tree associated to the sequences being examined. To do this Fourier calculus over a finite abelian group along with a fitting algorithm such as least-squares or the closest tree algorithm can be used. The following section contains the results of Székely, Steel and Erdős from their 1993 paper “Fourier Calculus on Evolutionary Trees” [60] as they apply to phylogenetic trees. In chapter 4 I will extend these results to splits networks.

The first lemma is a well known result which summarizes what is needed on characters and the Fourier transform. The statement of the lemma is taken from [60].

Lemma 2. *Let G be a finite abelian group, then*

- (i) *the character group \hat{G} is isomorphic to G .*

(ii) if $f : G \rightarrow C$ is a complex-valued function and $\hat{f} : \hat{G} \rightarrow C$ is defined by

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g) f(g),$$

then for all $g \in G$

$$f(g) = \frac{1}{|G|} \sum_{\chi \in \hat{G}} \chi(\bar{g}) \hat{f}(\chi).$$

(iii) The characters of a finite direct product of finite abelian groups are exactly the products of characters.

3.2.1 Preliminary Notation

The following notation will be given in general without respect to a phylogenetic tree, however its interpretation in the phylogenetic setting will be given when appropriate as it is used.

Let \mathbf{A} be a $p \times q$ matrix with integer entries. Let G be a finite abelian group with elements of G^q written in the vector form $\mathbf{x} = (x_1, x_2, \dots, x_q)^T$, where $x_j \in G$. Similarly let $\mathbf{y} \in G^p$ be the vector $\mathbf{y} = (y_1, y_2, \dots, y_p)$ such that

$$y_i = \sum_{j=1}^q a_{ij} x_j$$

so that $\mathbf{Ax} = \mathbf{y}$.

Next let $p_j : G \rightarrow C$ for $1 \leq j \leq q$ and let

$$F(\mathbf{x}) = \prod_{j=1}^q p_j(x_j). \quad (3.1)$$

For $\mathbf{y} \in G^p$, let

$$f(\mathbf{y}) = \sum_{\substack{\mathbf{x} \in G^q \\ \mathbf{Ax} = \mathbf{y}}} F(\mathbf{x}). \quad (3.2)$$

The next theorem is a general result which will be restated in the phylogenetic setting in theorem 9.

Theorem 6. [60] If $\chi = (\chi_1, \dots, \chi_p)^T \in \hat{G}^p$, then

$$\hat{f}(\chi) = \prod_{j=1}^q \sum_{x \in G^q} p_j(x_j) \prod_{i=1}^p \chi_i(a_{ij} x_j).$$

Proof. By part ii of Lemma 2,

$$\begin{aligned}\hat{f}(\chi) &= \sum_{\mathbf{y} \in G^p} \chi(\mathbf{y}) f(\mathbf{y}) \\ &= \sum_{\mathbf{y} \in G^p} \chi(\mathbf{y}) \cdot \sum_{\substack{\mathbf{x} \in G^q \\ A\mathbf{x}=\mathbf{y}}} F(\mathbf{x})\end{aligned}$$

By expanding and regrouping it can be seen that

$$\sum_{\mathbf{y} \in G^p} \chi(\mathbf{y}) \cdot \sum_{\substack{\mathbf{x} \in G^q \\ A\mathbf{x}=\mathbf{y}}} F(\mathbf{x}) = \sum_{\mathbf{x} \in G^q} F(\mathbf{x}) \chi(A\mathbf{x}).$$

Next notice that by definition,

$$\begin{aligned}\chi(A\mathbf{x}) &= \prod_{i=1}^p \chi_i((A\mathbf{x})_i) \\ &= \prod_{i=1}^p \chi_i \left(\sum_{j=1}^q a_{ij} x_j \right).\end{aligned}$$

Using Lemma 2 and expanding and regrouping once again gives,

$$\prod_{i=1}^p \chi_i \left(\sum_{j=1}^q a_{ij} x_j \right) = \prod_{j=1}^q \prod_{i=1}^p \chi_i(a_{ij} x_j).$$

Therefore,

$$\begin{aligned}\hat{f}(\chi) &= \sum_{\mathbf{x} \in G^q} F(\mathbf{x}) \chi(A\mathbf{x}) \\ &= \sum_{\mathbf{x} \in G^q} \prod_{j=1}^q p_j(x_j) \prod_{i=1}^p \chi_i(a_{ij} x_j) \\ &= \prod_{j=1}^q \sum_{\mathbf{x} \in G^q} p_j(x_j) \prod_{i=1}^p \chi_i(a_{ij} x_j)\end{aligned}$$

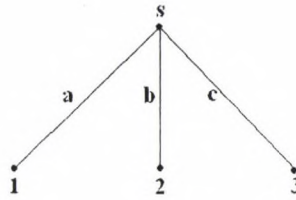
which is the desired equality. □

3.2.2 Notation related to phylogenetic trees

Let n be the number of taxa being considered. Then n is also the number of leaves in the phylogenetic tree T with leaf set L and $n - 1 = p$ is the number of non-root leaves. The root leaf is referred to as R and the number of edges in the tree is q .

The \mathbf{A} Matrix

The matrix \mathbf{A} was defined in section 3.2.1 as a $p \times q$ integer matrix. In this section the \mathbf{A} matrix will be defined as an $(n - 1) \times q$ integer matrix which depends on the phylogenetic tree, T . It is a matrix whose rows are indexed by the non-root leaves and columns indexed by number of edges in the tree. For example, consider the 3-claw tree shown below with root vertex labeled 3. This tree has $q = 3$ and $n - 1 = 2$.



The \mathbf{A} matrix corresponding to this tree is a 2×3 matrix with rows indexed by the non-root leaves, 1 and 2, and columns indexed by the edges, a , b and c . The matrix is defined by

$$\mathbf{A}_{i,k} = \begin{cases} 1 & \text{if } i \text{ and } n \text{ are on opposite sides of } \{A_k, B_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the \mathbf{A} matrix corresponding to the tree above is

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

It is possible to read information about the tree off of the columns of \mathbf{A} . In this example, the columns of \mathbf{A} show that if edge a is removed then leaf 1 is in the opposite split half from the root while leaf 2 is in the same split half as the root. Also, if edge b is removed then leaf 2 is in the opposite split half while 1 is in the same split half as the root. If edge c is removed then both leaves 1 and 2 are in the opposite split half from the root.

Information can also be read from the rows of the \mathbf{A} matrix. For example the ones in the columns indexed by a and c of the row indexed by 1 show that the path from the root to leaf 1

includes edges a and c . The row indexed by 2 shows that the path from the root to leaf 2 includes edges b and c . In general, the ones appearing in row i , for i a non-root leaf, indicate the edges contained in the path from the root to leaf i .

The \mathbf{A} matrix will be used along with the vectors \mathbf{x} and \mathbf{y} which were defined previously.

Leaf colorations

For each edge $e \in E(T)$ there exists an independent G -valued random variable ξ_e with distributions $p_e(g) := \text{Prob}(\xi_e = g)$ such that $\sum_{g \in G} p_e(g) = 1$. The set of p_e are known as a transition mechanism and is denoted by p .

The map $\sigma : L \setminus \{R\} \rightarrow G$ gives a leaf coloration. The set of leaf colorations is denoted G^{n-1} . The value of σ at leaf $l \in L$ is denoted σ_l . In order to produce a random G -coloration of the leaves of the tree evaluate ξ_e for every edge and assign to l the color which is the sum of the group elements along the unique Rl path. Let f_σ be the probability of obtaining the leaf coloration $\sigma : L \setminus \{R\} \rightarrow G$ in this way. Notice that f_σ is an entry in the vector $\mathbf{f}(T)$, the expected sequence spectrum discussed above.

Next let $\chi = (\chi_l \in \hat{G} : l \in L \setminus \{R\})$ be an ordered $(n-1)$ -tuple of characters. Then $\chi \in \hat{G}^{n-1}$, and χ acts on G^{n-1} according to Lemma 2 (iii). Given this define the set

$$L_e = \{l \in L : e \text{ separates } l \text{ from } R \text{ in } T\}$$

for $e \in E(T)$. Notice this is just the set of leaves in the opposite split half from the root after removing edge e .

Equations

For $e \in E(T)$ and $\chi \in \hat{G}^{n-1}$, set

$$\chi_e = \sum_{l \in L_e} \chi_l \tag{3.3}$$

so $\chi_e \in \hat{G}$. For $h \in \hat{G}$, $e \in E(T)$ define

$$l_e(h) = \sum_{g \in G} h(g) p_e(g), \tag{3.4}$$

$$r_\chi = \prod_{e \in E(T)} l_e(\chi_e). \quad (3.5)$$

At this point it is possible to state the Fourier inverse pair in the phylogenetic setting.

3.2.3 Results

Theorem 7. [60] With $\chi(\sigma) = \prod_{l \in L \setminus \{R\}} \chi_l(\sigma_l)$,

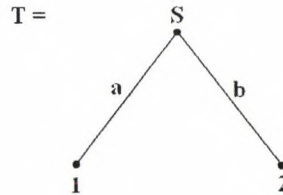
$$r_\chi = \sum_{\sigma \in G^{n-1}} \chi(\sigma) f_\sigma \quad (3.6)$$

$$f_\sigma = \frac{1}{|G|^{n-1}} \sum_{\chi \in \hat{G}^{n-1}} \chi(\bar{\sigma}) r_\chi. \quad (3.7)$$

See Székely et al for a proof to Theorem 9.

Theorem 9 provides an equation for r_χ which is a specific example of $\hat{f}(\chi)$ from theorem 6. The following small example shows that $r_\chi = \hat{f}(\chi)$ for a specific tree with probabilities each edge is assigned a specific group element.

Example 17. Consider the tree with root vertex 1 given below. (Note that the root vertex is no longer the leaf labeled with the greatest number.)



Make the following association:

$$A \leftrightarrow (0, 0)$$

$$C \leftrightarrow (1, 1)$$

$$G \leftrightarrow (0, 1)$$

$$T \leftrightarrow (1, 0)$$

Let

$$p_a(A) = p_b(A) = \frac{1}{2}$$

$$p_a(C) = p_b(C) = \frac{1}{4}$$

$$p_a(G) = p_b(G) = \frac{1}{8}$$

$$p_a(T) = p_b(T) = \frac{1}{8}$$

Since there is only one non-root leaf there are four σ maps. The four maps are σ_1 which colors the non-root leaf with A, σ_2 which colors the non-root leaf with C, σ_3 which colors the non-root leaf with G, and σ_4 which colors the non-root leaf with T. Without loss of generality I will assume that the root is labeled A.

Recall that f_σ is the probability of obtaining the leaf coloration σ . Given the probabilities above it is straight forward to compute f_{σ_i} for $i \in \{1, 2, 3, 4\}$. Below is the computation for f_{σ_1} . The four possible edge colorings that would yield $\sigma_{1_2} = A$, where σ_{1_2} is the value of σ_1 at leaf 2, are

$$a \mapsto A \text{ and } b \mapsto A$$

$$a \mapsto C \text{ and } b \mapsto C$$

$$a \mapsto G \text{ and } b \mapsto G$$

$$a \mapsto T \text{ and } b \mapsto T.$$

The probabilities of these edge colorings are $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$, $(\frac{1}{4})(\frac{1}{4}) = \frac{1}{16}$, $(\frac{1}{8})(\frac{1}{8}) = \frac{1}{64}$ and $(\frac{1}{8})(\frac{1}{8}) = \frac{1}{64}$, respectively. Therefore $f_{\sigma_1} = \frac{1}{4} + \frac{1}{16} + \frac{1}{32} = \frac{11}{32} = 0.34375$.

A similar computation shows $f_{\sigma_2} = 0.28125$, $f_{\sigma_3} = 0.1875$, and $f_{\sigma_4} = 0.1875$. Given the probabilities f_{σ_i} it is possible to compute $\hat{f}(\chi)$ and r_χ .

$$\hat{f}(\chi) = \sum_{\mathbf{x} \in G^q} \prod_{j=1}^q p_j(x_j) \prod_{i=1}^p \chi_i(a_{ij}x_j)$$

which for this example is

$$\begin{aligned}\hat{f}(\chi) &= \sum_{x \in G^q} p_a(x_1)\chi_1(x_1)p_b(x_2)\chi_1(x_2) \\ &= p_a(A)\chi_1(A)p_b(A)\chi_1(A) + p_a(A)\chi_1(A)p_b(C)\chi_1(C) + \dots \\ &\quad + p_a(T)\chi_1(T)p_b(G)\chi_1(G) + p_a(T)\chi_1(T)p_b(T)\chi_1(T).\end{aligned}$$

The value of the above equation depends on which group element is associated to the character χ_1 . If χ_1 is associated to A the sum is 1, if it is C the sum is 0.0625, if it is G the sum is 0.25, and if χ_1 is associated to T the sum is 0.0625.

Next I will compute r_χ for this example.

$$\begin{aligned}r_\chi &= \sum_{\sigma \in G^{n-1}} \chi_1(\sigma_1)f_\sigma \\ &= \chi_1(\sigma_{11})f_{\sigma_1} + \chi_1(\sigma_{21})f_{\sigma_2} + \chi_1(\sigma_{31})f_{\sigma_3} + \chi_1(\sigma_{41})f_{\sigma_4} \\ &= \pm 0.34375 \pm 0.28125 \pm 0.1875 \pm 0.1875\end{aligned}$$

The sign of each term depends on χ_1 . Again if χ_1 is associated to A the sum is 1, if it is C the sum is 0.0625, if it is G the sum is 0.25, and if χ_1 is associated to T the sum is 0.0625. These values are precisely the values from $\hat{f}(\chi)$ above.

In order to state the main result one more piece of notation is needed. For $e \in E(T)$ and $0 \neq g \in G$ define $\rho^{e,g} \in G^{n-1}$ so that $\rho_l^{e,g} = 0$ for $l \notin L_e$, where l is not the root and $\rho_l^{e,g} = g$ for $l \in L_e$. Given $\rho^{e,g}$ let $\mathcal{C}(T) = \{\rho^{e,g} : e \in E(T), 0 \neq g \in G\}$.

Theorem 8. [60] For $0_{G^{n-1}} \neq \rho \in G^{n-1}$, $\rho \notin \mathcal{C}(T)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = 1;$$

for $\rho = \rho^{e,g} \in \mathcal{C}(T)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{h \in \hat{G}} l_e(h)^{h(g)|G|^{n-2}};$$

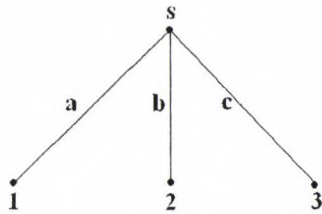
and for $\rho = 0_{G^{n-1}}$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{e \in E(T)} \prod_{h \in \hat{G}} l_e(h)^{|G|^{n-2}}.$$

This result will be stated and proven in the next chapter for a more general setting and therefore the proof has been omitted here. For a proof to this statement see [60].

3.2.4 An example assuming the Jukes-Cantor model

In order to keep the example manageable I will use the 3-claw tree with root vertex 1. For convenience it has been given below.



Also to keep things simple the Jukes and Cantor model will be assumed. Recall that the Jukes and Cantor nucleotide substitution model assumes that transitions and both transversions occur with equal probability. Theorem 11 can accommodate the Kimura three-substitution type nucleotide substitution model which allows for different probabilities for the transitions and each transversion. The three probabilities are denoted by p_1, p_2, p_3 , so for this example where the Jukes and Cantor model is assumed $p_1 = p_2 = p_3 = p$. The probability that no nucleotide substitution has taken place across an edge is $p_0 = 1 - 3p$. To begin assume that the probability of a nucleotide substitution occurring along edge a is the same as the probability of a substitution occurring along edges b and c . Later in this section we will see what happens when these probabilities are not equal.

Next recall that \mathbf{f} is the vector of length 4^{n-1} with entries f_σ , the probability of obtaining the leaf coloration σ . For this example let the root vertex be colored with the group element associated to A , then leaves 2 and 3 can be colored by any of A, C, G , or T . This implies there are 16 different colorations $\sigma_1, \sigma_2, \dots, \sigma_{16}$. To calculate f_{σ_i} for $1 \leq i \leq 16$ it is necessary to determine the four possible ways the coloration could be obtained.

For example, in order to obtain the coloration (A, C) where leaf 2 is colored A and leaf 3 is colored C , it is possible that no substitution took place on edges a and b and the substitution associated to C took place on edge c . To check that this gives the appropriate coloration check to see that sum of the group elements along the path from the root to the leaf in question is equal to the color of the leaf. Again recall that A, C, G , and T are associated to the elements of $\mathbb{Z}_2 \times \mathbb{Z}_2$ in the following way.

$$A \leftrightarrow (0, 0)$$

$$C \leftrightarrow (1, 1)$$

$$G \leftrightarrow (0, 1)$$

$$T \leftrightarrow (1, 0)$$

The probability of the substitutions given above occurring on the edges a, b and c is $(1 - 3p)(1 - 3p)p$. The remaining three substitutions resulting in (A, C) and their corresponding probabilities are given in the table.

$(A, C) :$	$a \mapsto A$	$a \mapsto C$	$a \mapsto G$	$a \mapsto T$
	$b \mapsto A$	$b \mapsto C$	$b \mapsto G$	$b \mapsto T$
	$c \mapsto C$	$c \mapsto A$	$c \mapsto T$	$c \mapsto G$
probability of given edge substitutions	$(1 - 3p)^2 p$	$(1 - 3p)p^2$	p^3	p^3

Therefore f_{σ^2} the probability of obtaining the edge coloring (A, C) is

$$(1 - 3p)^2 p + (1 - 3p)p^2 + 2p^3.$$

Doing a similar computation for the remaining fifteen σ yields the entries contained in \mathbf{f} below.

$$\mathbf{f} = \begin{bmatrix} (1-3p)^3 + 3p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ 3(1-3p)p^2 + p^3 \\ 3(1-3p)p^2 + p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ 3(1-3p)p^2 + p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ 3(1-3p)p^2 + p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \\ 3(1-3p)p^2 + p^3 \\ 3(1-3p)p^2 + p^3 \\ (1-3p)^2p + (1-3p)p^2 + 2p^3 \end{bmatrix}$$

If the probability of a substitution taking place across a given edge is $p = 0.01$ then the resulting vector \mathbf{f} is

$$\mathbf{f} = \begin{bmatrix} 0.912676 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.000292 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.000292 \\ 0.009508 \end{bmatrix}$$

In order to find the edge length spectrum according to Theorem 11 the matrix $\mathbf{H} = [\chi(\sigma)]$ is required. \mathbf{H} has rows corresponding to $\chi \in \hat{G}^{n-1}$ and columns corresponding to $\sigma \in G^{n-1}$ and

is equivalent to the sixteen by sixteen Hadamard matrix.

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Given \mathbf{f} and \mathbf{H} it is now possible to compute the edge length spectrum from Theorem 11. Below is the symbolic edge length spectrum from this example as well as the edge length spectrum assuming the probability of a nucleotide substitution taking place across an edge is 0.01.

$$[\mathbf{H}^{-1} \ln \mathbf{H} \mathbf{f}]_p = \begin{bmatrix} \frac{9}{4} \ln(1-4p) \\ -\frac{1}{4} \ln(1-4p) \\ -\frac{1}{4} \ln(1-4p) \\ -\frac{1}{4} \ln(1-4p) \\ -\frac{1}{4} \ln(1-4p) \\ -\frac{1}{4} \ln(1-4p) \\ 0 \\ 0 \\ -\frac{1}{4} \ln(1-4p) \\ 0 \\ -\frac{1}{4} \ln(1-4p) \\ 0 \\ -\frac{1}{4} \ln(1-4p) \\ 0 \\ 0 \\ -\frac{1}{4} \ln(1-4p) \end{bmatrix} = \begin{bmatrix} -0.0918494876 \\ 0.0102054986 \\ 0.0102054986 \\ 0.0102054986 \\ 0.0102054986 \\ 0.0102054986 \\ 0 \\ 0 \\ 0.0102054986 \\ 0 \\ 0.0102054986 \\ 0 \\ 0.0102054986 \\ 0 \\ 0 \\ 0.0102054986 \end{bmatrix} \text{ for } p = 0.01.$$

Notice there is some rounding error due to taking the natural logarithm.

The entries in the vectors above are indexed by $\rho^{e,g} \in G^{n-1}$, which can be thought of as leaf colorations on the non-root leaves. These leaf colorations correspond to the site patterns

that were discussed in the site patterns section 3.1.2. The correspondence works as follows. A site pattern, (A, B) , is made up of sets A and B each containing non-root leaves. If leaf $i \in A$ and $i \in B$ then a substitution of type C took place. If leaf $i \in A$ and $i \notin B$ then a substitution of type T took place, while if leaf $i \notin A$ and $i \in B$ then a substitution of type G took place. Therefore if $(\{2\}, \{2, 3\})$ is the site partition then substitution type C took place between the root and leaf 2 and substitution type G took place between the root and leaf 3. This implies that the site pattern $(\{2\}, \{2, 3\})$ corresponds to the coloration (C, G) . A benefit to this correspondence is that the same ordering described in section 3.1.2 can be used to order the colorations and therefore order $\rho^{e,g} \in G^{n-1}$.

Example 18. For the three claw example the ordering of the leaf colorations is: (A, A) , (A, C) , (A, G) , (A, T) , (C, A) , (C, C) , (C, G) , (C, T) , (G, A) , (G, C) , (G, T) , (T, A) , (T, C) , (T, T) .

Notice that Theorem 11 states that the entries of the edge length spectrum will have the form 0 , $[\mathbf{K}^{-1} \ln \mathbf{K} \mathbf{p}_e]_h$, or $\sum_{e \in E(T)} \sum_{h \in G} [\mathbf{K}^{-1} \ln \mathbf{K} \mathbf{p}_e]_h$ depending on ρ .

Chapter 4

HADAMARD CONJUGATION AND SPLITS NETWORKS: AN EXTENSION TO SPLITS NETWORKS

In 2005 David Bryant published a chapter in Loir Pachter and Bernd Sturmfels' text, Algebraic Statistics for Computational Biology [8] in which he claims to have extended Székely et al's result from phylogenetic trees to splits networks. Bryant attempts to do this by taking a slightly different approach from Székely et al. In verifying the results of this chapter I found that there appear to be a few errors with Bryant's result as it is published in [8]. There appear to be both an indexing error as well as a counterexample to a lemma Bryant uses to prove the extension of Székely et al's result. For further details regarding these errors see appendices A.2 and A.3 on page 119.

The goal of this chapter is to prove a generalization of Székely et al's result which will make it possible to take information derived from a set of nucleotide sequences and apply a Fourier transform in order to obtain a set of splits corresponding to a splits network. Some of the definitions and results from the previous chapter will be repeated here for clarity.

4.1 Notation

Choosing the appropriate notation is critical to extending Székely et al's results to splits networks. By carefully choosing notation and making the appropriate definitions it is possible to follow the approach of Székely et al in [60].

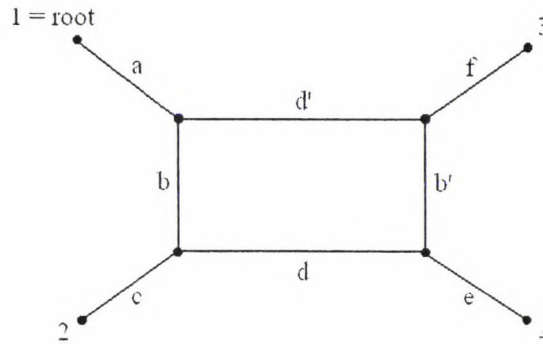
4.1.1 A Matrix

As in the previous chapter let \mathbf{A} be a $p \times q$ matrix with zeros and ones. Let G be a finite abelian group with elements of G^q written in the vector form $\mathbf{x} = (x_1, x_2, \dots, x_q)^T$, where $x_j \in G$ and let $\mathbf{y} \in G^p$ be the vector $\mathbf{y} = (y_1, y_2, \dots, y_p)$ such that

$$y_i = \sum_{j=1}^q a_{ij}x_j.$$

Although this property will be abbreviated as $\mathbf{Ax} = \mathbf{y}$ notice that multiplying a group element by 0 yields the identity element while multiplying a group element by 1 yields the group element.

In the context of phylogenetic networks the \mathbf{A} matrix is a $(n-1) \times q$ matrix with zeros and ones which depends on the network N . The number of taxa being considered is n , so $n-1$ is the number of non-root leaves while q is the number of color classes. For example, consider the splits network given below.



This splits network has three non-root leaves and six color classes. Notice that an isometric coloring of the graph forces edges b and b' as well as d and d' to have the same coloring. Therefore the \mathbf{A} matrix corresponding to this network is a 3×6 zero one matrix with rows

indexed by the leaves 2, 3, and 4 and columns indexed by the color classes, $\{a\}$, $\{b, b'\}$, $\{c\}$, $\{d, d'\}$, $\{e\}$ and $\{f\}$. Formally the entries in the matrix are given by the following definition.

$$\mathbf{A}_{i,k} = \begin{cases} 1 & \text{if } i \text{ and } n \text{ are on opposite sides of } \{A_k\}|\{B_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the \mathbf{A} matrix corresponding to the tree above is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

It is possible to read information about the splits network from the columns of \mathbf{A} . The entries in each column indicate which leaves are in the opposite split half from the root if the edges by which the column is indexed are removed. For example, if edge a is removed then all three leaves, 2, 3, and 4, are in the opposite split half from the root. Therefore the column indexed by $\{a\}$ contains all ones. The column indexed by $\{b, b'\}$ only has ones in the rows indexed by 2 and 4 since leaves 2 and 4 are in the opposite split half from the root when $\{b, b'\}$ is removed.

Information can also be read from the rows of the \mathbf{A} matrix. By reading across a row in the \mathbf{A} matrix it is possible to see which edges are contained in the shortest path between the root and the leaf by which the row is indexed. In the above example the row indexed by leaf 2 has ones in the columns indexed by $\{a\}$, $\{b, b'\}$ and $\{c\}$ since the shortest path from the root to leaf 2 is a, b, c or a, b', c .

4.1.2 Leaf Colorations

In order to introduce the Fourier inverse pair in the phylogenetic splits network setting some additional notation is needed. Namely it is necessary discuss leaf colorations.

For each color class $e_C \subset E(N)$, where $E(N)$ is the set of edges in the splits network N , there exists an independent G -valued random variable ξ_{e_C} with distributions $p_{e_C}(g) := \text{Prob}(\xi = g)$

such that $\sum_{g \in G} p_{e_C}(g) = 1$. The set of p_{e_C} are known as a transition mechanism and is denoted by p .

The map $\sigma : L \setminus \{R\} \rightarrow G$ is a leaf coloration, that is a map which assigns to each non-root leaf a group element g . The set of leaf colorations is denoted G^{n-1} . The value of σ at leaf $l \in L$ is denoted σ_l . In order to produce a random G -coloration of the leaves of the network evaluate ξ_{e_C} for every edge and assign to l the color which is the sum of the group elements along the unique Rl path. Let f_σ be the probability of obtaining the leaf coloration $\sigma : L \setminus \{R\} \rightarrow G$ in this way.

Next let $\chi = (\chi_l \in \hat{G} : l \in L \setminus \{R\})$ be an ordered $(n-1)$ -tuple of characters. Then $\chi \in \hat{G}^{n-1}$, and χ acts on G^{n-1} according to Lemma 2 (iii). Given this define the set

$$L_{e_C} = \{l \in L : e_C \text{ separates } l \text{ from } R \text{ in } N\}$$

for $e_C \subset E(N)$. Notice this is just the set of leaves in the opposite split half from the root after removing color class e_C .

For $e_C \subset E(N)$ and $\chi \in \hat{G}^{n-1}$, set

$$\chi_{e_C} = \sum_{l \in L_{e_C}} \chi_l \tag{4.1}$$

so $\chi_{e_C} \in \hat{G}$. For $h \in \hat{G}$, $e_C \subset E(N)$ define

$$l_{e_C}(h) = \sum_{g \in G} h(g) p_{e_C}(g), \tag{4.2}$$

$$r_\chi = \prod_{e_C \subset E(N)} l_{e_C}(\chi_{e_C}). \tag{4.3}$$

4.1.3 Preliminary Result

At this point it is possible to state the Fourier inverse pair in the phylogenetic setting.

Theorem 9. [60] With $\chi(\sigma) = \prod_{l \in L \setminus \{R\}} \chi_l(\sigma_l)$,

$$r_\chi = \sum_{\sigma \in \mathcal{G}^{n-1}} \chi(\sigma) f_\sigma \quad (4.4)$$

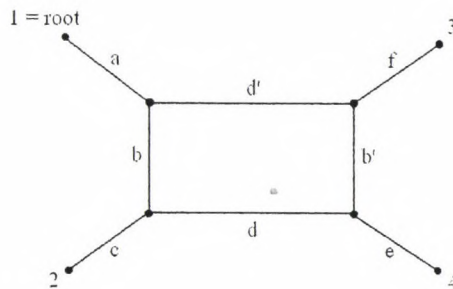
$$f_\sigma = \frac{1}{|G|^{n-1}} \sum_{\chi \in \mathcal{G}^{n-1}} \overline{\chi(\sigma)} r_\chi. \quad (4.5)$$

The proof in Székely et al [60] of Theorem 9 carries over to splits networks mutatis mutandis. This is a consequence of the very specific structure of splits networks. If each edge in a phylogenetic tree is given a probability that it will be assigned a given character state then it is possible to determine the probability of obtaining a given leaf coloration. This is because the color of each leaf is equal to the sum of the edge labelings along the path between the root and that leaf. Therefore the probability that a leaf in the tree will be colored a certain way depends on the probabilities along the path from the root to that leaf. In the case of a splits network it is required that edges in the same color class must be assigned the same probabilities. The result is that even though there may be more than one path between the root and a leaf, the probabilities along each of those paths must be the same. Essentially, the property that in a phylogenetic tree there is a unique path between any two vertices corresponds to the fact that in splits networks although there may be more the one path all paths must be labeled the same way.

Theorem 9 provides an equation for r_χ which is equivalent to $\hat{f}(\chi)$ from theorem 6. The next section provides an example of this for a specific splits network.

4.2 An illustrated example

Consider the splits network below.



As mentioned on page 66 the \mathbf{A} matrix corresponding to this splits network is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

For simplicity assume a 2-state evolutionary model. This implies that when considering a specific site of a set of aligned sequences edges and leaves of a splits network are labeled by purines (R) and pyrimidines (Y) rather than by the nucleotides A , C , G and T . The purines and pyrimidines correspond to the group elements of \mathbb{Z}_2 by associating

$$R \leftrightarrow 0$$

$$Y \leftrightarrow 1.$$

If a mutation takes place across an edge the edge is labeled by the group element 1, whereas if no mutation takes place the edge is labeled by the identity element 0. Each edge labeling determines exactly one leaf labeling. For example if color classes $\{a\}$, $\{b, b'\}$ and $\{c\}$ are labeled 1, 0, 1 respectively and the root vertex is labeled by a purine, R , then leaf 2 must also be labeled R since adding the group elements along the path from the root to leaf 2 is $1 + 0 + 1 = 0$. Again notice that the paths a, b, c and a, b', c are indistinguishable.

Next in order to calculate r_χ and $\hat{f}(\chi)$ it is necessary to give the probability that a given color class is labeled by either a 0 (R) or a 1 (Y). For this example let $p_{\{a\}}(R) = 0.95$, $p_{\{b, b'\}}(R) = 0.9$, $p_{\{c\}}(R) = 0.85$, $p_{\{d, d'\}}(R) = 0.8$, $p_{\{e\}}(R) = 0.75$ and $p_{\{f\}}(R) = 0.7$ and note that $p_{e_c}(R) + p_{e_c}(Y) = 1$. In the 2-state model it is necessary for $p_{e_c}(R) > 0.5$. In the biological setting this corresponds to the probability of a mutation taking place being less than one half.

Recall from Theorem 6 that

$$\hat{f}(\chi) = \sum_{\mathbf{x} \in G^q} \prod_{j=1}^q p_j(x_j) \prod_{i=1}^p \chi_i(a_{ij}x_j)$$

and let p equal the number of non-root leaves and q equal the number of color classes. Since $\chi = (\chi_l \in \hat{G} : l \in L \setminus \{R\})$ there are eight possible values for χ and therefore eight $\hat{f}(\chi)$. These eight values are contained in the \hat{f} vector.

The last item to note is that since we are considering a 2-state model with character states R and Y corresponding to the group elements of \mathbb{Z}_2 the character table used to determine values of χ is

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

the character table for \mathbb{Z}_2 . Given this it is possible to calculate the values of $\hat{f}(\chi)$ using the equation above. Since the length of the computation is quite long only one term of the sum which yields $\hat{f}(\chi)$ for $\chi = (\chi_R, \chi_Y, \chi_Y)$ is calculated. Let $\mathbf{x} = (1, 1, 1, 1, 1, 1)$, then the term in the sum corresponding \mathbf{x} is

$$\begin{aligned} \prod_{j=1}^6 p_j(x_j) \prod_{i=1}^3 \chi_i(a_{ij}x_j) &= p_1(x_1)\chi_1(a_{1,1}x_1)\chi_2(a_{2,1}x_1)\chi_3(a_{3,1}x_1) \cdot \dots \cdot \\ &\quad p_6(x_6)\chi_1(a_{1,6}x_6)\chi_2(a_{2,6}x_6)\chi_3(a_{3,6}x_6) \\ &= p_1(1)\chi_1(1 \cdot 1)\chi_2(1 \cdot 1)\chi_3(1 \cdot 1) \cdot \dots \cdot \\ &\quad p_6(1)\chi_1(0 \cdot 1)\chi_2(1 \cdot 1)\chi_3(0 \cdot 1) \\ &= 0.05 \cdot \chi_1(1)\chi_2(1)\chi_3(1) \cdot \dots \cdot 0.3 \cdot \chi_1(0)\chi_2(1)\chi_3(0) \\ &= 0.05(1)(-1)(-1) \cdot \dots \cdot 0.3(1)(-1)(1). \end{aligned}$$

Repeating the computation for each term in the sum and for each χ yields the following values for $\hat{f}(\chi)$:

$$\begin{aligned}
\hat{f}(\chi_R, \chi_R, \chi_R) &= 1 \\
\hat{f}(\chi_R, \chi_R, \chi_Y) &= 0.216 \\
\hat{f}(\chi_R, \chi_Y, \chi_R) &= 0.216 \\
\hat{f}(\chi_R, \chi_Y, \chi_Y) &= 0.16 \\
\hat{f}(\chi_Y, \chi_R, \chi_R) &= 0.504 \\
\hat{f}(\chi_Y, \chi_R, \chi_Y) &= 0.21 \\
\hat{f}(\chi_Y, \chi_Y, \chi_R) &= 0.1344 \\
\hat{f}(\chi_Y, \chi_Y, \chi_Y) &= 0.126
\end{aligned}$$

In order to compute r_χ the value of f_σ must be computed for each leaf coloring σ . This can be done by looking at all possible edge labelings of the splits network and then determining which edge labelings correspond to which leaf colorings. For the above example there are $2^6 = 64$ possible edge labelings which correspond to $2^3 = 8$ possible leaf colorings. That means there are eight different edge labelings each producing a given σ .

For example, suppose there are eight distinct edge labelings which result in the leaf coloring σ_i . Then f_{σ_i} is equal to the sum of the probabilities of each of the eight edge labelings. To determine the probability that a given edge labeling occurs the probabilities of obtaining each color class' label are multiplied together.

Recall that Theorem 9 states that

$$r_\chi = \sum_{\sigma \in G^{n-1}} \chi(\sigma) f_\sigma$$

for $\chi(\sigma) = \prod_{l \in L \setminus \{R\}} \chi_l(\sigma_l)$. Just as with $\hat{f}(\chi)$ there are eight values of r_χ corresponding to the eight possible χ . Given the values of f_σ , that can be found in example 20 on page 79, the values of r_χ for each χ are

$$\begin{aligned}
r_{(\chi_R, \chi_R, \chi_R)} &= 1 \\
r_{(\chi_R, \chi_R, \chi_Y)} &= 0.216 \\
r_{(\chi_R, \chi_Y, \chi_R)} &= 0.216 \\
r_{(\chi_R, \chi_Y, \chi_Y)} &= 0.16 \\
r_{(\chi_Y, \chi_R, \chi_R)} &= 0.504 \\
r_{(\chi_Y, \chi_R, \chi_Y)} &= 0.21 \\
r_{(\chi_Y, \chi_Y, \chi_R)} &= 0.1344 \\
r_{(\chi_Y, \chi_Y, \chi_Y)} &= 0.126.
\end{aligned}$$

Notice that for each χ the values of $\hat{f}(\chi)$ and r_χ are exactly the same illustrating that $\hat{f}(\chi)$ and r_χ are equal.¹

4.3 A Result

The following section contains a theorem that generalizes a result from Székely et al [60] from phylogenetic trees to splits networks. The generalization is proved using earlier results along with some additional notation.

For $e_C \subset E(N)$ and $0 \neq g \in G$ define $\rho^{e_C, g} \in G^{n-1}$ so that $\rho_l^{e_C, g} = 0$ for $l \notin L_{e_C}$, where l is not the root and $\rho_l^{e_C, g} = g$ for $l \in L_{e_C}$. Given $\rho^{e_C, g}$ let $\mathcal{C}(N) = \{\rho^{e_C, g} : e_C \in E(N), 0 \neq g \in G\}$.

Theorem 10. For $0_{G^{n-1}} \neq \rho \in G^{n-1}$, $\rho \notin \mathcal{C}(N)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = 1;$$

for $\rho = \rho^{e_C, g} \in \mathcal{C}(N)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{h \in \hat{G}} l_{e_C}(h)^{h(g)|G|^{n-2}};$$

and for $\rho = 0_{G^{n-1}}$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_\chi^{\chi(\rho)} = \prod_{e_C \in E(N)} \prod_{h \in \hat{G}} l_{e_C}(h)^{|G|^{n-2}}.$$

Proof. Let i equal the number of color classes in the splits network and $n - 1$ equal the number of non-root leaves. Then the number of possible vectors $\chi \in \hat{G}^{n-1}$ is $|\hat{G}|^{n-1} = m$. Denote these vectors by $\chi_1, \chi_2, \dots, \chi_m$.

Equation (4.3) states that

$$r_\chi = \prod_{e_C \in E(N)} l_{e_C}(\chi_e).$$

¹Actual values of $\hat{f}(\chi)$ and r_χ were computed using software written in G by Jonathan McBee.

Using the notation above I will show that equation (4.3) implies the following:

$$\prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} = \prod_{e_c \subset E(N)} \prod_{h \in \hat{G}} l_{e_c}(h)^{\Sigma\{\chi(\rho): \chi_{e_c}=h\}}.$$

First notice that $r_{\chi} = \prod_{e_c \subset E(N)} l_{e_c}(\chi_{e_c})$ implies

$$\begin{aligned} r_{\chi}^{\chi(\rho)} &= \left(\prod_{e_c \subset E(N)} l_{e_c}(\chi_{e_c}) \right)^{\chi(\rho)} \\ &= l_{e_1}(\chi_{e_1})^{\chi(\rho)} \cdot l_{e_2}(\chi_{e_2})^{\chi(\rho)} \cdot \dots \cdot l_{e_i}(\chi_{e_i})^{\chi(\rho)}. \end{aligned}$$

This in turn implies

$$\begin{aligned} \prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} &= \left(l_{e_1}(\chi_{1_{e_1}})^{\chi_1(\rho)} l_{e_2}(\chi_{1_{e_2}})^{\chi_1(\rho)} \dots l_{e_i}(\chi_{1_{e_i}})^{\chi_1(\rho)} \right) \cdot \\ &\quad \left(l_{e_1}(\chi_{2_{e_1}})^{\chi_2(\rho)} l_{e_2}(\chi_{2_{e_2}})^{\chi_2(\rho)} \dots l_{e_i}(\chi_{2_{e_i}})^{\chi_2(\rho)} \right) \cdot \dots \cdot \\ &\quad \left(l_{e_1}(\chi_{m_{e_1}})^{\chi_m(\rho)} l_{e_2}(\chi_{m_{e_2}})^{\chi_m(\rho)} \dots l_{e_i}(\chi_{m_{e_i}})^{\chi_m(\rho)} \right) \\ &= \left(l_{e_1}(\chi_{1_{e_1}})^{\chi_1(\rho)} l_{e_1}(\chi_{2_{e_1}})^{\chi_2(\rho)} \dots l_{e_1}(\chi_{m_{e_1}})^{\chi_m(\rho)} \right) \cdot \\ &\quad \left(l_{e_2}(\chi_{1_{e_2}})^{\chi_1(\rho)} l_{e_2}(\chi_{2_{e_2}})^{\chi_2(\rho)} \dots l_{e_2}(\chi_{m_{e_2}})^{\chi_m(\rho)} \right) \cdot \dots \cdot \\ &\quad \left(l_{e_i}(\chi_{1_{e_i}})^{\chi_1(\rho)} l_{e_i}(\chi_{2_{e_i}})^{\chi_2(\rho)} \dots l_{e_i}(\chi_{m_{e_i}})^{\chi_m(\rho)} \right) \\ &= \prod_{e_c \subset E(N)} \left(l_{e_c}(\chi_{1_{e_c}})^{\chi_1(\rho)} \cdot l_{e_c}(\chi_{2_{e_c}})^{\chi_2(\rho)} \cdot \dots \cdot l_{e_c}(\chi_{m_{e_c}})^{\chi_m(\rho)} \right). \end{aligned}$$

Consider a term $l_{e_c}(\chi_{e_c})^{\chi(\rho)}$. Notice that $l_{e_c}(\chi_{e_c})^{\chi(\rho)} = l_{e_c}(h)^{\chi(\rho)}$ for $\chi_{e_c} = h$. It is possible for more than one map χ to have the property that $\chi_{e_c} = h$ for some fixed $h \in G$. Therefore the above can be simplified to,

$$\begin{aligned} \prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} &= \prod_{e_c \subset E(N)} l_{e_c}(h_1)^{\Sigma\{\chi(\rho): \chi_{e_c}=h_1\}} l_{e_c}(h_2)^{\Sigma\{\chi(\rho): \chi_{e_c}=h_2\}} \dots l_{e_c}(h_{|G|})^{\Sigma\{\chi(\rho): \chi_{e_c}=h_{|G|}\}} \\ &= \prod_{e_c \subset E(N)} \prod_{h \in \hat{G}} l_{e_c}(h)^{\Sigma\{\chi(\rho): \chi_{e_c}=h\}}. \end{aligned}$$

Therefore

$$\prod_{\chi} r_{\chi}^{\chi(\rho)} = \prod_{e_c \subset E(N)} \prod_{h \in \hat{G}} l_{e_c}(h)^{\sum \{\chi(\rho) : \chi_{e_c} = h\}}.$$

By equations (4.1) and (4.2) along with the fact that $\chi(\rho) = \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l)$

$$\sum \{\chi(\rho) : \chi_{e_c} = h\} = \sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus \{R\}}} \left\{ \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l) : \sum_{l \in L_{e_c}} \chi_l = h \right\}. \quad (4.6)$$

Next it is necessary to show that $\sum \{\chi(\rho) : \chi_{e_c} = h\}$ equals the appropriate value for different ρ .

Case 1: $\rho = 0_{G^{n-1}}$

If $\rho = 0_{G^{n-1}}$, then

$$\sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus R}} \left\{ \prod_{l \in L_{e_c}} \chi_l(\rho_l) : \sum_{l \in L_{e_c}} \chi_l = h \right\} = \sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus R}} \left\{ 1 : \sum_{l \in L_{e_c}} \chi_l = h \right\}.$$

To determine the number of ways $\sum_{l \in L_{e_c}} \chi_l = h$ where h is fixed first let $L_{e_c} = \{l_1\}$ and $\chi = (\chi_1, \chi_2, \dots, \chi_{n-1})$. Then χ_{l_1} is fixed and χ_{l_k} for $k \in \{2, 3, \dots, n-1\}$ can be one of $|G|$ possibilities. Since there are $n-2$ values of χ_k , there are $|G|^{n-2}$ possible χ that satisfy $\sum_{l \in L_{e_c}} \chi_l = h$.

Next let $L_{e_c} = \{l_1, l_2, \dots, l_q\}$ for $1 < q \leq n-1$. Then there are $|G|$ possibilities for each of the $\chi_{l_1}, \chi_{l_2}, \dots, \chi_{l_{q-1}}$ and χ_{l_q} is fixed so that $\sum_{l \in L_{e_c}} \chi_l = h$. The remainder of the χ_{l_k} for $k \in \{q+1, \dots, n-1\}$ may vary amongst the $|G|$ possibilities. Therefore only one of the χ_{l_k} out of $n-1$ terms is fixed. Hence there are $|G|^{n-2}$ possible χ that satisfy $\sum_{l \in L_{e_c}} \chi_l = h$. So

$$\sum_{\substack{\chi_l \in \hat{G}: \\ l \in L \setminus R}} \left\{ 1 : \sum_{l \in L_{e_c}} \chi_l = h \right\} = |G|^{n-2}$$

implies if $\rho = 0_{G^{n-1}}$, then $\prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} = \prod_{e_c \subset E(N)} \prod_{h \in \hat{G}} l_{e_c}(h)^{|G|^{n-2}}$.

Case 2: $\rho = \rho^{e_c, g} \in \mathcal{C}(N)$

If $\rho = \rho^{e_c, g} \in \mathcal{C}(N)$ then

$$\sum_{\substack{\chi_l \in \hat{G} \\ l \in L \setminus R}} \left\{ \prod_{l \in L \setminus R} \chi_l(\rho_l) : \sum_{l \in L_{e_c}} \chi_l = h \right\} = \sum_{\chi_l} \{ \chi(\rho^{e_c, g}) = h(g) \}.$$

It is the case that $\chi(\rho^{e_c, g}) = h(g)$ because

$$\begin{aligned} \chi(\rho^{e_c, g}) &= \prod_{l \in L \setminus R} \chi_l(\rho_l) \\ &= \chi_{l_1}(\rho_{l_1}) \chi_{l_2}(\rho_{l_2}) \cdots \chi_{l_{n-1}}(\rho_{l_{n-1}}) \end{aligned}$$

where each

$$\chi_{l_i}(\rho_{l_i}) = \begin{cases} \chi_{l_i}(g) & \text{if } l_i \in L_{e_c} \\ \chi_{l_i}(0) = 1 & \text{if } l_i \notin L_{e_c}, l_i \neq R. \end{cases}$$

Since $\sum_{l \in L_{e_c}} \chi_l = h$ then $\chi_{l_1} + \chi_{l_2} + \cdots + \chi_{l_w} = h$, which implies $\chi_{l_1}(g) \cdot \chi_{l_2}(g) \cdots \chi_{l_w}(g) = h(g)$ and hence $\chi(\rho^{e_c, g}) = h(g)$. As a result

$$\sum_{\substack{\chi_l \in \hat{G} \\ l \in L \setminus R}} \left\{ \prod_{l \in L \setminus R} \chi_l(\rho_l) : \sum_{l \in L_{e_c}} \chi_l = h \right\} = \sum_{\chi_l} \{ h(g) : \sum_{l \in L \setminus R} \chi_l = h \}$$

and by the same argument as in case 1, there are $|G|^{n-2}$ vectors χ such that $\sum \chi_l = h$. Therefore

$$\sum_{\chi_l} \{ h(g) : \sum \chi_l = h \} = h(g) |G|^{n-2}.$$

So for $\rho = \rho^{e_c, g} \in \mathcal{C}(N)$,

$$\prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} = \prod_{h \in \hat{G}} l_{e_c}(h)^{h(g) |G|^{n-2}}.$$

Case 3: $0_{G^{n-1}} \neq \rho \in G^{n-1}$, $\rho \notin \mathcal{C}(N)$.

Since $\rho \notin \mathcal{C}(N)$ and $\rho \neq 0_{G^{n-1}}$ one of two things must be true. Either

(α) There exists $l \notin L_{e_c}, l \neq R$ and $\rho_l \neq 0_G$.

(β) There exists $l, j \in L_{e_c}$ such that $\rho_l \neq \rho_j$. Notice that if $\rho \in \mathcal{C}(N), \rho_l = \rho_j = g$.

Subcase (α): Take $\eta \in \hat{G}$ such that $\eta(\rho_l) \neq 1$. One must exist because $[\chi(g)]$ is invertible which implies it cannot be an all ones matrix. Change χ from $\chi = (\chi_1, \chi_2, \dots, \chi_{n-1})$ to $\chi = (\chi_1, \chi_2, \dots, \eta + \chi_l, \dots, \chi_{n-1})$.

Next consider equation (4.6),

$$\sum_{\substack{\chi_l \in \hat{G} \\ l \in L \setminus \{R\}}} \left\{ \prod_{l \in L \setminus \{R\}} \chi_l(\rho_l) : \sum_{l \in L_{e_c}} \chi_l = h \right\}$$

with the χ above. On one hand the sum is fixed because letting $\chi = (\chi_1, \dots, \eta + \chi_l, \dots, \chi_{n-1})$ is like permuting the terms in the sum. This is because the sum ranges over all choices of χ_l , and since $\eta + \chi_l$ is just like choosing a different value for a χ_l you get the same sum out. The terms have just been added in a different order.

On the other hand taking $\chi = (\chi_1, \dots, \eta + \chi_l, \dots, \chi_{n-1})$ is like multiplying the sum by $\eta(\rho_l) \neq 1$. This is because $\chi(\rho^{e_c, g}) = \chi_{l_1}(\rho_{l_1}) \dots \eta(\rho_{l_1}) \chi_{l_1}(\rho_{l_1}) \dots \chi_{l_{n-1}}(\rho_{l_{n-1}})$ and each term in the sum has a factor of $\eta(\rho)$ so it can be factored out. Hence the sum must be zero.

Subcase (β): There exists $l, j \in L_{e_c}$ such that $\rho_l \neq \rho_j$. Suppose there exists $l, j \in L_{e_c}$ such that $\rho_l \neq \rho_j$. Take η such that $\eta(\rho_j - \rho_l) = \eta(\rho_j)\eta^{-1}(\rho_l) \neq 1$. Again one must exist since $[\chi(g)]$ is regular.

Replace the vector $\chi = (\chi_1, \dots, \chi_l, \dots, \chi_j, \dots, \chi_{n-1})$ by $\chi = (\chi_1, \dots, \chi_l - \eta, \dots, \chi_j + \eta, \dots, \chi_{n-1})$. Again on one hand the terms of the sum in (4.6) have just been permuted and therefore the sum is fixed.

On the other hand notice that

$$\begin{aligned} \chi(\rho^{e_c, g}) &= \chi_1(\rho_1)\chi_2(\rho_2)\dots\chi_l - \eta(\rho_l)\dots\chi_j + \eta(\rho_j)\dots\chi_{n-1}(\rho_{n-1}) \\ &= \chi_1(\rho_1)\dots\chi_l(\rho_l)\eta^{-1}(\rho_l)\dots\chi_j(\rho_j)\eta(\rho_j)\dots\chi_{n-1}(\rho_{n-1}) \\ &= \eta^{-1}(\rho_l)\eta(\rho_j)(\chi_1(\rho_1)\dots\chi_{n-1}(\rho_{n-1})) \end{aligned}$$

Therefore as in (α) the sum must be zero. So if $0_{G^{n-1}} \neq \rho \in G^{n-1}$, $\rho \notin \mathcal{C}(N)$

$$\sum_{\chi_l} \{ \prod \chi_l(\rho_l) : \sum \chi_l = h \} = 0$$

which implies

$$\begin{aligned} \prod_{\chi \in \hat{G}^{n-1}} r_{\chi}^{\chi(\rho)} &= \prod_{e_c \in E(N)} \prod_{h \in \hat{G}} l_{e_c}(h)^0 \\ &= \prod_{e_c \in E(N)} \prod_{h \in \hat{G}} 1 \\ &= 1. \end{aligned}$$

The three equalities have been shown. □

An alternative formulation of Theorem 10 is given below. Let $K = [h(g)]$ denote the matrix in which rows correspond to $h \in \hat{G}$ and columns correspond to $g \in G$. Let $H = [\chi(\sigma)]$ denote the matrix in which rows correspond to $\chi \in \hat{G}^{n-1}$ and columns correspond to $\sigma \in G^{n-1}$. The vector \mathbf{f} denotes the vector of f_{σ} 's and \mathbf{p}_{e_c} is the vector of $p_{e_c}(g)$'s for every $e_c \subset E(N)$. Finally, the logarithm of a vector is vector of logarithms of the components.

Theorem 11.

$$\mathbf{q}(N) = [H^{-1} \log H \mathbf{f}]_{\rho} = \begin{cases} 0, & \text{if } 0 \neq \rho \notin \mathcal{C}(N), \\ [K^{-1} \log K \mathbf{p}_{e_c}]_h, & \text{if } \rho = \rho^{e_c, h} \in \mathcal{C}(N), \\ \sum_{e_c \in E(N)} \sum_{h \in G} [K^{-1} \log K \mathbf{p}_{e_c}]_h, & \text{if } \rho = 0 \end{cases}$$

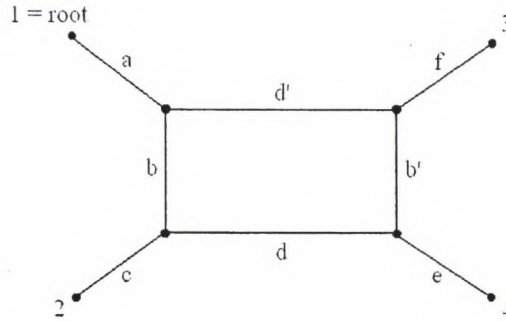
The indexing of the vector \mathbf{q} makes it easy to read off the most likely splits occurring in a splits network. The following examples discuss the indexing as well as Theorem 11.

Example 19. *Suppose the historical relationships between four taxa are being considered by examining four aligned sequences representing the taxa. As in previous examples let each sequence be made up of 2-state characters, purines (R) and pyrimidines (Y). Assume that sequence 1 is the root sequence and that one site of that sequence will be compared to the same site in the*

other three sequences. Notice that the three sites being examined which are not from the root sequence must be in state R or Y . Therefore there are $2^3 = 8$ possibilities for the leaf states at the site being considered. For example the ordered triple (R, R, R) indicates that the site being considered in sequences 2, 3 and 4 are all at state R . The ordered triple (Y, Y, R) indicates that the site being considered in sequences 2 and 3 is at state Y while the site in sequence 4 is at state R . The vectors of length 2^3 are indexed by the eight ordered triples given in alphabetical order. In general the vectors of length 2^{n-1} are indexed by $(n-1)$ -tuples, with entries R and Y in the case of 2-state sequences and entries A, C, G and T in the case of 4-state sequences. The tuples are given in alphabetical order.

The ordered $(n-1)$ -tuples used in indexing vectors have multiple interpretations. The most obvious interpretation is that the $(n-1)$ -tuples correspond to the possible leaf colorations by either R and Y or by A, C, G , and T . A less obvious interpretation of the tuples is that they correspond to a color class. In order to make this correspondence precise recall the notation below.

For $e_C \subset E(N)$ and $0 \neq g \in G$ define $\rho^{e_C, g} \in G^{n-1}$ so that $\rho_l^{e_C, g} = 0$ for $l \notin L_{e_C}$, where l is not the root and $\rho_l^{e_C, g} = g$ for $l \in L_{e_C}$. If the splits network below is assumed and R and Y correspond to elements of \mathbb{Z}_2 with R the identity element, the values of $\rho^{e_C, g}$ are as follows:



$$\begin{array}{lll} \rho^{\{a\}, Y} = (Y, Y, Y) & \rho^{\{b, b'\}, Y} = (Y, R, Y) & \rho^{\{c\}, Y} = (Y, R, R) \\ \rho^{\{d, d'\}, Y} = (R, Y, Y) & \rho^{\{e\}, Y} = (R, R, Y) & \rho^{\{f\}, Y} = (R, Y, R). \end{array}$$

Notice that for each ρ^{e_C} the positions of the non-identity group elements give the leaves in the opposite split half from the root when e_C is removed. Also notice that not all possible triples are listed above. Only the triples corresponding to color classes are given.

The following example is a continuation of the example presented in section 4.2 and illustrates the relationship between \mathbf{f} and \mathbf{q} given in Theorem 11.

Example 20. Given the four leaf splits network from above with edge probabilities given in section 4.2 recall that the vector \mathbf{f} contains the values of f_σ for each σ . For this example \mathbf{f} is equal to

$$\mathbf{f} = \begin{bmatrix} 0.3208 \\ 0.1428 \\ 0.1617 \\ 0.1267 \\ 0.0772 \\ 0.0672 \\ 0.0483 \\ 0.0553 \end{bmatrix}$$

Multiplying the 8×8 Hadamard matrix by \mathbf{f} yields

$$H\mathbf{f} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0.3208 \\ 0.1428 \\ 0.1617 \\ 0.1267 \\ 0.0772 \\ 0.0672 \\ 0.0483 \\ 0.0553 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.216 \\ 0.216 \\ 0.16 \\ 0.504 \\ 0.21 \\ 0.1344 \\ 0.126 \end{bmatrix}$$

Notice that $H\mathbf{f}$ is equal to r_x from section 4.2. Taking the natural log of the entries of $H\mathbf{f}$ and multiplying by the inverse of the Hadamard matrix yields the vector below. Since the entries are expressed as a decimal there is some rounding error resulting from taking the natural log.

$$\mathbf{q} = H^{-1} \log H\mathbf{f} = \begin{bmatrix} -1.4 \\ 0.3466 \\ 0.4581 \\ 0.2554 \\ 0.1783 \\ 0.1116 \\ 0 \\ 0.0527 \end{bmatrix}$$

The entry indexed by (Y, Y, R) represents the split $\{2, 3\}$ since the nonidentity group elements are in the first and second positions of the triple. The fact that this entry is zero implies the split $\{2, 3\}$ does not exist in the splits network being considered. The six positive entries indicate that there are six splits which exist in the splits network while the first entry in the vector is the negative value of the sum of the six positive entries.

The example above is fairly small however it illustrates how probabilities of leaf colorings can be transformed using Hadamard conjugation into a vector containing information about the existence of different splits in a splits network. In larger examples assuming different nucleotide substitution models even more information such as the probabilities of different mutations taking place across edges can be determined.

Although this paper assumes a splits network in order to compute \mathbf{f} , \mathbf{f} can also be computed from nucleotide sequences meaning that given a set of nucleotide sequences representing existent taxa it is possible to produce a splits tree or network representing historical relationships between taxa.

Until this point it has been assumed that there is an underlying group based evolutionary model describing how mutations take place. The Kimura three-substitutions type model whose three types of mutations correspond to the non-identity elements of $\mathbb{Z}_2 \times \mathbb{Z}_2$ was used in many examples as well as the 2-state model corresponding to the group \mathbb{Z}_2 . The next chapter discusses a paper by Bryant which provides another formulation of the equations resulting from Hadamard conjugation. Bryant mentions that a motivation for developing a new formulation is the desire to be able to apply Hadamard conjugation to non-group-based models.

Chapter 5

A DIFFERENT PERSPECTIVE ON HADAMARD CONJUGATION

Hadamard conjugation as it is used in combinatorial phylogenetics was first introduced by Hendy and Penny in 1989 and since then has been derived many different ways giving rise to several proofs of the same result. Although there are many ways to view Hadamard conjugation there are two main approaches to deriving the formula. One approach involves the use of pathsets and the other approach involves a Fourier transform on a finite abelian group. In 1989 Hendy and Penny used pathsets to derive the Hadamard conjugation formula. Their results were later extend by Hendy and Snir in 2008 [33]. In the early 1990s Steel et al (1992) [53], Evans and Speed (1993) [20] and Székely et al (1993) [60] viewed the transform as an example of a Fourier transform on abelian groups.

In 2009 “Hadamard Phylogenetic Methods and the n -taxon Process” by David Bryant [9] used a continuous time Markov chain to describe the evolution of the leaf colorations over time and derives the probability matrix $\mathbf{P}[\tau]$ for this process given a vector τ of branch lengths. $\mathbf{P}[\tau]$ is the exponential $\exp(\mathbf{Q}[\tau])$ of a linear combination of rate matrices for the branches. Bryant claims the formula for Hadamard conjugation falls straight out of the formula for $\mathbf{P}[\tau]$.

5.1 Bryant’s approach

The following subsections outline the approach taken by Bryant in his derivation of the Hadamard transform. Bryant’s approach makes clear the use of the continuous Markov chain and the need to be able to diagonalize the rate matrices.

5.1.1 Models of evolution along one branch

Bryant focuses on two nucleotide substitution models throughout his paper. The first model is a 2-state model known as the binary symmetric model or Neyman model. The binary symmetric model is used to analyze sequences of purines and pyrimidines. There is only one substitution type in this model, the substitution taking a purine to a pyrimidine or a pyrimidine to a purine and this substitution happens with rate 1. Consequently, the instantaneous rate matrix for this model is a two by two matrix with entries of -1 on the diagonal and 1 elsewhere.

The second nucleotide substitution model Bryant uses is the Kimura three-substitution type model. The rate matrices for each of these models is given below.

$$Q^{(Neyman)} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

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

The parameters α , β , and γ are chosen so that $\alpha + \beta + \gamma = 1$.

$P_{ij}(t)$, the probability that the state at the end of a branch of length t is j given that the state at the beginning of the branch is i , is given by the matrix exponential $\exp(Qt)$. In the case of the binary symmetric model

$$P(t) = \exp\left(\begin{bmatrix} -t & t \\ t & -t \end{bmatrix}\right) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2t} & \frac{1}{2} - \frac{1}{2}e^{-2t} \\ \frac{1}{2} - \frac{1}{2}e^{-2t} & \frac{1}{2} + \frac{1}{2}e^{-2t} \end{bmatrix}.$$

5.1.2 The n -taxon process

Bryant works with a continuous time Markov chain whose state space depends on the number of states in the substitution model to describe the evolutionary process along the branches of a tree with n non-root leaves. The state space for the Markov process is the set of vectors assigning a state to every taxon. If the binary symmetric model is assumed, the state space is

made up of 2^n vectors. The state space of the process assuming Kimura three-substitution is made up of 4^n vectors since the Kimura three-substitution type model is a four state model.

Terminology and Notation

The transition probabilities depend on which lineages at a particular time are ancestral to which taxa and at each time point t_0, t_1, \dots the rates of substitution for the n -taxon process will change. As a result of this there are several rate matrices that need to be considered. These matrices are described below.

1. Q is the rate matrix for the underlying substitution process.
2. $Q^{(i)}$ is the rate matrix for the n -taxon process during the time interval $[t_{i-1}, t_i]$.
3. $R^{(e)}$ denotes the rate matrix for the n -taxon process restricted to substitutions occurring on a given branch e .

Branch Rate Matrices

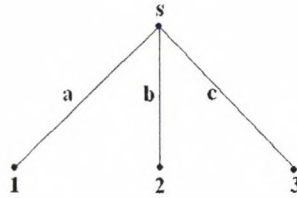
In order to construct a branch rate matrix for branch e only substitutions occurring along branch e are considered. Substitutions occurring at the same time, but along different branches are accounted for in the other branch rate matrices. Also, only taxa that are descendants of a given branch are affected by substitutions occurring along that branch. A taxon is a descendant of a branch if it is a descendant of the ancestors represented by that branch. Let the set A be the set of descendants of the branch being considered.

For simplicity consider creating a branch rate matrix for a tree assuming the binary symmetric evolutionary model. The branch rate matrix is indexed by the 2^n vectors assigning a state to every taxon. In the case of the 3-claw tree the matrix is indexed by the vectors $[00], [01], [10], [11]$, where the components in the vector represent the coloration of the two non-root leaves. For example, consider the vector $[01]$. A 0 in the first component of $[01]$ indicates leaf one is a purine while a 1 in the second component indicates leaf two is a pyrimidine.

By letting $\mathbf{v}[i]$ be the state of the ancestor of taxon i it is possible to define the branch rate matrix for branch e assuming the binary symmetric model.

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

Notice that there are three types of entries in the branch rate matrix. The first type of entry corresponds to substitutions taking place across an edge. In the case of the binary symmetric model substitutions take place at rate 1, hence if $\mathbf{u}[i] \neq \mathbf{v}[i]$ when $i \in A$ the corresponding entry is 1. The second type of entry occurs along the diagonal when $\mathbf{u} = \mathbf{v}$. These entries are set equal to -1 so that row sums are equal to zero. The final entries occurring in the matrix represent substitutions that cannot take place. For instance, consider the 3-claw tree given below.



Let leaf 1 be the root of this tree and consider the branch rate matrix for edge a . If the root is labeled 0 then a substitution taking place across branch a changes both leaves 2 and 3 from 0 to 1. This is because leaves 2 and 3 are both descendants of branch a . Consequently, a substitutions along branch a will never affect only one of the two descendants of a and therefore the entries

$$\mathbf{R}_{[00],[01]}^{(a)} = \mathbf{R}_{[00],[10]}^{(a)} = \mathbf{R}_{[01],[00]}^{(a)} = \mathbf{R}_{[01],[11]}^{(a)} = \mathbf{R}_{[10],[00]}^{(a)} = \mathbf{R}_{[10],[11]}^{(a)} = \mathbf{R}_{[11],[10]}^{(a)} = \mathbf{R}_{[11],[10]}^{(a)} = 0.$$

The branch rate matrices for the 3-claw tree assuming the binary symmetric model are

$$\mathbf{R}^{(a)} = \begin{bmatrix} -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \mathbf{R}^{(b)} = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \mathbf{R}^{(c)} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Bryant states two lemmas which provide a method of constructing the branch rate matrices. The first lemma is specific to the binary symmetric model while the second lemma is stated in terms of the Kimura three-substitutions type model.

Lemma 3. *Let A be the set of taxa that are descendants of the population represented by branch e . Let $E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. For each $i = 1, 2, 3, \dots, n$ set $M^{(i)} = E$ if $i \in A$ and $M^{(i)} = \mathbf{I}$ otherwise. Then*

$$\mathbf{R}^{(e)} = M^{(1)} \otimes M^{(2)} \otimes \dots \otimes M^{(n)} - \mathbf{I}.$$

The matrix E represents the one type of substitution that can take place in the binary symmetric model. In the case of the Kimura three-substitution types model there are three substitution types, α , β , and γ . These three substitutions are represented by the three matrices E_I , E_{II} , and E_{III} , where $Q = \alpha E_I + \beta E_{II} + \gamma E_{III} - (\alpha + \beta + \gamma) \mathbf{I}$.

$$E_I = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad E_{II} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad E_{III} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Bryant states the following lemma for the Kimura three-substitutions type model.

Lemma 4. *Let A be the set of taxa that are descendants of the population represented by branch e . For each $i = 1, 2, \dots, n$ set $M_I^{(i)} = E_I$ if $i \in A$ and $M^{(i)} = I$ otherwise; likewise for $M_{II}^{(i)}$ and $M_{III}^{(i)}$. Then*

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

Diagonalizing a branch rate matrix

Bryant considers the standard group-based models that have an abelian permutation group acting regularly on the set of states ¹. He is able to diagonalize the branch rate matrices using Hadamard matrices. The following lemma and proof are directly from Bryant [9].

¹For more about group-based models see section 6.1 on page 91

Lemma 5. Let $\mathbf{H} = \mathbf{H}^{(n)}$, the n^{th} order Hadamard matrix, and let $\mathbf{R}^{(e)}$ be the rate matrix for the n -taxon process restricted to branch e , binary symmetric case. Let A be the set of taxa that are descendants of branch e . Then

$$\Lambda^{(e)} := \mathbf{H}\mathbf{R}^{(e)}\mathbf{H}^{-1}$$

is a diagonal matrix with

$$\Lambda_{\mathbf{u}\mathbf{u}}^{(e)} = (-1)^{|\{i \in A: \mathbf{u}[i]=1\}|} - 1$$

for all state vectors \mathbf{u} .

Proof. Define matrices $M^{(i)}$ as in Lemma 3. Then

$$\mathbf{H}(\mathbf{R}^{(e)} + \mathbf{I})\mathbf{H}^{-1} = (HM^{(1)}H^{-1}) \otimes \dots \otimes (HM^{(n)}H^{-1}).$$

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

For the diagonal values, note that

$$\Lambda_{\mathbf{u}\mathbf{u}}^{(e)} = \prod_{i=1}^n (HM^{(i)}H^{-1})_{\mathbf{u}[i]\mathbf{u}[i]}$$

and that

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

□

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

$$\exp(\mathbf{R}^{(e)}t) = \mathbf{H}^{-1}\exp(\Lambda^{(e)})\mathbf{H}.$$

Bryant also includes a similar lemma for the Kimura three-substitution types model. Like with the binary symmetric model conjugation by the Hadamard matrix diagonalizes the branch rate matrices. In the case of the Kimura three-substitution model the resulting diagonal matrix has diagonal entries

$$\Lambda_{\mathbf{u},\mathbf{u}}^{(b)} = \alpha(-1)^{|\{i \in A: \mathbf{u}[i]=C \text{ or } T\}|} + \beta(-1)^{|\{i \in A: \mathbf{u}[i]=G \text{ or } T\}|} + \gamma(-1)^{|\{i \in A: \mathbf{u}[i]=C \text{ or } G\}|} - 1.$$

Transition probabilities over multiple lineages

Frequently multiple lineages are evolving independently during a given time interval. The rate matrix for the substitution process over all lineages in each time interval, $\mathbf{Q}^{(i)}$, is the sum of the rate matrices, $\mathbf{R}^{(e)}$, for the individual branches present at that time point. Between time points t_0 and the present time t_k there is a sequence of rate matrices $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \dots, \mathbf{Q}^{(k)}$. Each rate matrix $\mathbf{Q}^{(i)}$ equals the sum of the rate matrices $\mathbf{R}^{(e)}$ for all branches e present during the interval $[t_{i-1}, t_i]$. During each interval, the probability transitions are given by the standard exponential formula

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

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

$$\mathbf{P} = \mathbf{P}^{(1)}\mathbf{P}^{(2)} \dots \mathbf{P}^{(k)}.$$

Since each branch rate matrix is diagonalized by the same Hadamard matrix the sums \mathbf{Q}^i are also diagonalized and the matrices \mathbf{Q}^i must commute. Notice that if two matrices \mathbf{X} and \mathbf{Y} commute that $\exp(\mathbf{X})\exp(\mathbf{Y}) = \exp(\mathbf{X} + \mathbf{Y})$ and therefore

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

Theorem 12. ² Let $\mathbf{P}[\boldsymbol{\tau}]$ be the matrix of transition probabilities in the n -taxon process for the binary symmetric case given a branch length vector $\boldsymbol{\tau}$. Define

$$\mathbf{Q}[\boldsymbol{\tau}] = \sum_e \mathbf{R}^{(e)}\tau_e$$

where e ranges over branches in the tree, $\mathbf{R}^{(e)}$ is the matrix given in Lemma 3, and τ_e is the length of branch e . Then $\mathbf{H}\mathbf{Q}[\boldsymbol{\tau}]\mathbf{H}^{-1}$ is a diagonal matrix and

$$\mathbf{P}[\boldsymbol{\tau}] = \exp(\mathbf{Q}[\boldsymbol{\tau}]).$$

²The same theorem exists for the Kimura three-substitution types model. See Bryant (2009) [9] theorem 14.

Commenting on the above theorem Bryant states the following. The probability distribution for a tree can be recovered from the above equation by noting that at the root, the process is $\mathbf{0} = [0, 0, \dots, 0]$ with probability $\pi_0 = \frac{1}{2}$ and $\mathbf{1} = [1, 1, \dots, 1]$ with probability $\frac{1}{2}$. If \mathbf{u} is the pattern of states at the leaves, then the probability of observing \mathbf{u} equals

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

Using Theorem 12 Bryant derives the Hadamard conjugation formula originally derived by Hendy and Penny in 1989.

Theorem 13. *Suppose that the tree has taxa at the root with state 0. For each non-zero vector \mathbf{u} (indexed by the remaining taxa), let $\mathbf{q}_{\mathbf{u}}$ be the length of the branch with descendants $\{i : \mathbf{u}[i] = 1\}$, if there is such a branch in the tree, and zero otherwise. Let \mathbf{q}_0 be the negative of the sum of all the branch lengths in the tree. Let $\mathbf{p}_{\mathbf{u}}$ be the probability of observing the pattern \mathbf{u} at the leaves. Then*

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

Here the exponential is entry-wise.

Proof. [9] Let $\mathbf{0}$ denote the vector $[0, 0, \dots, 0]$. Both \mathbf{P} and \mathbf{Q} are indexed by vectors of states: By $\mathbf{0}$ -row or $\mathbf{0}$ -column, we mean the row or column with index 0. We seek the probabilities $\mathbf{p}_{\mathbf{u}} = \mathbf{P}_{\mathbf{0}\mathbf{u}}[\tau]$. As $\mathbf{P}[\tau]$ is symmetric, $\mathbf{P}_{\mathbf{0}\mathbf{u}}[\tau] = \mathbf{P}_{\mathbf{u}\mathbf{0}}[\tau]$.

The vector \mathbf{q} is the $\mathbf{0}$ -column of $\mathbf{Q}[\tau]$. Let $\Lambda = \mathbf{H}\mathbf{Q}[\tau]\mathbf{H}^{-1}$, so that $\mathbf{H}\mathbf{Q} = \Lambda\mathbf{H}$. The $\mathbf{0}$ -column of \mathbf{H} is all ones, so the $\mathbf{0}$ -column of $\Lambda\mathbf{H}$ is made up of the diagonal entries of Λ . Hence, the entries on $\mathbf{H}\mathbf{q}$ are the entries along the diagonal of Λ . Taking entry-wise exponentials, we have that $\exp(\mathbf{H}\mathbf{q})$ equals the entries along the diagonal of $\exp(\Lambda)$ and so $\exp(\mathbf{H}\mathbf{q})$ is the first column of $\exp(\Lambda)\mathbf{H}$. Hence, $\mathbf{H}^{-1}\exp(\mathbf{H}\mathbf{q})$ is the $\mathbf{0}$ -column of $\mathbf{H}^{-1}\exp(\Lambda)\mathbf{H}$, which by Theorem 12 equals \mathbf{P} . □

5.1.3 Example

Below an example using the approach outlined in Bryant's 2009 paper [9].

Jukes-Cantor model using Bryant 2009

Let

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}.$$

If the probability any type of substitution takes place along a given edge is $p = 0.01$ then since $P = \exp(Qt)$, $0.01 = -\frac{1}{4}e^{-4\alpha t} + \frac{1}{4}$. This implies $\alpha t \approx 0.010205$. αt is equal to the number of one type of mutation taking place over some time t . Since there are three mutation types in the general group-based model the total number of mutations along a branch over time t is equal to $3 \alpha t \approx 0.030615$.

The matrices representing the possible substitutions are:

$$E1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad E2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad E3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

If the 3-claw is assumed then the branch rate matrices are:

$$\mathbf{R}^{(a)} = \alpha (E1 \otimes E1) + \alpha (E2 \otimes E2) + \alpha (E3 \otimes E3) - \mathbf{I}$$

$$\mathbf{R}^{(b)} = \alpha (I \otimes E1) + \alpha (I \otimes E2) + \alpha (I \otimes E3) - \mathbf{I}$$

$$\mathbf{R}^{(c)} = \alpha (E1 \otimes I) + \alpha (E2 \otimes I) + \alpha (E3 \otimes I) - \mathbf{I}.$$

Each of these matrices have a -1 on the diagonal and three additional nonzero entries in each row. Notice that the row sums for each matrix are always zero. Each of these matrices can also be diagonalized by the Hadamard matrix. Diagonalizing each matrix, adding them together, and multiplying by $\tau \approx 0.030615$ yields $Q[\tau]$ as defined in Theorem 12. Exponentiating $Q[\tau]$ produces a diagonal matrix Bryant refers to as $\mathbf{P}[\tau]$ with the following entries on the diagonal.

$$\mathbf{v} = \begin{bmatrix} 1 \\ 0.92160 \\ 0.92160 \\ 0.92160 \\ 0.92160 \\ 0.92160 \\ 0.88473 \\ 0.88473 \\ 0.92160 \\ 0.88473 \\ 0.92160 \\ 0.88473 \\ 0.92160 \\ 0.88473 \\ 0.88473 \\ 0.92160 \end{bmatrix}$$

Multiplying \mathbf{v} by the 16×16 inverse Hadamard matrix produces the leaf coloration probability vector.

$$\mathbf{H}^{-1}\mathbf{v} = \mathbf{f} = \begin{bmatrix} 0.912676 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.009508 \\ 0.000292 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.009508 \\ 0.000292 \\ 0.000292 \\ 0.009508 \end{bmatrix}$$

Notice that this vector is equal to the vector obtained in section 3.2.4 on page 61.

If the 3-claw tree being considered had different edge lengths the above computations would be very similar. The only difference would be that instead of having a single edge length value τ , there would be a value τ_i for each edge i in the tree.

Chapter 6

EXTENDING HADAMARD CONJUGATION

One of the major limitations of Hadamard conjugation is its restriction to group-based substitution models. Currently the literature regarding the use of Hadamard conjugation assumes any substitution model being used is a group-based model or a submodel of a group-based model. Bryant states in [9] that if the substitution model is assumed to be time-reversible then the only three-parameter group-based nucleotide substitution model is the Kimura three-substitution type model and the only nucleotide substitution models that are available to use with Hadamard conjugation are special cases of the Kimura three-substitution type model. Therefore although there are 203 different time-reversible nucleotide substitution models only a handful are currently used with Hadamard conjugation [36].

In this chapter I will examine what it means to be a group-based model and look at the possibility of extending Hadamard conjugation beyond the Kimura three-substitution type model and submodels of Kimura three-substitution type model. After first focusing on nucleotide based evolutionary models I will look at the possibility of extending Hadamard conjugation to other types of evolutionary models.

6.1 Group-based substitution models

Current literature on Hadamard conjugation assumes that the technique is restricted to group-based substitution models, such as the Kimura three-substitution type model. Although the Kimura three-substitution type model was first published in 1981, the relationship between

the Klein four group and the model was not recognized until the early 1990s. One of the first papers recognizing the group structure of the Kimura three-substitution type model was the 1993 paper by Steven Evans and T.P. Speed [20]. In this paper they described 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, $\mathbb{Z}_2 \times \mathbb{Z}_2$. One possible isomorphism is given by $A \leftrightarrow (0,0)$, $C \leftrightarrow (0,1)$, $G \leftrightarrow (1,0)$ and $T \leftrightarrow (1,1)$. Each of the substitution types can also be associated to a group element by associating to each substitution type the difference between the group element associated to the starting nucleotide and the group element of the ending nucleotide. For example if γ is the substitution which takes A to C , C to A , G to T and T to G , then using the isomorphism above, γ corresponds to the group element $(0,0) - (0,1) = (1,0) - (1,1) = (0,1)$. The substitution types α , β and γ along with the identity substitution, ϵ , produce a group under composition which acts on the nucleotide set $\{A, C, G, T\}$. From this perspective the fact that the Kimura three-substitution type model is abelian group-based means that there exists a permutation group acting regularly on the four bases.

Bryant gives the following equivalent definition of a group-based model in his chapter of Lior Pachter and Bernd Sturmfels text [8].

Definition 25. *A mutation model on state space $\{1, 2, \dots, r\}$ is said to be a group-based model if there exists an abelian group G with elements g_1, g_2, \dots, g_r and a function $\psi : G \rightarrow \mathbb{R}$ such that the instantaneous rate matrix Q satisfies*

$$Q_{ij} = \psi(g_i - g_j)$$

for all i and j .

The definition implies that there is a regular abelian group of automorphisms, defined to be permutations of the bases such that $Q_{\sigma(i),\sigma(j)} = Q_{ij}$, acting on the four bases.

Theorem 14. *Let G be an abelian group with $G = \{g_1, g_2, \dots, g_r\}$. A model is group-based as defined by Bryant if and only if there exists an abelian group, A , of permutations of $\{1, 2, \dots, r\}$ that are automorphisms of the model and A acts regularly on $\{1, 2, \dots, r\}$.*

Proof. Let G be an abelian group with $G = \{g_1, g_2, \dots, g_r\}$. Consider a group-based model as defined by Bryant. Notice that there is a one-to-one correspondence between the states $\{1, 2, \dots, r\}$ and the group elements $\{g_1, g_2, \dots, g_r\}$. Writing the correspondence formally causes the notation to become overly complicated, so to simplify the notation identify state i with group element g_i for $i \in \{1, 2, \dots, r\}$.

By Bryant's definition the instantaneous rate matrix Q satisfies $Q_{ij} = \Psi(g_i - g_j)$ for all i and j . Let $\pi_i : G \rightarrow G$ be the permutation $x \mapsto x + g_i$ for all $x \in G$. Given this map

$$\begin{aligned} Q_{\pi_k(i)\pi_k(j)} &= \Psi((g_i + g_k) - (g_j + g_k)) \\ &= \Psi(g_i - g_j) \\ &= Q_{ij}, \end{aligned}$$

where in the second to last equation we use the fact that G is abelian. So $\pi : G \rightarrow S_r$, $g_i \mapsto \pi_i$ for all i is a homomorphism, one-to-one and $Q_{\pi(g)(i),\pi(g)(j)} = Q_{ij}$ for all i and j . Therefore $\pi(G)$ is a group of automorphisms of the model which are defined to be permutations of $\{1, 2, \dots, r\}$ such that $Q_{\sigma(i),\sigma(j)} = Q_{ij}$ for all i and j . Hence $\pi(G)$ is abelian and via π , G acts regularly on the set $\{1, 2, \dots, r\}$ of states.

Next suppose that there exists an abelian group, A , of permutations of $\{1, 2, \dots, r\}$ that are automorphisms of the model and let A act regularly on $\{1, 2, \dots, r\}$. Then for all $\sigma \in A$, $Q_{\sigma(i),\sigma(j)} = Q_{i,j}$. Label state m with the identity element $e \in A$. Next identify the states with the elements of A so that $a \in A$ corresponds to $a(m)$. Let i, j be states labeled with $a_i, a_j \in A$

and i', j' be states labeled with $a_{i'}, a_{j'} \in A$ such that $a_j \circ a_i^{-1} = a_{j'} \circ a_{i'}^{-1}$. This implies that $a_{i'} \circ a_i^{-1} = a_{j'} \circ a_j^{-1}$ since A is abelian. Next let $b = a_{i'} \circ a_i^{-1}$. Then

$$\begin{aligned} b(i) &= a_{i'}(a_i^{-1}(i)) \\ &= a_{i'}(a_i^{-1}(a_i(m))) \\ &= a_{i'}(m) \\ &= i' \end{aligned}$$

and

$$\begin{aligned} b(j) &= a_{i'}(a_i^{-1}(j)) \\ &= a_{i'}(a_i^{-1}(a_j(m))) \\ &= a_{j'}(a_j^{-1}(a_j(m))) \\ &= a_{j'}(m) \\ &= j'. \end{aligned}$$

So $Q_{ij} = Q_{i'j'}$. Therefore there exists ψ such that $Q_{ij} = \psi(a_i - a_j)$ for all i and j . Hence the model is group-based by Bryant's definition. \square

Recall that the Kimura two-substitution type model is a submodel of the Kimura three-substitution type model and has rate matrix given by

$$\mathbf{Q}^{(K2ST)} = \begin{bmatrix} -K & \gamma & \alpha & \gamma \\ \gamma & -K & \gamma & \alpha \\ \alpha & \gamma & -K & \gamma \\ \gamma & \alpha & \gamma & -K \end{bmatrix}$$

where $K = \alpha + 2\gamma$. The Kimura two-substitution type model and other submodels of Kimura three-substitution type model can be obtained from the Kimura three-substitution type model by setting parameters equal to each other. Unlike the Kimura three-substitution type model, in the Kimura two-substitution type model there does not exist a permutation group acting regularly

on the four bases. Therefore the Kimura two-substitution type model is not group-based in the same sense as the Kimura three-substitution type model. From this point of view Hadamard conjugation applies to group-based models as well as submodels of group-based models.

6.2 Simultaneous Diagonalization

In hopes of expanding the number of evolutionary models that can be used with Hadamard conjugation it is necessary to consider properties of the models that make the implementation of Hadamard conjugation possible. Since 1989 there have been several different derivations all leading to the invertible formulas known as Hadamard conjugation. The utility of Hadamard conjugation comes from the fact that it provides an analytic formula relating observed pattern frequencies from DNA data to a vector containing information about the structure of the phylogenetic tree. In general analytic formulas which relate this information do not exist.

In Bryant's derivation of the Hadamard conjugation formula given in his 2009 paper [9], which was discussed in chapter 5, he relies on the fact that the n -taxon process is a continuous time Markov chain that describes the evolution of a vector of ancestral states. He derives the probability matrix $\mathbf{P}[\tau]$ as being the exponential $\exp(\mathbf{Q}[\tau])$. The vector \mathbf{q} as given in Hendy and Penny's derivation of the of Hadamard conjugation formula is the first column of $\mathbf{Q}[\tau]$.

In deriving the equation $\mathbf{P}[\tau] = \exp(\mathbf{Q}[\tau])$ Bryant relies on the fact that the rate matrices, \mathbf{Q} , are simultaneously diagonalizable. A quote from Bryant's proof illustrating this fact is given below.

“Now comes a key step in the proof. The rate matrices $\mathbf{R}^{(b)}$ down each branch are all diagonalized by the Hadamard matrix \mathbf{H} and, therefore, so are the sums $\mathbf{Q}^{(i)}$. Since every matrix $\mathbf{Q}^{(i)}$ in the product is diagonalized by the same matrix \mathbf{H} , the rate matrices $\mathbf{Q}^{(i)}$ all commute. If two matrices \mathbf{X} and \mathbf{Y} commute, then $\exp(\mathbf{X})\exp(\mathbf{Y}) = \exp(\mathbf{X} + \mathbf{Y})$ [9].”

He goes on to say that applying this identity gives the formula.

The existence of the matrix exponential in the proof requires that the rate matrices be simultaneously diagonalizable. Recall from section 1.1.7 on page 16 that the matrix exponential of $A \in \mathbb{F}^{n \times n}$ is defined as the matrix

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

If $A = \text{diag}(A_1, \dots, A_k)$, where $A_i \in \mathbb{F}^{n_i \times n_i}$ for all $i = 1, \dots, k$, then $e^A = \text{diag}(e^{A_1}, \dots, e^{A_k})$. However, if A is not a diagonal matrix the matrix exponential produces a matrix with entries that will not correspond to the Hadamard conjugation formula given in Bryant 2009 [9] or Székely et al [60].

As well as needing simultaneous diagonalization to obtain the correct formulas, simultaneous diagonalization is also needed to reduce the computational complexity of calculating the matrix exponential. The computational complexity becomes even more significant when analyzing evolutionary models with more than four states. This was noted in a paper by a 2007 paper by Mayrose, Doron-Faigenboim, Bacharach and Pupko in their discussion of the use of codon models [46]. They state, “These models tend to be computationally intensive as they involved exponentiating a large (61 by 61) rate matrix.”

The ability to simultaneously diagonalize rate matrices appears to be of significant importance and so will be contained in the assumptions made in later sections.

6.3 Substitution Models and Groups

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 $\mathbb{R}V$. The first result of this section illustrates this shift.

Before stating the first result an explanation of terminology must be given. Let 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 be known as a Q space. For example, given the Kimura three-substitution type model whose rate matrix is give below

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

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

Theorem 15. *If there exists a real invertible four by four matrix X that simultaneously diagonalizes A , a three parameter Q space, 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 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 . \square

Just as the Kimura two-substitution type model is a submodel of the Kimura three-substitution type model and can be used with Hadamard conjugation, submodels of M_{37} can also be used with Hadamard conjugation.

Corollary 3. *Every subspace of a nucleotide substitution model in which all 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 15. To see this consider a three parameter Q algebra so that theorem 15 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 Kimura three-substitution type model and submodels. As the following example will show, it allows the use of additional models with Hadamard conjugation.

6.3.1 An example with model M_{37}

Consider the rate matrix for the M_{37} nucleotide substitution model with diagonal entries chosen so that the row sums are equal to zero.

$$Q^{(M_{37})} = \begin{bmatrix} - & \alpha & \beta & \beta \\ \alpha & - & \beta & \beta \\ \beta & \beta & - & \eta \\ \beta & \beta & \eta & - \end{bmatrix}$$

Next suppose the M_{37} substitutions take place with the following probabilities:

α type substitution with probability 0.01

β type substitution with probability 0.02

η type substitution with probability 0.03

Then since $P = \exp(Qt)$ for a transition probability matrix of the form

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

the following equations must hold.

$$\begin{aligned} p_0 &= \frac{1}{4} + \frac{1}{4}e^{-4t\beta} + \frac{1}{2}e^{-2t\beta-2t\alpha} \\ p_1 &= \frac{1}{4} + \frac{1}{4}e^{-4t\beta} - \frac{1}{2}e^{-2t\beta-2t\alpha} \\ p_2 &= \frac{1}{4} - \frac{1}{4}e^{-4t\beta} \\ p_3 &= \frac{1}{4} + \frac{1}{4}e^{-4t\beta} + \frac{1}{2}e^{-2t\beta-2t\eta} \\ p_4 &= \frac{1}{4} + \frac{1}{4}e^{-4t\beta} - \frac{1}{2}e^{-2t\beta-2t\eta} \end{aligned}$$

These equations imply that

$$\alpha t \approx 0.0100922996$$

$$\beta t \approx 0.0208454022$$

$$\eta t \approx 0.0318348556.$$

If the average of the off diagonal row sums of Q are equal to 1, then $\frac{\alpha+2\beta+\eta+2\beta}{2} = 1$ which implies

$$t \approx 0.062654382.$$

Therefore

$$\alpha \approx 0.1610789106$$

$$\beta \approx 0.3327046175$$

$$\eta \approx 0.5081026192.$$

The next step is to convert the M_{37} model into the form of the Kimura three-substitution types model. This will be done by conjugating by a specific matrix. To do this consider the M_{37} Q algebra, $M = \{Q^{(M_{37})} + \delta I | \alpha, \beta, \eta, \delta \in \mathbb{R}\}$ and 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 algebra¹. Therefore there is an isomorphism between the M algebra and the group algebra $\mathbb{R}V$. Notice that

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

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

Conjugating by the matrix YH yields a matrix with the form of the Kimura three-substitution type rate matrix.

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

If the parameters of the Kimura three-substitution type model are κ , λ , and ω then the equations relating α , β , and η to κ , λ , and ω are given by

¹ 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.

$$\begin{aligned}
\kappa &= \frac{1}{2}\eta + \frac{1}{2}\alpha \\
\lambda &= -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta \\
\omega &= \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta \\
-(\kappa + \lambda + \omega) &= -\frac{1}{2}\alpha - 2\beta - \frac{1}{2}\eta.
\end{aligned}$$

This implies that

$$\begin{aligned}
\kappa t &= \frac{1}{2}\eta + \frac{1}{2}\alpha \approx 0.020963 \\
\lambda t &= -\frac{1}{2}\alpha + \beta + \frac{1}{2}\eta \approx 0.017391 \\
\omega t &= \frac{1}{2}\alpha + \beta - \frac{1}{2}\eta \approx 0.009974
\end{aligned}$$

and

$$Q^{(K3ST)} = \begin{bmatrix}
-(\kappa + \lambda + \omega) & \kappa & \lambda & \omega \\
\kappa & -(\kappa + \lambda + \omega) & \omega & \lambda \\
\lambda & \omega & -(\kappa + \lambda + \omega) & \kappa \\
\omega & \lambda & \kappa & -(\kappa + \lambda + \omega)
\end{bmatrix}.$$

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

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

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

AA	0.83062
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	0.00896

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

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

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 15 and corollary 3 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 so the full force of Hadamard conjugation. 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.

6.4 Algebras and Association Schemes

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 using Bryant's n -taxon process [9]. The question remaining however, is how to find an appropriate commutative algebra. One way is to use an association scheme.

Recall from section 1.1.11 that if \mathcal{A} is the linear span over the reals of the adjacency matrices A_0, \dots, 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. The following theorem given in C.D. Godsil's text [23] provides a method of determining whether a set of rate matrices form an association scheme.

Theorem 16. [23] *Given a finite dimensional vector space, M , of real symmetric matrices containing I, J and closed under both regular matrix and Schur multiplication then M has a unique basis of orthogonal Schur idempotents and these matrices form an association scheme.*

Given this it is possible to look at the form of 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 17. *A time-reversible s -parameter nucleotide substitution model with rate matrix Q such that all entries Q_{ii} are equal for $1 \leq i \leq 4$ and $Q_{ii} \neq Q_{ij}$ for $i \neq j$, corresponds to an association scheme with s associate classes on a set $\mathfrak{X} = \{A, C, G, T\}$.*

Proof. Consider the undirected complete graph on four vertices, K_4 , with vertices labeled A, C, G , and T . Since we are considering an s -parameter model for $1 \leq s \leq 3$, each Q_{ij} for $i \neq j$ is equal to one of s rates $\{\theta_1, \dots, \theta_s\}$ and $Q_{ii} = -(\theta_1 + \dots + \theta_s)$.

Let $E(i, j)$ represent the edge from i to j in the K_4 . Color the edge of the complete graph so that $E(i, j)$ is colored θ_n if $Q_{ij} = \theta_n$.

Notice that the rates correspond to the set of colors on the edges of K_4 . Since the nucleotide model is an s -parameter model, the graph K_4 is colored so that each of the s colors is used at least once.

Since the diagonal entries of the Q matrix are equal, the sum of the off diagonal entries in each row must be equal. This implies that each vertex $x \in \mathfrak{X}$ is incident to the same number of edges of color $\theta_1, \dots, \theta_s$. Therefore there are integers a_{θ_i} for $i \in \{1, \dots, s\}$ such that each vertex is contained in exactly a_{θ_i} edges of color θ_i .

Next, choose two vertices in K_4 , say x and y , which are connected by an edge of color θ_k .

Case 1: Let $s = 3$. Since $s = 3$ each vertex is incident to exactly one edge of each color. Also, in K_4 there are exactly two paths of length two from vertex x to vertex y . Since each vertex is incident to one edge of each color, one path will consist of the edge colored θ_i followed by the edge colored θ_j and the other path will consist of the edge colored θ_j followed by the edge colored θ_i . Therefore whenever (x, y) is an edge of color θ_k

$$|\{z \in \mathfrak{X} : (x, z) \text{ has color } \theta_i \text{ and } (z, y) \text{ has color } \theta_j\}| = 1 = p_{\theta_i, \theta_j}^{\theta_k}$$

Case 2: Let $s = 2$. If $s = 2$, each vertex is incident to one edge of color θ_m and two edges of color θ_n . If edge (x, y) is colored θ_m then in K_4 there are two paths of length two from x to y such that each edge is colored θ_n . If (x, y) is colored θ_n then there is one path of length two from x to y with an edge colored θ_m followed by an edge colored θ_n and there is one path of length two from x to y with an edge colored θ_n followed by an edge colored θ_m . Therefore there exists an integer $p_{\theta_i, \theta_j}^{\theta_k}$ such that whenever (x, y) is an edge of color θ_k

$$|\{z \in \mathfrak{X} : (x, z) \text{ has color } \theta_i \text{ and } (z, y) \text{ has color } \theta_j\}| = p_{\theta_i, \theta_j}^{\theta_k}$$

Case 3: Let $s = 1$. If $s = 1$ then all edges of K_4 are colored the same color. That implies that for any edge (x, y) colored θ_k there will be exactly two paths of length two from x to y such that each edge in the path is colored θ_k . Therefore there exists an integer $p_{\theta_i, \theta_j}^{\theta_k}$ such that whenever (x, y) is an edge of color θ_k

$$|\{z \in \mathfrak{X} : (x, z) \text{ has color } \theta_i \text{ and } (z, y) \text{ has color } \theta_j\}| = p_{\theta_i, \theta_j}^{\theta_k} = 2$$

□

The above result implies that the Kimura two-substitution type and Kimura three-substitution type models must correspond to association schemes on four points. The following two examples illustrate the correspondence between the models and the association schemes.

Example 21. Consider the Kimura two-substitution types nucleotide substitution model given by the rate matrix

$$Q^{(K2ST)} = \begin{bmatrix} -K & \beta & \alpha & \beta \\ \beta & -K & \beta & \alpha \\ \alpha & \beta & -K & \beta \\ \beta & \alpha & \beta & -K \end{bmatrix}$$

where $K = \alpha + 2\beta$. The Kimura two-substitution types model corresponds to the association scheme $\{A_0, A_1, A_2\}$ where

$$A_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Notice that $A_0 + A_1 + A_2 = J$, $A_i = A_i^T$ for $0 \leq i \leq 2$ and for any i, j the product $A_i A_j$ is a linear combination of A_0, A_1, A_2 .

Example 22. Consider the Kimura three-substitution types nucleotide substitution model given by the rate matrix

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

where $K = \alpha + \beta + \gamma$. The Kimura three-substitution types model corresponds to the association scheme $\{A_0, A_1, A_2, A_3\}$ where

$$A_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Notice that $A_0 + A_1 + A_2 + A_3 = J$, $A_i = A_i^T$ for $0 \leq i \leq 3$ and for any i, j the product $A_i A_j$ is a linear combination of A_0, A_1, A_2, A_3 .

In the case of nucleotide substitution models the existence of an association scheme implies that the model is a submodel of the Kimura three-substitution type model.

Theorem 18. *If the rate matrix Q of a time-reversible s -parameter nucleotide substitution model corresponds to an association scheme with s associate classes on a set $\mathfrak{X} = \{A, C, G, T\}$, then the nucleotide substitution model is a submodel of the Kimura three-substitution type model.*

Proof. Consider the definition of an association scheme that involves the coloring of the edges of a complete undirected graph with vertex set $\{A, C, G, T\}$.

Case 1: The nucleotide substitution model is a one-parameter model. There is exactly one coloring using a single color of the complete graph on four vertices. This association scheme corresponds to the Jukes-Cantor nucleotide substitution model.

Case 2: The nucleotide substitution model is a two-parameter model. By the definition of an association scheme the complete graph on four vertices must be colored using two colors. Since there are six edges in the graph and there are integers a_i for i in $\{1, \dots, s\}$ such that each vertex is contained in exactly a_i edges color i , each vertex in K_4 must be contained in two edges of color a and one edge of color b . This implies there are two disjoint edges of color b in the graph. The remaining edges are colored a . If the edges colored b are (AG) and (CT) then the model corresponds to the Kimura two-substitution type model. If different edges are colored b then the nucleotide substitution model could be obtained by setting the appropriate parameters in the Kimura three-substitution type model equal to each other.

Case 3: The nucleotide substitution model is a three-parameter model. In this case K_4 is colored by three colors and each vertex must be contained in exactly one edge of each color. This forces a specific coloration of the graph corresponds to the Kimura three-substitution type model. □

6.5 Larger evolutionary models

Thus far the focus of this chapter has been on time-reversible nucleotide substitution models. As mentioned in chapter 2 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 become 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 following result fails for models that have a number of states not equal to a power of two. This explains why amino acid and codon models, except for those on 64 states, differ from nucleotide substitution models. Additionally, all results in this section assume the evolutionary models being considered are time-reversible. This implies all rate matrices are symmetric.

Theorem 19. *If there exists a real invertible $2^n \times 2^n$ matrix X that simultaneously diagonalizes A , a $2^n - 1$ parameter Q space, then $A \cong \mathbb{R}V$ where V is an elementary abelian group of order 2^n .*

Proof. Let D be the set of all $2^n \times 2^n$ diagonal matrices and A be a $2^n - 1$ parameter Q algebra. $XAX^{-1} \subseteq D$, 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 2^n dimensional algebra which is diagonalizable, $\mathbb{R}V \cong D$. Therefore $A \cong \mathbb{R}V$. To see that V must be elementary abelian notice that the abelian group must have a group algebra diagonalizable over the real numbers. Therefore the diagonal entries must only involve real roots of unity. This implies that the group must have exponent 2, which in turn implies that the group is elementary abelian. \square

Corollary 4. *Every 2^n state evolutionary model in which all Q matrices are simultaneously diagonalizable is a submodel of a group algebra where the group is an elementary abelian group of order 2^n .*

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

Moving to complex scalars, a similar result holds.

Theorem 20. *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. $X\bar{A}X^{-1} \subseteq D$, 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 $\bar{A} \cong \mathbb{C}V$. \square

Corollary 5. *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 $\mathbb{C}V$. \square

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. Although in the case of nucleotide substitution models using an association scheme only produces the Kimura three-substitution model and submodels of the Kimura three-substitution model, 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 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.

Below are a few examples of association schemes on twenty through twenty-two and sixty-one through sixty-four points. Recall from section 1.2.2 on page 29 that there is some uncertainty regarding the appropriate number of states to consider in amino acid and codon evolutionary models. For references which include a list of known association schemes and distance regular graphs see the *CRC Handbook of Combinatorial designs* by Colbourn and Dinitz [17] and *Distance-Regular Graphs* by Brouwer, Cohen and Neumaier [7].

Association Scheme on 20 points

Recall the Johnson scheme which was discussed in section 1.1.10 on page 23. The Johnson scheme $J(6, 3)$ is an association scheme on twenty points with three associate classes [7].

Association Scheme on 21 points

The Johnson scheme $J(7, 2)$ is an association scheme on twenty-one points with two associate classes [7].

Association Scheme on 22 points

Consider the field $GF(11) = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and the set of squares in $GF(11)$ $D = \{t^2 | t \in GF(11)^*\} = \{1, 3, 4, 5, 9\}$. D is a difference set. A (v, k, λ) *difference set* is a subset D of a group G such that the order of G is v , the size of D is k , and every non-identity element of G can be expressed as a product $d_1 d_2^{-1}$ of elements of D in exactly λ ways. Given the difference set D and $g \in G$, $gD = \{gd : d \in D\}$ is also a difference set, and is called a translate of D . The set of all translates of D forms a symmetric design. The design contains v points and v blocks. Each block consists of k points and each point is contained in k blocks. Any two blocks have λ elements in common and any two points are joined by λ blocks.

The 2-(11,5,2) design corresponds to an incidence graph. This graph has one vertex for each point and each block. For this example that implies the incidence graph has $11 + 11 = 22$ vertices. There also exists one edge for every incidence between a point and a block. This graph is distance regular and corresponds to an association scheme on twenty-two points [7].

Association Scheme on 61 points

There exists an association scheme on sixty-one points that corresponds to a Paley graph of order sixty-one. If q is a prime power with $q \equiv 1 \pmod{4}$ then the Paley graph $P(q)$ is a graph with vertices labeled by the elements of the finite field $GF(q)$ and edges between two vertices if the difference between the vertices is a square in $GF(q)$. The Paley graph is strongly regular with parameters $srg(q, \frac{1}{2}(q-1), \frac{1}{4}(q-5), \frac{1}{4}(q-1))$ [61].

Association Schemes on 62, 63, and 64 points

See [17] for examples.

Each of the above association schemes correspond to an amino acid or codon model of evolution. The examples produce a twenty amino acid model with three parameters, a twenty-one amino acid model with two parameters, and a twenty-two amino acid model with two parameters. There also exists a sixty-one codon model with two parameters, a sixty-two codon model with three parameters a sixty-three codon model with three parameters, and a sixty-four codon model with two or three parameters.

6.6 Networks and Hadamard conjugation

As can be seen from above, Hadamard conjugation can be applied to a number of models beyond the Kimura three-substitution type model and submodels. In the case of a nucleotide substitution model like M_{37} it is possible to use the isomorphism between it and the Kimura three-substitution type model to apply the results from Székley et al [60]. This implies that the extension of Székley's results to splits networks also applies to the M_{37} model. Since Hadamard

conjugation can also exist for larger evolutionary models splits networks can be used in these settings as well.

6.7 Strand Symmetric Models and Hadamard Conjugation

Recall from section 2.1.9 on page 41 that Strand symmetric substitution implies that both strands of the genome segment undergo any given type of substitution at the same rate, hence complementary substitution rates are equal. This implies that when there are no biases in mutation there are six substitution types.

$$\begin{aligned}
 a : A &\leftrightarrow T \\
 b : A &\rightarrow G, & T &\rightarrow C \\
 c : G &\rightarrow A, & C &\rightarrow T \\
 d : G &\rightarrow T, & C &\rightarrow A \\
 e : A &\rightarrow C, & T &\rightarrow G \\
 f : G &\leftrightarrow C.
 \end{aligned}$$

The rate matrix for the most general strand symmetric model is give below. Recall that the rows and columns are indexed by A, C, G and T respectively.

$$Q^{(SSM)} = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 \\ q_4 & q_5 & q_6 & q_7 \\ q_7 & q_6 & q_5 & q_4 \\ q_3 & q_2 & q_1 & q_0 \end{bmatrix}$$

Looking at the eight matrices obtained by defining matrix $M_n, 0 \leq n \leq 7$ to be a four by four matrix with entry $m_{ij} = 1$ if $Q_{ij}^{(SSM)} = q_n$ and 0 otherwise it simple to check that M_0, M_1, \dots, M_7 do not commute. Therefore by theorem 5, it is not possible to simultaneously diagonalize the above matrices. Consequently unlike with the three-parameter models analyzed earlier it is not possible to diagonalize $Q^{(SSM)}$ for all possible rates. In the case of the general strand symmetric model it is only possible to block diagonalize $Q^{(SSM)}$.

Notice that the matrix

$$M = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$$

block diagonalizes the rate matrix into a block diagonal matrix with two 2 blocks. Specifically,

$$M^{-1}Q^{(SSM)}M = \begin{bmatrix} q_0 + q_3 & q_1 + q_2 & 0 & 0 \\ q_4 + q_7 & q_5 + q_6 & 0 & 0 \\ 0 & 0 & q_5 - q_6 & q_4 - q_7 \\ 0 & 0 & q_1 - q_2 & q_0 - q_3 \end{bmatrix}$$

The diagonalization of rate matrix of a given evolutionary model makes it much simpler to find the exponential of the rate matrix. Recall that the transition probability matrix P is related to the rate matrix Q by the equation $P = \exp(Qt)$. Given a matrix H that diagonalizes the rate matrix notice that if $P = \exp(Qt)$ then

$$H^{-1}PH = H^{-1}\exp(Qt)H \quad (6.1)$$

$$= \exp(H^{-1}QtH). \quad (6.2)$$

Also for an $n \times n$ diagonal matrix A , $\exp(A)$ is equal to the diagonal matrix with diagonal entries equal to $\exp(a_i)$ for $1 \leq i \leq n$. The exponential of a matrix is much more difficult to compute if the matrix is not diagonal. In general the matrix exponential of a matrix $A \in \mathbb{F}^{n \times n}$ is defined as

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

If the matrix A can be block diagonalized into 2×2 blocks formulas exist to compute e^A however they are rather complicated.

Theorem 21. [5] Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2 \times 2}$, and define $\gamma = (a - d)^2 + 4bc$ and $\delta = \frac{1}{2}|\gamma|^{\frac{1}{2}}$. Then,

$$e^A = \begin{cases} e^{\frac{a+d}{2}} \begin{bmatrix} \cos(\delta) + \frac{a-d}{2\delta} \sin(\delta) & \frac{b}{\delta} \sin(\delta) \\ \frac{c}{\delta} \sin(\delta) & \cos(\delta) - \frac{a-d}{2\delta} \sin(\delta) \end{bmatrix}, & \gamma < 0, \\ e^{\frac{a+d}{2}} \begin{bmatrix} 1 + \frac{a-d}{2} & b \\ c & 1 - \frac{a-d}{2} \end{bmatrix}, & \gamma = 0, \\ e^{\frac{a+d}{2}} \begin{bmatrix} \cosh(\delta) + \frac{a-d}{2\delta} \sinh(\delta) & \frac{b}{\delta} \sinh(\delta) \\ \frac{c}{\delta} \sinh(\delta) & \cosh(\delta) - \frac{a-d}{2\delta} \sinh(\delta) \end{bmatrix}, & \gamma > 0. \end{cases}$$

Given the fact that the general strand symmetric model is not diagonalizable over the real numbers and rather only diagonalizable into 2×2 blocks the formulas that result do not appear to be useful. Comparing the n -taxon derivation of Hadamard conjugation provided in chapter 5 to the formulas given in Székely et al [60] implies that in order to interpret the results in terms of the relationship between the observed sequence spectrum and the edge length spectrum the rate matrices must be simultaneously diagonalizable.

6.7.1 Conclusion

The 1993 paper by Evans and Speed [20] had a significant impact on the way mathematicians and biologists think about nucleotide substitution models and Hadamard conjugation. Their observations regarding the Kimura three-substitution type model allows the process of mutation at a site using the model to be seen as providing a probability distribution on the Klein four group, rather than a probability distribution on the nucleotides. A probability distribution on the Klein four group, V , can be interpreted as an element of the real group algebra $\mathbb{R}V$. The same observation allows the edges in the tree to be labelled by elements of V with certain probabilities, and these, too, can be interpreted as elements of the group algebra. These observations suggest the idea that it is the algebra, rather than the group, that is significant.

The same observation of Evans and Speed allows leaf colorations to be seen as colorations by elements of V , rather than by nucleotides. Again a probability distribution can be seen as an element of a group algebra: this time of $\mathbb{R}V^{n-1}$, where n is the number of states in the model. It is this interpretation of leaf colorations that survives in the new setting: the tree is labelled with an element of $\mathbb{R}V_1^{n-1}$, for the new copy V_1 of the Klein four group, as illustrated in the example with model M_{37} . Here the subtle distinction between equality and isomorphism allows the for the interpretation to be transferred between M_{37} and Kimura three-substitution type model, while still providing biological, if not mathematical, novelty.

In algebraic combinatorics, the replacement of a group by regularity hypotheses that still allow an algebra to be constructed has a long and successful history. Applying this to certain

phylogenetic models has been one of the underlying ideas in the preceding chapter. In many ways, the group is still present, but only as a kind of phantom. One way to summarize Don Higman's ideas in algebraic combinatorics is to take the ideas in Schur and Wielandt on centralizer algebras as demonstrating that instead of permutations preserving a property being of leading importance, linear maps preserving that property play the primary role. In our setting, this moves attention from the centralizer of all Q -matrices in the symmetric group on the states (the automorphism group of the model) to the centralizer of all Q -matrices in the general linear group of degree equal to the number of states of the model.

Then, it turns out, that the essential ingredient in the full force of Hadamard conjugation is not an abelian permutation group acting regularly on the states, but rather an abelian linear group acting regularly on a basis of the (expanded) Q -algebra. This is the "phantom" referred to earlier.

The group can be thought of as a phantom because it is really the commutative algebra that matters; that it is a subalgebra of a commutative group algebra is true, but inessential to the argument. This is best illustrated in the case of codon models, where a 2-parameter model arising from the strongly regular Paley graph on sixty-one vertices would have to be expanded to a 60-parameter model to realize the group algebra and so the group, however, the simultaneous diagonalization of this commutative algebra can be described without resorting to the group.

Finally if the phylogenetic model is structured so that it relates to an association scheme, then the Bose-Mesner algebra of that association scheme can serve as the commutative algebra. This observation is only significant for models with sufficiently many states such as amino acid models in proteomics and codon models. We have displayed models that have few parameters and shown that Hadamard conjugation can be extended to these situations, but they are not biologically realistic models. In the attempt to create biologically realistic models the challenge will be to find commutative association schemes whose structure is biologically meaningful.

Chapter 7

PROBLEMS RAISED

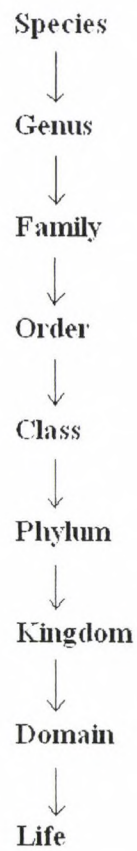
Given the connection between association schemes and evolutionary models it would be interesting to consider the question of how codon models corresponding to association scheme could be developed such that they are more biologically realistic. Given that each labeling of the points in an association scheme by codons produce biologically distinct models there are many models from which to choose. In order to ensure biologically meaningful models are created a dialog between mathematicians and biologists must take place to determine the properties such models should possess. Once these properties have been determined computational algebra expertise could be applied to search for appropriate models.

Other questions of interest that arise are:

- Which models exist where Hadamard conjugation works, but the rate matrix does not correspond to an association scheme?
- What, if any, is the biological meaning behind the correspondence between evolutionary models and association schemes?

Appendix A

A.1 Hierarchy



The eight major taxonomic ranks from the hierarchy of biological classification are listed above.

A.2 An indexing error

In [8] Bryant makes the following definition.

Definition 26. For $x, y \in G^m$ let

$$\hat{y}(x) = \prod_{i=1}^m \hat{y}_i(x_i).$$

Notice that x and y are vectors of the same length. Bryant then goes on to state a lemma in which $z \in G^q$ and $u \in G^{n-1}$. In the phylogenetic setting q represents the number of edges or color classes of a graph and $n - 1$ represents the number of non-root leaves. Consequently z and u are not of the same length. Despite this the term $\hat{z}(u)$ appears in the lemma. It is not clear how $\hat{z}(u)$ should be defined.

A.3 A counterexample to Bryant's Lemma

In [8] Bryant states the lemma provided below.

Lemma 6. Suppose that $z \in G^q$ and $y \in G^{n-1}$. Let \mathbf{A} be an $(n - 1) \times q$ integer matrix. Either

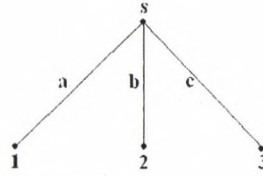
$$\sum_{x \in G^q: \mathbf{A}x=y} \hat{z}(x) = 0,$$

or there is $u \in G^{n-1}$ such that $z = \mathbf{A}^T u$ and so

$$\sum_{x \in G^q: \mathbf{A}x=y} \hat{z}(x) = r^{q-(n-1)} \hat{z}(u).$$

In Bryant's notation q is equal to the number of edges in the tree and n is equal to the number of leaves. Since Bryant assumes that the root must always be a leaf and never an interior vertex $n - 1$ is equal to the number of leaves minus the root vertex. Oftentimes vectors are indexed by the set of $n - 1$ non-root leaves.

In the lemma \mathbf{A} is defined to be an $(n - 1) \times q$ integer matrix, however applications of this lemma require \mathbf{A} to have a specific form that is dependent upon a predetermined tree. For example, consider the 3-claw tree shown below with root vertex labeled 1. This tree has $q = 3$ and $n - 1 = 2$.



The \mathbf{A} matrix corresponding to this tree is a 2×3 matrix with rows indexed by the non-root leaves, 2 and 3, and columns indexed by the edges, a , b and c . The matrix is defined by

$$\mathbf{A}_{i,k} = \begin{cases} 1 & \text{if } i \text{ and } n \text{ are on opposite sides of } \{A_k, B_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the \mathbf{A} matrix corresponding to the tree above is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

In the lemma the sum is taken over the set of x such that $\mathbf{A}x = y$, where y is indexed by the non-root leaves. The entries in y correspond to the mutation that has taken place along the path between the root and the leaf that indexes a particular entry in y . In the case of the Kimura-Three-Substitution types model there are four types of mutations that can take place and these mutations are represented by the elements of $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{(0,0), (1,1), (0,1), (1,0)\}$. These group elements can also be associated to the set of nucleotides, A , C , G , and T in the following way. Let $A = (0,0)$, $C = (1,1)$, $G = (0,1)$ and $T = (1,0)$.

For example, suppose the tree above is labeled such that the root is A and leaves 2 and 3 are C and T respectively. Since the root is labeled A and leaf 2 is labeled C the mutation changing A to C has taken place along the edges a and b . The mutation is referred to as $A - C = (0,0) - (1,1) = (1,1)$. A similar computation gives the mutation taking place between the root and leaf 3. Therefore the resulting y is $[(0,0) - (1,1), (0,0) - (1,0)]^T = [(1,1), (1,0)]^T$.

As a result $\mathbf{A}x = y$ implies that the group elements corresponding to the edges along a path from the root to a given leaf sum to the group element representing the mutation that has occurred from the root to that leaf. Looking at the example, if $y = [(1, 1), (1, 0)]^T$ then $\mathbf{A}x = y$ for the following x .

$$x_1 = \begin{bmatrix} (0,0) \\ (1,1) \\ (1,0) \end{bmatrix}, \quad x_2 = \begin{bmatrix} (1,1) \\ (0,0) \\ (0,1) \end{bmatrix}, \quad x_3 = \begin{bmatrix} (0,1) \\ (1,0) \\ (1,1) \end{bmatrix}, \quad x_4 = \begin{bmatrix} (1,0) \\ (0,1) \\ (0,0) \end{bmatrix}$$

The first entry of y is $(1, 1)$ which means that the mutation C has taken place along the path from the root to leaf 2. That means that the first two entries of x , x_a and x_b , must sum to $(1, 1)$. There are four ways this can happen; $(0,0) + (1, 1)$, $(1, 1) + (0, 0)$, $(0, 1) + (1, 0)$ or $(1, 0) + (0, 1)$. Each of these possibilities correspond to one of the four x vectors.

In the lemma, both x and z belong to G^q . As before I will assume that $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and that $q = 3$. To determine the value of $\widehat{z}(x)$ a z and x must be specified. Let $x = [(0, 0), (1, 1), (1, 0)]^T$ and $z = [(0, 1), (1, 0), (1, 1)]^T$. The character table for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ is also required to determine the values of $\widehat{z}(x)$ and so it is given below.

	A = (0,0)	C = (1,1)	G = (0,1)	T = (1,0)
A = (0,0)	1	1	1	1
C = (1,1)	1	-1	1	-1
G = (0,1)	1	1	-1	-1
T = (1,0)	1	-1	-1	1

Then by Bryant's definition,

$$\begin{aligned} \widehat{z}(x) &= \widehat{z}_1(x_1) \cdot \widehat{z}_2(x_2) \cdot \widehat{z}_3(x_3) \\ &= \widehat{z}_1((0, 0)) \cdot \widehat{z}_2((1, 1)) \cdot \widehat{z}_3((1, 0)) \\ &= 1 \cdot -1 \cdot -1 \\ &= 1 \end{aligned}$$

The Counterexample: The following example shows there is an error in the lemma. In this example $\sum_{x \in G^q: \mathbf{A}x=y} \widehat{z}(x) \neq 0$, however it is also not equal to $r^{q-(n-1)} \widehat{z}(u)$. Given the 3-claw tree, the corresponding \mathbf{A} matrix, $y = [(1, 1), (1, 0)]^T$ and the values of x listed above, let $z = [(0, 0), (1, 0), (1, 0)]^T$. In order to calculate $\widehat{z}(u)$ I assume the following definition (I also checked other possible values for $\widehat{z}(u)$).

Definition 27. For $z \in G^q$, $u \in G^{n-1}$,

$$\widehat{z}(u) = \prod_i^{\min(q, n-1)} \widehat{z}_i(u_i)$$

Next consider the following computation.

$$\begin{aligned} \sum_{x \in G^q: \mathbf{A}x=y} \widehat{z}(x) &= \widehat{z}(x_1) + \widehat{z}(x_2) + \widehat{z}(x_3) + \widehat{z}(x_4) \\ &= \widehat{z}_1(0,0) \cdot \widehat{z}_2(1,1) \cdot \widehat{z}_3(1,0) + \widehat{z}_1(1,1) \cdot \widehat{z}_2(0,0) \cdot \widehat{z}_3(0,1) \\ &\quad + \widehat{z}_1(0,1) \cdot \widehat{z}_2(1,0) \cdot \widehat{z}_3(1,1) + \widehat{z}_1(1,0) \cdot \widehat{z}_2(0,1) \cdot \widehat{z}_3(0,0) \\ &= (1 \cdot -1 \cdot 1) + (1 \cdot 1 \cdot -1) + (1 \cdot 1 \cdot -1) + (1 \cdot -1 \cdot 1) \\ &= -4 \end{aligned}$$

Therefore a $u \in G^{n-1}$ exists such that $z = \mathbf{A}^T u$ and that u is $[(1,0), (1,0)]^T$. However, notice that

$$\begin{aligned} r^{q-(n-1)} \widehat{z}(u) &= 4^1 \cdot \widehat{z}_1(1,0) \widehat{z}_2(1,0) \\ &= 4(-1 \cdot -1) \\ &= 4 \end{aligned}$$

and clearly $-4 \neq 4$.

Bibliography

- [1] Alexandre Ambrogelly, Sotiria Palioura, and Dieter Söll, *Natural expansion of the genetic code*, *Nature Chemical Biology* **3** (2007), no. 1, 29–35.
- [2] Jorgen Backelin and Svante Linusson, *Parity splits by triple point distances in x -trees*, *Annals of Combinatorics* **10** (2006), 1–18.
- [3] R.A. Bailey, *Association schemes, designed experiments, algebra and combinatorics*, Cambridge University Press, 2004.
- [4] E. Bannai and T. Ito, *Algebraic combinatorics i: Association schemes*, Benjamin-Cummings, 1984.
- [5] Dennis S. Bernstein, *Matrix mathematics, theory, facts, and formulas with application to linear systems theory*, Princeton University Press, 2005.
- [6] Joseph P. Bielawski and John R. Gold, *Mutation patterns of mitochondrial h- and l- strand dna in closely related cyprinid fishes*, *Genetics* **161** (2002), 1589–1597.
- [7] A.E. Brouwer, A.M. Cohen, and A. Neumaier (eds.), *Distance-regular graphs*, Springer-Verlag, 1989.
- [8] D. Bryant, *Extending tree models to splits networks*, *Algebraic Statistics For Computational Biology* (Lior Pachter and Bernd Sturmfels, eds.), Cambridge University Press, 2005.

- [9] David Bryant, *Hadamard phylogenetic methods and the n -taxon process*, *Bulletin of Mathematical Biology* **71** (2009), 339–351.
- [10] P. Buneman, *The recovery of trees from measures of dissimilarity*, *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, 1971.
- [11] Peter J. Cameron, *Association schemes and permutation groups*, <http://designtheory.org/library/encyc/topics/as.pdf>, 2003, Extended essay from *The Encyclopedia of Design Theory*.
- [12] ———, *Coherent configurations, association schemes and permutation groups*, *Groups, Combinatorics and Geometry*, World Scientific, 2003, pp. 55–72.
- [13] M. Casanellas, L.D. Garcia, and S. Sullivant, *Catalog of small trees*, *Algebraic Statistics For Computational Biology* (Lior Pachter and Bernd Sturmfels, eds.), Cambridge University Press, 2005.
- [14] M. Casanellas and S. Sullivant, *The strand symmetric model*, *Algebraic Statistics For Computational Biology* (Lior Pachter and Bernd Sturmfels, eds.), Cambridge University Press, 2005.
- [15] M.A. Charleston and R.D.M. Page, *Spectral analysis - a brief introduction*, *Molecular Systematics and Plant Evolution* (Peter Hollingsworth, Richard Bateman, and Richard Gornall, eds.), Taylor and Francis, 1999.
- [16] Benny Chor, Michael D. Hendy, and Sagi Snir, *Maximum likelihood jukes-cantor triplets: Analytic solutions*, *Molecular Biology and Evolution* **3** (2006), no. 23, 626–632.
- [17] Charles J. Colbourn and Jeffrey H. Dinitz (eds.), *The crc handbook of combinatorial designs*, CRC Press, 1996.
- [18] Andreas Dress and Daniel H. Huson, *Constructing splits graphs*, *IEEE Transactions on Computational Biology and Bioinformatics* **1** (2004), no. 3, 109–115.

- [19] David S. Dummit and Richard M. Foote, *Abstract algebra*, Wiley, 2004.
- [20] Steven N. Evans and T.P. Speed, *Invariants of some probability models used in phylogenetic inference*, *The Annals of Statistics* **21** (1993), no. 1, 355–377.
- [21] Warren J. Ewens and Gregory R. Grant, *Statistical methods in bioinformatics: An introduction*, Springer, 2005.
- [22] Joseph Felsenstein, *Inferring phylogenies*, Sinauer Associates, Inc, 2004.
- [23] C. D. Godsil, *Algebraic combinatorics*, Chapman and Hall Mathematics, 1993.
- [24] Nick Goldman and Ziheng Yang, *A codon-based model of nucleotide substitution for protein-coding dna sequences*, *Molecular Biology and Evolution* **11** (1994), no. 5, 725–736.
- [25] Jonathan Gross and Jay Yellen (eds.), *Handbook of graph theory*, CRC Press, 2004.
- [26] Charles Constantine Gumas, *A century old, the fast hadamard transform proves useful in digital communications*, <http://archive.chipcenter.com/dsp/DSP000517F1.html>.
- [27] Masami Hasegawa, Hirohisa Kishino, and Taka-Aki Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial dna*, *Journal of Molecular Evolution* **22** (1985), no. 2, 160–174.
- [28] M.D. Hendy, D. Penny, and M.A. Steel, *A discrete fourier analysis for evolutionary trees*, *The Proceedings of the National Academy of Sciences* **91** (1994), 3339–3343.
- [29] Michael D. Hendy, *The relationship between simple evolutionary tree models and observable sequence data*, *Systematic Zoology* **38** (1989), no. 4, 310–321.
- [30] ———, *Spectral analysis of phylogenetic data*, *Journal of Classification* **10** (1993), 5–24.
- [31] ———, *Hadamard conjugation: An analytic tool for phylogenetics*, *Mathematics of Evolution and Phylogeny* (Oliver Gascuel, ed.), Oxford University Press, 2005.

- [32] Michael D. Hendy and Michael A. Charleston, *Hadamard conjugation: a versatile tool for modelling nucleotide sequence evolution*, *New Zealand Journal of Botany* **31** (1993), 231–237.
- [33] Michael D. Hendy and Sagi Snir, *Hadamard conjugation for the kimura 3st model: Combinatorial proof using path sets*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5** (2008), no. 3, 461–471.
- [34] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, 1985.
- [35] Katharina T. Huber and Vincent Moulton, *Phylogenetic networks*, *Mathematics of Evolution and Phylogeny* (Oliver Gascuel, ed.), Oxford University Press, 2005.
- [36] John P. Huelsenbeck, Bret Larget, and Michael E. Alfaro, *Bayesian phylogenetic model selection using reversible jump markov chain monte carlo*, *Molecular Biology and Evolution* **21** (2004), no. 6, 1123–1133.
- [37] Gordon James and Martin Liebeck, *Representations and characters of groups*, second ed., Cambridge University Press, 2001.
- [38] T. H. Jukes and C. R. Cantor, *Evolution of protein molecules*, Academy Press, 1969.
- [39] Tamura K and Nei M, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees*, *Molecular Biology and Evolution* **10** (1993), no. 3, 512–526.
- [40] J.E. Karro, M. Peifer, R.C. Hardison, M. Kollmann, and H.H. von Grunberg, *Exponential decay of gc-content detected by strand-symmetric substitution rates influences the evolution of isochore structure*, *MBE Advance Access* (2007), 1–35.
- [41] T.W. Körner, *Fourier analysis*, Cambridge University Press, 1988.

- [42] J.R. Lobry and C. Lobry, *Evolution of dna base composition under no-strand-bias conditions when the substitution rates are not constant*, *Molecular Biology and Evolution* **6** (1999), no. 16, 719–723.
- [43] Kimura M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*, *Journal of Molecular Evolution* **16** (1980), no. 2, 111–120.
- [44] ———, *Estimation of evolutionary distances between homologous nucleotide sequences*, *The Proceedings of the National Academy of Sciences* **78** (1981), no. 1, 454–458.
- [45] Ruth Mace, Clare J. Holden, and Stephen Shennan (eds.), *The evolution of cultural diversity a phylogenetic approach*, UCL Press, 2005.
- [46] Itay Mayrose, Adi Doron-Faigenboim, Eran Bacharach, and Tal Pupko, *Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates*, *Bioinformatics* **23** (2007), 319–327.
- [47] Spencer V. Muse and Brandon S. Gaut, *A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome*, *Molecular Biology and Evolution* **11** (1994), no. 5, 715–724.
- [48] J. 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.
- [49] David Penny, Michael D. Hendy, Peter J. Lockhart, and Michael A. Steel, *Corrected parsimony, minimum evolution, and hadamard conjugation*, *Systematic Biology* **45** (1996), no. 4, 596–606.
- [50] David Posada and Keith A. Crandall, *Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (hiv-1)*, *Molecular Biology and Evolution* **18** (2001), no. 6, 897–906.

- [51] Charles Semple and Mike Steel, *Phylogenetics*, Oxford University Press, 2005.
- [52] Adam Siepel and David Haussler, *Phylogenetic estimation of context-dependent substitution rates by maximum likelihood*, *Molecular Biology and Evolution* **3** (2003), no. 21, 468–488.
- [53] M.A. Steel, M.D. Hendy, L.A. Székely, and P.L. Erdos, *Spectral analysis and a closest tree method for genetic sequences*, *Applied Mathematics Letters* **5** (1992), no. 6, 63–67.
- [54] Mike Steel, Michael D. Hendy, and David Penny, *Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results*, *Discrete Applied Mathematics* **88** (1998), 367–396.
- [55] Korbinian Strimmer, Carsten Wiuf, and Vincent Moulton, *Recombination analysis using directed graphical models*, *Molecular Biology and Evolution* **18** (2001), no. 1, 97–99.
- [56] Daniel W. Stroock, *An introduction to markov processes*, Springer, 2005.
- [57] Noboru Sueoka, *Intrastrand parity rules of dna base composition and usage biases of synonymous codons*, *Journal of Molecular Evolution* **40** (1995), 318–325.
- [58] David L. Swofford, Gary J. Olsen, Peter J. Waddell, and David M. Hillis, *Phylogenetic inference*, *Molecular Systematics* (David M. Hillis, Craig Moritz, and Barbara K. Mable, eds.), Sinauer Associates, Inc, 1996.
- [59] L.A. Székely, P.L. Erdos, M.A. Steel, and D. Penny, *A fourier inversion fomula for evolutionary trees*, *Applied Mathematics Letters* **6** (1993), no. 2, 13–16.
- [60] L.A. Székely, M.A. Steel, and P.L Erdos, *Fourier calculus on evolutionary trees*, *Advances in Applied Mathematics* **14** (1993), 200–216.
- [61] J.H. van Lint and R.M. Wilson, *A course in combinatorics*, second ed., Cambridge University Press, 2001.

- [62] R. Wetzel, *Zur visualisierung abstrakter ähnlichkeitsbeziehungen*, Ph.D. thesis, Universität Bielefeld, Germany, 1995.
- [63] Ziheng Yang, *Computational molecular evolution*, Oxford University Press, 2006.
- [64] Yang Z., *Estimating the pattern of nucleotide substitution*, *Journal of Molecular Evolution* **39** (1994), no. 1, 105–111.