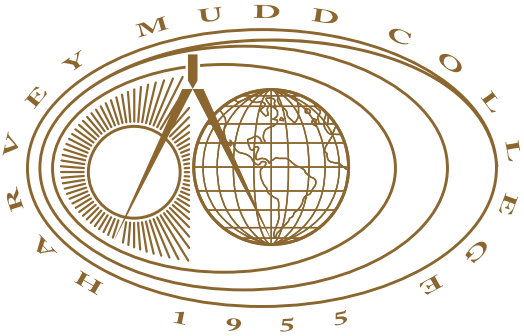
2007

Algebra and Phylogenetic Trees

Michael Hansen
Harvey Mudd College



Algebraic Invariants of Phylogenetic Trees

Mike Hansen

Francis Su, Advisor

Michael Orrison, Reader

May, 2007

HARVEY MUDD
COLLEGE
Department of Mathematics

Abstract

One of the restrictions used in all of the works done on phylogenetic invariants for group-based models has been that the group be abelian. In my thesis, I aim to generalize the method of invariants for group-based models of DNA sequence evolution to include nonabelian groups. By using a nonabelian group to act one the nucleotides, one could capture the structure of the symmetric model for DNA sequence evolution. If successful, this line of research would unify the two separated strands of active research in the area today: Allman and Rhodes's invariants for the symmetric model and Strumfels and Sullivant's toric ideals of phylogenetic invariants. Furthermore, I want to look at the statistical properties of polynomial invariants to get a better understanding of how they behave when used with real, "noisy" data.

Acknowledgments

I would like to thank Professor Francis Su for his helpful advice, suggestions, and encouragement over the course of this past year.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Phylogenetic Trees

Charles Darwin, in his *Origin of the Species*, hypothesized that life evolves due to natural selection and gene variation. The field of phylogenetics looks to understand these evolutionary relationships between both living and extinct species today. Trying to unravel these relationships is often known as phylogenetic inference since there is an unknown true history of evolution which we are trying to infer. It involves collecting and analyzing data in an attempt to obtain a best estimate for this true history.

The field of phylogenetics has applications to molecular biology, genetics, evolution, epidemiology, ecology, conservation biology, and forensics to name a few. Phylogenies are the historical and evolutionary relationships among organisms. Scientists can use this data to better understand how viruses spread or to study common biological processes between different species of life.

The overall goal within phylogenetics is to obtain the true phylogenetic tree, or phylogeny, for a given group of species, or taxa. A *phylogenetic tree* is a mathematical object which consists of a series of nodes and edges. A phylogenetic tree of vascular plants is shown below in Figure 1.1.

The leaves of the tree represent the taxa, or species, in consideration. The lines in the tree are called branches or edges and represent lines of evolutionary descent. Any point on an edge corresponds to the point of time in the life of an ancestor of some individual taxon. The internal nodes of the tree represent times in which the line of evolutionary descent for the taxa diverged. Leaves that are closer on the tree are more closely related than those which are far apart.

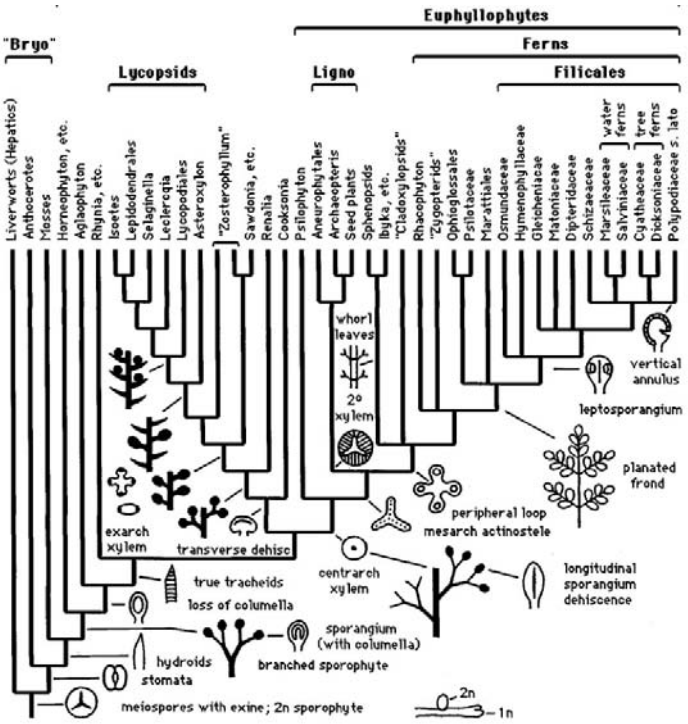


Figure 1.1: A phylogenetic tree of vascular plants. (Doyle, 1998)

1.2 Methods of Phylogenetic Tree Reconstruction

The earliest of techniques to reconstruct phylogenetic trees for a number of species used the physical, or morphological, characteristics of the organisms as a means to classify them. These techniques were extremely limited due to the simplicity of the data. With the advent of modern DNA techniques, the field of phylogenetics took a different turn. DNA sequences consists of two strands of molecules made up of the four DNA *nucleotides*: adenine, cytosine, guanine, and tyhmine. Sequences of DNA could be obtained for each of the organisms under consideration; assuming that organisms with similar DNA sequences are closely related, one could determine these relationships solely by looking at the DNA sequences.

One of the first methods of phylogenetic tree reconstruction based on DNA sequences was *maximum parsimony*. It seeks to determine the tree so that the total number of changes in the sequences along the edges is minimized. The technique is very simple and straightforward. For this reason,

it was one of the most popular method for phylogenetic tree reconstruction for a number of years. Unfortunately, maximum parsimony was shown not to be statistically consistent by Joseph Felsenstein in 1978. This means that for certain trees, maximum parsimony will produce an incorrect tree no matter how much data one had.

Other methods of phylogenetic tree reconstruction based on DNA sequences were developed in order to overcome this problem of statistical inconsistency. The most popular one, even to this day, is *maximum likelihood*. It uses an explicit model for DNA sequence evolution. With maximum likelihood, the DNA sequence evolution is modeled by some stochastic Markov process where the DNA nucleotides can mutate along the edges of the tree. Additionally, the lengths of the edges in a tree are proportional to the probability that a mutation occurs along that edge. For a given tree, one can estimate the likelihood (which is proportional to the probability) of that particular tree producing the DNA sequence data in question. The tree with the highest likelihood is the one selected by maximum likelihood.

One of the major drawbacks for maximum likelihood is that one must estimate all of the parameters in the model. In addition to the tree topology, maximum likelihood needs to estimate the rates of change between the nucleotides along each edge as well as the length of each of the edges. Maximum likelihood only works well when the assumptions about the model of sequence evolution actually hold for the given set of DNA sequence data.

The *method of invariants* for phylogenetic tree reconstruction is related to maximum likelihood in that it too requires a model of DNA sequence evolution. With a tree and that model of sequence evolution, one can calculate the probability of seeing a given pattern of nucleotides at the leaves of the tree. The *invariants* for the tree and model are multivariate polynomials with one indeterminate for each of the possible patterns of nucleotides along the leaves. The invariants vanish when they are evaluated at any of these pattern probabilities which come from the model of DNA sequence evolution. Unlike maximum likelihood, the method of invariants has the advantage of not requiring one to estimate all of the model parameters.

If one had all the invariant polynomials for a tree and model of sequence evolution, one could test whether or not a given set of DNA sequence data comes from that tree (and model). One can obtain estimates for the pattern probabilities by looking at the frequency that a given pattern appears in the DNA sequences.

1.3 Literature Review

The method of invariants was first introduced in Lake's "A rate independent technique for the analysis of nucleic acid sequences: evolutionary parsimony" (1987) and in Cavender and Felsenstein's "Invariants of phylogenies: a simple case with discrete states" (1987). They were able to produce invariants for a tree with four leaves considering only two-state character data as opposed to the four-state DNA data. With two-state data, the only Markov model of sequence evolution is the Jukes-Cantor model which specifies a probability for a character to change state and the complementary probability for the character to stay the same. The invariants found by Lake and Cavender–Felsenstein were linear invariants (i.e. the polynomials did not have quadratic or higher order terms).

In 1993, both Evans and Speed and Székely, Steel, and Erdös published work on the invariants for group-based models of DNA sequence evolution. A group-based model is one that can be represented by an algebraic group acting on the nucleotides. A group element is randomly selected and applied to a nucleotide to yield a new (not necessarily different) nucleotide. The Jukes-Cantor model, for both two and four character data, and Kimura's (1981) two or three parameter models are group-based models. They both employ the use of the Fourier transform in order to construct invariants for the group-based models.

The audiences for Evans and Speed (1993) and Székely et al. (1993) were different. Székely, Steel, and Erdös published their work in *Advances in Applied Mathematics*; thus, it presents the ideas from the perspective of theoretical mathematics. As Evans and Speed was written for the *Annals of Statistics* it makes heavy use of the language of probability and statistics. The Fourier transform is thought of in terms of an expected value.

In 2004, Sturmfels and Sullivant published "Toric Ideals of Phylogenetic Invariants" (2005). Their paper continues the work of Evans and Speed (1993) and Székely, Steel, and Erdös (1993) in the area of invariants for group-based models of DNA sequence evolution. They too make use of the Fourier transform and use it to define a change of coordinates for the pattern probabilities. Under this new coordinate system, every single invariant for the group-based models turns out to be a binomial. Furthermore, Sturmfels and Sullivant are able to describe all of these invariants in terms of the structure of the tree. That is, one can compute all of the invariants in this new coordinate system by simply looking at the tree. This paper caps off the research done in the invariants for group-based models giving a complete, elegant characterization of all the invariants.

In 2003, Elizabeth Allman and John Rhodes published "Phylogenetic invariants for the general Markov model of sequence mutation" (Allman and Rhodes, 2003). In it, they studied the invariants for the general or symmetric model of sequence evolution. Under this model, the rates at which one character (DNA nucleotide) can change into another can be independent for every pair of characters. They use four different techniques to arrive at their invariants using commutation and symmetry relations among the transition matrices for the model. This work is the furthest attempt at finding all of the invariants for the symmetric model of DNA sequence evolution.

1.4 My Work

One of the restrictions used in all previous work on invariants for group-based models has been that the group be abelian. In my thesis, I have attempted to generalize the method of invariants for group-based models of DNA sequence evolution to include nonabelian groups. By using a nonabelian group to act one the nucleotides, one could capture the structure of the symmetric model for DNA sequence evolution. If successful, this line of research would unify the two separated strands of active research in the area today: Allman and Rhodes's invariants for the symmetric model and Strumfels and Sullivant's toric ideals of phylogenetic invariants.

In addition to investigating the invariants for the symmetric model of DNA sequence evolution, I have studied the statistical properties of algebraic invariants for phylogenetic trees (for any of the models of sequence evolution). The technique of phylogenetic tree reconstruction via algebraic invariants can be made much more robust by understanding how these polynomials behave when working with real data. In the technique's current state, ad-hoc procedures must be used to tell when an algebraic invariant is "close enough" to zero.

I'd like to close this section with a quote from Joseph Felsenstein in his work "Inferring Phylogenies" (2003).

> But, these uses aside, invariants are worth attention, not for what they do now, but what they might lead to in the future. They are a very different way of considering tree topologies and branch lengths. Instead of crawling about in a tree space, trying to find the tree of best fit, they have us look at the relationship of pattern probabilities in a space of pattern frequencies, and build up our inferences of the tree in that space. For the cases

in which both invariants and the Hadamard conjugation apply, this is essentially the same as looking at which partitions show support in the Hadamard transform analysis. Both invariants and the Hadamard conjugation lead to interesting mathematics, and both give us a view of phylogenetic inference from a new direction. That alone would be enough to justify continued development of these interesting methods.

Chapter 2

Models of DNA Sequence Evolution

In order to use the method of invariants for phylogenetic tree reconstruction we need a model to describe how the DNA nucleotides evolve over time. The most often used models for DNA substitution are *Markov models*. These models assume that the probability that the character i changes to character j at a given site does not depend on its past history, only its current state. For these Markov models, we will further assume that each of the sites in a sequence evolve independently of each other. This condition can be relaxed with work, but we will maintain it for this thesis.

For DNA sequences, there are at most 16 substitution rates which we must consider: the probability that each character in $\{A, G, C, T\}$ is replaced by an alternative one. These rates are most often written as a $4x4$ instantaneous rate matrix. The i, jth element of the matrix represents the rate of change from character i to character j over an infinitesimal amount of time dt.

According to Swofford et al. (1996), the most general form of this matrix is

$$Q = \begin{pmatrix} -w & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -x & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -y & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -z \end{pmatrix}$$

where the rows and columns correspond to the bases A, C, G, and T respectively. The variables w, x, y, and z are just the sums of the rest of the

parameters in their respective row:

$$\begin{aligned} w &= \mu(a\pi_C + b\pi_G + c\pi_T) \\ x &= \mu(g\pi_A + d\pi_G + e\pi_T) \\ y &= \mu(h\pi_C + j\pi_C + f\pi_T) \\ z &= \mu(i\pi_A + k\pi_C + l\pi_G) \end{aligned}$$

The factor μ represents the average substitution rate, which is modified by the relative rate parameters a, b, ..., l. The product of the average rate parameter and a relative rate parameter constitute a *rate parameter*. The remaining parameters, π_A, π_C, π_G, π_T, are *frequency-parameters* which correspond to the frequencies of the bases over time. We generally assume that the frequencies are constant over time.

All DNA substitution models can be seen as some restriction of this most general model. In almost all cases, we want the model to be *time-reversible*, which means that the rate that a character i changes to j is the same as the rate at which j changes to i. This restriction corresponds to setting $g = a$, $h = b$, $i = c$, $j = d$, $k = e$, and $l = f$. A nice consequence to using time-reversible models is that calculations are not dependent on where a tree is a rooted.

The Jukes-Cantor model assumes that all the base frequencies are equal and that all substitutions occur at the same rate. The rate matrix for the Jukes-Cantor model looks like

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

We can divide the substitutions into two different classes: transitions and transversions. Transitions are substitutions from one purine (A, G) to another or one pyrimidine (C, T) to pyrimidine (i.e., $A \leftrightarrow G, C \leftrightarrow T$). Transversions are substitutions from a purine to a pyrimidine or vice verse. The Kimura 3-parameter model allows for transitions and 2 classes of transversions to occur at different rates. Its rate matrix is of the form

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

A number of other DNA substitution models along with their assumptions can be seen in the figure below. We will be concerned with extending the technique of spectral analysis to be used with the symmetric (SYM) model of DNA substitution.

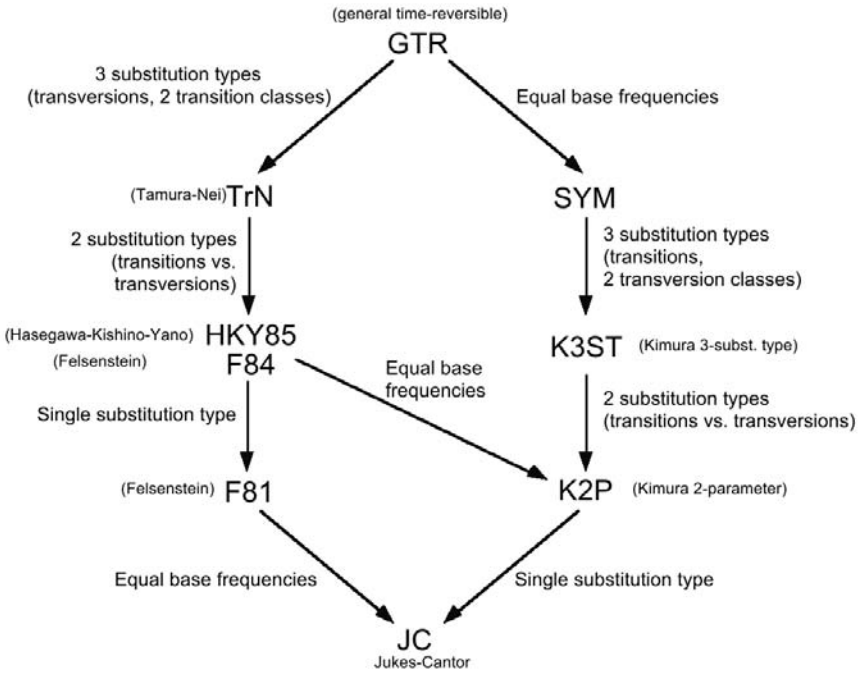


Figure 2.1: Relationships among some substitution models. The arrows go from a more general model to a more specific one with the assumptions stated on the side. (Swofford et al., 1996)

2.0.1 Group-based Models

One approach used in determining the invariants for various groups has been to use algebraic groups to model the DNA substitutions models. If we think of the elements of some group acting on the set of DNA bases, then we can think of the group elements as representing DNA mutations. We next assign values to each group element representing the probability that the substitution represented by the group element occurs over a period of time.

For example, the Jukes-Cantor model can be represented by the group

$Z_2 = S_2$. The identity element leaves the bases unchanged while the non-identity element switches a purine to a pyrimidine and vice verse. Note that this group does not take into consideration transitions and assumes that all transversions occur at the same rate.

By having the group $Z_2 \times Z_2$ act on the set of DNA bases, one is able to represent Kimura's 3-parameter model. The probabilities are assigned to group elements in the following manner:

$$\begin{aligned}(0,0) &\Leftrightarrow \text{No substitution}\\ (0,1) &\Leftrightarrow \text{Transition}\\ (1,0) &\Leftrightarrow \text{Transversion 1 (A-T, G-C)}\\ (1,1) &\Leftrightarrow \text{Transversion 2 (A-C, G-T)}\end{aligned}$$

This group and group action is the one that is considered in previous works on the invariants for group-based models.

If we do not require that the group acting on the bases be abelian, then the group S_4 can be used to represent SYM model of DNA substitution. It is still unclear whether or not the technique remains valid if the we relax the assumption of equal base frequencies in the SYM model to get the GTR model. We assign the probability that the bases remain unchanged to the identity element. The probability that base X mutates to base Y and vice verse is assigned to the transposition $(XY) \in S_4$. For example, the probability that an A changes to a C is assigned to the group element (AC). All of the other group elements are assigned a value of zero.

All of the work previously done on invariants required the use of abelian group-based models. This restricted the models to Kimura's 3-parameter model (1981) and its various submodels. Because the technique only works when the assumptions specified by the model hold, it is beneficial to be able to extend the technique to more general models.

2.0.2 DNA Nucleotides and Homogeneous Spaces

In using the method of invariants, one must calculate the frequency that a nucleotide pattern occurs in a set of DNA. This is a function defined on strings of nucleotide. For the Fourier techniques used with group-based models to work, we need a function that is defined on the elements of the group – not the nucleotides. This issue is not discussed in depth in any of the literature.

In the previous work done in this area, the primary group acting on the nucleotides is $Z_2 \times Z_2$. Since there were 4 group elements and 4 nu-

cleotides, each nucleotide was associated with a single group element:

$$\begin{aligned} A &\leftrightarrow (0,0) \\ G &\leftrightarrow (0,1) \\ C &\leftrightarrow (1,0) \\ T &\leftrightarrow (1,1). \end{aligned}$$

The arbitrariness of this identification is highlighted in "Small Phylogenetic Trees" (http://www.math.tamu.edu/~lgp/small-trees/) where the Garcia and Porter point out that the identification traditionally used did not follow the assumptions of Kimura's 2 parameter model. Thus, a replacement character table representing the correct identification is given.

When we deal with groups with more than four elements acting on the nucleotides, we cannot simply make this one to one association. Thus, it becomes more important to understand what is really going on when converting from functions defined on the nucleotides to functions defined on group elements.

The solution comes in terms of homogeneous spaces. In order to make precise this notion, we need some definitions.

Definition 2.1. *Let G be a group that acts on a set X. We say the group action is* transitive *if for every $x, y \in X$, there exists a $g \in G$ such that $g \cdot x = y$.*

This amounts to there being only one orbit under the group action. Since we want a nucleotide to be able to mutate into any of the other nucleotides, all the group actions under consideration will be transitive.

Definition 2.2. *A* homogeneous space *is a set upon which a group acts transitively.*

Thus, we can think about the nucleotides as being elements of a homogeneous space.

Definition 2.3. *The* stabilizer subgroup *S of an element x of a homogeneous space X is the set of all elements $s \in S$ such that $s \cdot x = x$.*

Suppose we have a stabilizer subgroup $S \subset G$ for an element $x \in X$. Each of the elements of X can be identified with a coset of S in G. More precisely, there exists a bijection $\phi : X \to G/S$ such that

$$\phi(g \cdot x) = g\phi(x)$$

for all $x \in X$ and $g \in G$.

A function defined on X can then be converted to a function defined on G which is constant on the cosets of S. Suppose $f : X \to \mathbb{C}$. We define a function $h : G \to \mathbb{C}$ by

$$h(g) = f\left(\phi^{-1}(gS)\right).$$

This gives us a method for handling groups with more than four elements acting on the nucleotides.

Example 2.1. *Consider the group $Z_2 \times Z_2$ acting on the nucleotides $X = \{A, G, C, T\}$. We notice that the only group element which leaves A fixed is $(0,0)$. Thus, the stabilizer of A is $S_A = \{(0,0)\}$. Then, we can identify the elements of X with the cosets of S_A:*

$$\begin{aligned} A &\leftrightarrow (0,0)S_A = \{(0,0)\} \\ G &\leftrightarrow (0,1)S_A = \{(0,1)\} \\ C &\leftrightarrow (1,0)S_A = \{(1,0)\} \\ T &\leftrightarrow (1,1)S_A = \{(1,1)\}. \end{aligned}$$

Note that this is the same identification as previously given.

Example 2.2. *Consider the group S_4 acting on the nucleotide $X = \{A, G, C, T\}$. The group elements that leave A fixed are the permutations which do not involve A. Thus, the stabilizer of A is*

$$\begin{aligned} S_A &= \{(), (GC), (GT), (CT), (GCT), (GTC)\} \\ &\cong S_3. \end{aligned}$$

Therefore, we can identify the elements of X with the cosets of S_A:

$$\begin{aligned} A &\leftrightarrow ()S_A = S_A \\ G &\leftrightarrow (AG)S_A \\ C &\leftrightarrow (AC)S_A \\ T &\leftrightarrow (AT)S_A. \end{aligned}$$

Chapter 3

Invariants for Abelian Group-based Models

I will illustrate the method of invariants using some simple examples. With these in mind, we will be able to build up the theory of invariants for both abelian and nonabelian groups in the next chapter. Although the examples deal with only two character 0/1 sequences, the ideas carry over naturally to the four character DNA case.

Before we begin with the example, we will need a few definitions. Suppose we have the following three aligned DNA sequences:

Taxon 1:	A	G	A	C	G	T	T	A	C	G	T	A	...
Taxon 2:	A	G	A	G	C	A	A	C	T	T	T	G	...
Taxon 3:	A	A	A	C	G	A	T	A	C	G	C	A	...

Then, we define a *pattern* σ to be the sequence of characters we get when we look at a single site (column) of our sequence data. In the sequences above, we can look at the second site in the sequences and see the pattern "GGA". A *pattern frequency* $\bar{p}_\sigma$ is the percentage of time that σ appears in our set of sequence data. In the above sequences, we see that the pattern "AAA" has a frequency of $2/12 = 1/6$ among the visible nucleotides.

Finally, if we assume that the DNA sequences come from a particular tree and model of DNA sequence evolution, then we define the *pattern probability* p_σ to be the percentage of time that expect to see the given pattern σ under that model of sequence evolution.

Now, we can finally define what an algebraic invariant of a phylogenetic tree and model of sequence evolution actually is.

Definition 3.1. *An* invariant *f for a phylogenetic tree and a model of sequence evolution is a polynomial in indeterminates x_σ such that*

$$f((p_\sigma)) = 0.$$

In other words, an invariant is a polynomial (with indeterminates for each pattern) such that when one plugs the pattern probabilities into their associated indeterminates, the polynomial evaluates to 0.

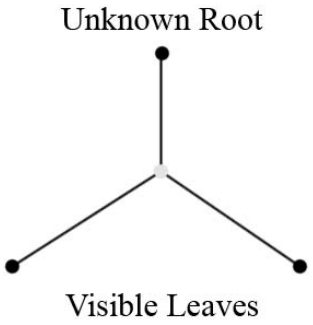
Example 3.1. *One example of an invariant for all trees and all models of sequence evolution is the* trivial invariant*:*

$$\left(\sum x_\sigma\right) - 1.$$

This is an invariant because we know that in any probability model the sum of the probabilities for all possible outcomes is 1.

If we have all the invariants for a tree and a particular model of sequence evolution, we can determine whether or not a set of sequences comes from that particular tree. In order to do so, we simply compute the pattern frequencies $\bar{p}_\sigma$ from the sequences. Since we assumed that each site in the sequence evolved independently and identically to the other sites, the pattern frequencies serve as estimators for the pattern probabilities p_σ. We then evaluate each of the invariants using the pattern frequencies. If the sequence data came from the tree and model of evolution associated with the invariants, then we would expect the evaluated polynomials not to differ significantly from zero.

A First Example



We will again be looking at 0/1 sequences with the group Z_2 acting on each of the "nucleotides" (characters). We will assume that the characters

0 and 1 are uniformly distributed at the root of the tree. Along each edge, we will random select an element of Z_2 with the probability of randomly selecting the element $1 \in Z_2$ being p and the probability of selecting the identity $0 \in Z_2$ being $q = 1 - p$. These act naturally on the characters 0 and 1. Thus, we can propagate a character at the root down the tree according to the group action. The pattern probabilities are the probabilities of seeing a given pattern at the leaves obtained in this manner.

With these assumptions, we can then explicitly calculate the pattern probabilities for this tree and model of sequence evolution:

$$p_{00} = \frac{1}{2}\left(q^3 + p^3\right) + \frac{1}{2}\left(pq^2 + qp^2\right)$$

$$p_{11} = \frac{1}{2}\left(q^3 + p^3\right) + \frac{1}{2}\left(pq^2 + qp^2\right)$$

$$p_{01} = pq^2 + qp^2$$

$$p_{10} = pq^2 + qp^2.$$

We can think about this as a parametrization of a curve in a four dimensional space with coordinates given by x_{00},x_{01},x_{10}, and x_{11}. As we let p range from 0 to 1, the curve will be traced out:

$$x_{00}(p) = p_{00} = \frac{1}{2}\left((1-p)^3 + p^3 + p(1-p)^2 + (1-p)p^2\right)$$

$$x_{01}(p) = p_{11} = \frac{1}{2}\left((1-p)^3 + p^3 + p(1-p)^2 + (1-p)p^2\right)$$

$$x_{10}(p) = p_{01} = p(1-p)^2 + (1-p)p^2$$

$$x_{11}(p) = p_{10} = p(1-p)^2 + (1-p)p^2.$$

This curve is a 1-dimensional *algebraic variety* sitting in a 4-dimensional "ambient space". As such it can be described as the roots of three different multivariate polynomials in the variables x_{00}, x_{01}, x_{10}, and x_{11}. These polynomials are the invariants

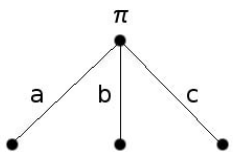
$$2x_{01} + 2x_{11} - 1, 2x_{10} + 2x_{11} - 1, x_{00} - x_{11}.$$

No matter what the value of p is, when we plug the pattern probabilities into the invariants, we will always get zero. Thus, the invariants depend only on the tree topology and the particular model of sequence evolution used; they do not depend on any particular parameters of the model.

The problem of computing these polynomials from the parametrization is known as the *implicitization problem* in algebraic geometry. For small trees, generic methods using Gröbner bases can be used to determine these invariants. For larger trees or models with many characters, more specialized and sophisticated methods must be employed.

Example: The 1,3 Claw Tree (Sturmfels and Sullivant, 2005)

The tree that we will be considering is the following:



We call π the root of the tree. At the root, we assume that the "nucleotides" (0 or 1) are randomly distributed according to the following distribution:

$$P(\pi = 0) = \pi_0$$
$$P(\pi = 1) = \pi_1.$$

Then, along each of the edges α, β, and γ, we randomly choose a group element from $Z_2 = \{0, 1\}$ according to the distributions

$$P(\alpha = 0) = \alpha_0$$
$$P(\alpha = 1) = \alpha_1$$
$$P(\beta = 0) = \beta_0$$
$$P(\beta = 1) = \beta_1$$
$$P(\gamma = 0) = \gamma_0$$
$$P(\gamma = 1) = \gamma_1.$$

These random group elements will act naturally on the nucleotide at the root to yield the nucleotides at the three leaves of the tree.

We can explicitly write out the pattern probabilities, that is the probability of seeing a given pattern of 0's and 1's along the leaves, in terms of the parameters along the edges. We can think of this as a parametrization of the pattern probabilities and will call this the parametrization in *probability coordinates*.

$$\begin{array}{ll} p_{000} = \pi_0\alpha_0\beta_0\gamma_0 + \pi_1\alpha_1\beta_1\gamma_1 & p_{001} = \pi_0\alpha_0\beta_0\gamma_1 + \pi_1\alpha_1\beta_1\gamma_0 \\ p_{010} = \pi_0\alpha_0\beta_1\gamma_0 + \pi_1\alpha_1\beta_0\gamma_1 & p_{011} = \pi_0\alpha_0\beta_1\gamma_1 + \pi_1\alpha_1\beta_0\gamma_0 \\ p_{100} = \pi_0\alpha_1\beta_0\gamma_0 + \pi_1\alpha_0\beta_1\gamma_1 & p_{101} = \pi_0\alpha_1\beta_0\gamma_1 + \pi_1\alpha_0\beta_1\gamma_0 \\ p_{110} = \pi_0\alpha_1\beta_1\gamma_0 + \pi_1\alpha_0\beta_0\gamma_1 & p_{111} = \pi_0\alpha_1\beta_1\gamma_1 + \pi_1\alpha_0\beta_0\gamma_0 \end{array}$$

The key insight in the Fourier-based methods is to use the Fourier transform as a change of coordinates. (The Fourier transform will be discussed in much more detail in the next chapter; for now, we can just think of it as a black box that provides a change of coordinates.) Along each of the edges, we have a function defined on a group (the probability of selecting an individual group element), and we can apply the Fourier transform to this function. Additionally, we can think of the patterns as being elements in $Z_2 \times Z_2 \times Z_2$. Thus, the pattern probabilities are functions defined on $Z_2 \times Z_2 \times Z_2$. We will also use the Fourier transform on the pattern probability function as well. These change of parameters and coordinates are shown below:

$$\begin{array}{ll} \pi_0 = \frac{1}{2}(r_0 + r_1) & \pi_1 = \frac{1}{2}(r_0 - r_1) \\ \alpha_0 = \frac{1}{2}(a_0 + a_1) & \alpha_1 = \frac{1}{2}(a_0 - a_1) \\ \beta_0 = \frac{1}{2}(b_0 + b_1) & \beta_1 = \frac{1}{2}(b_0 - b_1) \\ \gamma_0 = \frac{1}{2}(c_0 + c_1) & \gamma_1 = \frac{1}{2}(c_0 - c_1) \end{array}$$

$$p_{ijk} = \sum_{r=0}^{1}\sum_{s=0}^{1}\sum_{t=0}^{1}(-1)^{ir+js+kt}q_{rst}.$$

Under this new coordinate system, the each coordinate is parametrized by a monomial:

$$\begin{array}{ll} q_{000} = r_0a_0b_0c_0 & q_{001} = r_1a_0b_0c_1 \\ q_{010} = r_1a_0b_1c_0 & q_{011} = r_0a_0b_1c_1 \\ q_{100} = r_1a_1b_0c_0 & q_{101} = r_0a_1b_0c_1 \\ q_{110} = r_0a_1b_1c_0 & q_{111} = r_1a_1b_1c_1. \end{array}$$

We call this parametrization the parametrization in *Fourier coordinates*.

In this new coordinate system, the invariants are much easier to describe. The set of all invariants is known as a *toric ideal* and is generated by binomials. The invariants for this tree and model are shown below:

$$\left\{q_{001}q_{110} - q_{000}q_{111}, q_{010}q_{101} - q_{000}q_{111}, q_{100}q_{011} - q_{000}q_{111}\right\}.$$

If we wanted, we could convert these polynomials back to the probability coordinates. In doing so, each of the binomials would become a

quadratic polynomial with eight terms. One of these is shown below:

$$p_{001}p_{010} + p_{001}p_{100} - p_{000}p_{011} - p_{000}p_{101} +$$

$$p_{100}p_{111} - p_{101}p_{110} + p_{010}p_{111} - p_{001}p_{110}.$$

Chapter 4

Representation Theory

4.1 Introduction to Representation Theory

In order to understand the Fourier transform for finite groups, we will need to know about the representation theory of finite groups. An excellent introduction to this area is Jean-Pierre Serre's "Linear Representations of Finite Groups" (1977).

Let V be a vector space over the complex numbers and $GL(V)$ be the group of invertible linear transformations from V to itself.

Definition 4.1. *A linear representation of a finite group G in V is a homomorphism ϕ from the group G to $GL(V)$. Thus, we map each group element g to an element $\phi(g) \in GL(V)$ such that*

$$\phi(g_1 g_2) \quad = \quad \phi(g_1)\phi(g_2)$$

for any $g_1, g_2 \in G$. We define the degree of the representation d_ϕ to be $\dim V$.

Note that we can identify each linear transformation in $GL(V)$ with an invertible square matrix by choosing a basis for V.

Example 4.1. *Every group has a trivial representation T_d of degree d which maps every group element to the $d \times d$ identity matrix.*

Example 4.2. *Let $G = Z_2$, the cyclic group of order two. A one-dimensional representation ϕ_1 and a two-dimensional representation ϕ_2 are given below:*

$$\begin{array}{ll} \phi_1(0) = 1 & \phi_1(1) = -1 \\ \phi_2(0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \phi_2(0) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \end{array}$$

Notice that ϕ_2 looks like the direct sum of the degree 1 trivial representation T_1 and ϕ_1. In general, if ϕ and ψ are representations of a group G, the the map ρ defined by

$$\rho(g) \quad = \quad \phi(g) \oplus \psi(g)$$

is also a representation of G.

Definition 4.2. *We say a linear representation ϕ of G is* irreducible *if V is not 0 and no vector subspace of V is closed under the action of G, except of course 0 and V.*

Example 4.3. *The only subspaces that a one-dimensional vector space V can have are the trivial vector space and V itself. Those are the only vector subspaces that are stable under the action of G. Therefore, every one-dimensional representation is irreducible.*

These irreducible representations are the fundamental building blocks for every possible representation of G.

Theorem 4.1. *Every representation is a direct sum of irreducible representations.*

If we want to understand all the representations for a finite group, we simply need to understand the irreducible representations for that group.

Next, we turn our attention to direct products of groups. The irreducible representations for a direct product H be constructed from the irreducible representations of its factors.

Theorem 4.2. *Let $H = G_1 \times G_2 \times \cdots \times G_n$. Then, every irreducible representations for H is of the form*

$$\phi_1 \otimes \phi_2 \otimes \cdots \otimes \phi_n$$

where ϕ_i is an irreducible representation for G_i for $1 \leq i \leq n$. Additionally, if

$$\Phi = \phi_1 \otimes \phi_2 \otimes \cdots \otimes \phi_n$$

and

$$\Psi = \psi_1 \otimes \psi_2 \otimes \cdots \otimes \psi_n$$

are representations of H with $\phi_i \neq \psi_i$ for some $1 \leq i \leq n$, then $\Phi \neq \Psi$.

Note that if each G_i has n_i irreducible representations, then $H = G_1 \times G_2 \times \cdots \times G_n$ has

$$\prod_{i=1}^{n} n_i$$

irreducible representations.

It is easier to think about tensor products of irreducible representations as *Kronecker products* of matrices.

Definition 4.3. *Suppose A is an $m \times n$ matrix and B is a $p \times q$ matrix. We define the Kronecker product $A \otimes B$ to be the $mp \times nq$ matrix given by*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Now, we return to our direct product of groups $H = G_1 \times G_2 \times \cdots \times G_n$. Suppose that ϕ_i is an irreducible representation of the group G_i for $1 \leq i \leq n$. Then, $\phi_1 \otimes \cdots \otimes \phi_n$ is an irreducible representation for H and is defined by

$$(\phi_1 \otimes \cdots \otimes \phi_n)(g_1 \times \cdots \times g_n) = \phi_1(g_1) \otimes \cdots \otimes \phi_n(g_n).$$

All of the irreducible representations for H can be constructed in this manner.

We can also think about tensor products of representations for a single group G. If ϕ and ψ are representations for a group G, then $\phi \otimes \psi$, defined by $(\phi \otimes \psi)(g) = \phi(g) \otimes \psi(g)$ for all $g \in G$, is a representation of G.

Example 4.4. *Let $H = Z_2 \times Z_2$. Since Z_2 is abelian, all its irreducible representations are one-dimensional. As we saw before, Z_2 has the trivial representation T_1 and the alternating representation A_1 of degree one. Therefore, H has four irreducible representations:*

$$T_1 \otimes T_1$$

$$T_1 \otimes A_1$$

$$A_1 \otimes T_1$$

$$A_1 \otimes A_1.$$

Characters

None of the previous work done on invariants for group-based models makes use of the language of irreducible representations. Instead, the mathematics is framed in terms of characters. As we will see, the language of characters is particularly well-suited to abelian groups.

The irreducible representations of an abelian group are well characterized by the following theorem.

Theorem 4.3. *Every irreducible representation for an abelian group is one dimensional.*

With this in the back of our mind, we are now ready to define the character of a representation.

Definition 4.4. *The character χ_ρ of a representation ρ is the map from the group G to the complex numbers defined by*

$$\chi_\rho(g) \quad = \quad \mathrm{Tr}(\rho(g)).$$

Since every irreducible representation of an abelian group is one dimensional, the characters and irreducible representations of an abelian group are equivalent. In generalizing results from the abelian case to the non-abelian case, we will often need to view the abelian group characters as irreducible representations instead.

Fourier Transform

The Fourier transform plays a key role in the method of invariants for group-based models of DNA sequence evolution. The Fourier transform provides a change of coordinates

Theorem 4.4. *Let G be a group, $\hat{G}$ be the set of all irreducible representations of G, and f be a function from G to the complex numbers. Then, we define the Fourier transform $\hat{f}$ of f at an irreducible representation ϕ by*

$$\hat{f}(\phi) \quad = \quad \sum_{g \in G} f(g)\phi(g).$$

The inverse Fourier transform is given by

$$f(g) \quad = \quad \frac{1}{|G|} \sum_{\phi \in \hat{G}} d_\phi \mathrm{Tr}(\hat{f}(\phi)\phi(g^{-1}))$$

where d_ϕ is the degree of ϕ.

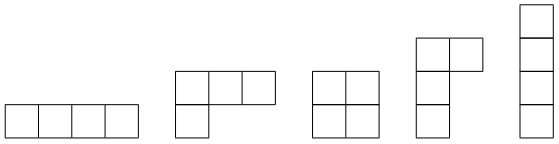


Figure 4.1: Ferrer's diagrams for symmetric group 4.

Furthermore, if G is an abelian group, then the set of its irreducible complex representations $\hat{G}$ forms a group under multiplication. When endowed with this group structure, $\hat{G}$ is isomorphic to G itself.

4.2 Representation Theory of the Symmetric Group

Since the group S_4 acts naturally on the four DNA nucleotides, it will be useful to know its representation theory. The encyclopedic text for this topic is "The Representation Theory of the Symmetric Group" written by James and Kerber (1984).

The irreducible representations for the symmetric group are indexed by integer partitions. For example, there are 5 integer partitions of the number 4:

$$\begin{aligned} 4 &= 4 \\ &= 3+1 \\ &= 2+2 \\ &= 2+1+1 \\ &= 1+1+1+1. \end{aligned}$$

Therefore, the group S_4 has 5 distinct irreducible representations. These partitions are typically represented as Ferrer's diagrams as in Figure 4.1.

We can specify the *shape* of a partition by specifying the number of boxes in each row. The shapes of the 5 integer partitions of 4 are $\lambda_1 = (4)$, $\lambda_2 = (3,1)$, $\lambda_3 = (2,2)$, $\lambda_4 = (2,1,1)$, and $\lambda_5 = (1,1,1,1)$.

The dimension of each representation can be determined from these Young diagrams and is given in terms of *standard Young tableaux*. A standard Young tableaux is a filling of a Young diagram with the numbers 1 though n such that the numbers increase as one moves down and to the right. The dimension of a representation of shape λ is the number of stan-

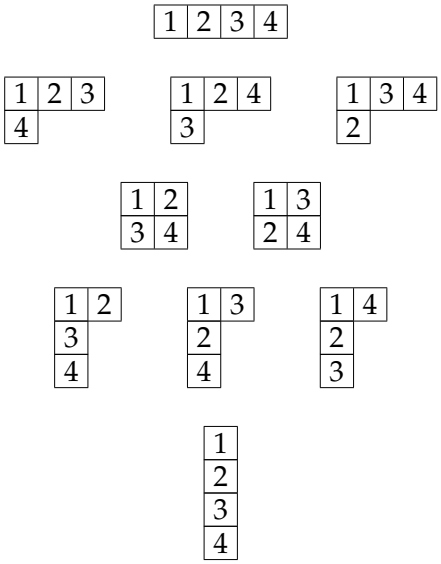


Figure 4.2: Standard Young tableaux for the partitions of 4. The number of standard Young tableaux gives the dimension of the corresponding representation.

dard Young tableaux of shape λ. The standard Young tableaux for the partitions of 4 are shown in Figure 4.2.

As can be seen in the figure, the dimension of the irreducible representation corresponding to partition $\lambda = (2, 2)$ is 2.

The trivial partition ($4 = 4$) corresponds to the one dimensional trivial representation. That is, $\rho_{(4)}(\pi) = 1$ for all $\pi \in S_4$. The all-ones partition ($4 = 1 + 1 + 1 + 1$) corresponds to the one-dimensional *alternating representation* which is defined in terms of the sign of a permutation. The sign $\text{sgn}(\pi)$ of a permutation π is (well-)defined to be 1 if π can be written as a product of an even number of transpositions; otherwise, it is defined to be -1. The mapping from S_4 to $\{1, -1\}$ defined by

$$\rho_{(1,1,1,1)}(\pi) \quad = \quad \text{sgn}(\pi)$$

is a homomorphism and is known as the alternating representation.

For the higher-dimensional representations, there are an infinite number of ways to choose the specific matrices. Each of these ways corresponds to choosing a basis for the space. Young's seminormal form and Young's orthogonal form are two of the most common forms for these matrices.

Although the particular form chosen will not play a role in the theoretical aspects of my work, one will need to choose a form for any practical implementation of the ideas. It is typically easier to work with Young's orthogonal form as all of the matrices are orthogonal. Young's Orthogonal representations of the generators for the symmetric groups S_2, S_3, and S_4 can be found in Appendix A.

Chapter 5

Invariants for Nonabelian Group-based Models

In this chapter, I will present the progress I have made in generalizing the Fourier-based methods for determining the invariants of nonabelian group-based models. I will try to make my notation consistent with the notation used in Székely et al. (1993).

5.0.1 Basic Model

We will start with a tree T with a set of leaves L^* where $|L^*| = n$. We will denote one of the leaves as the *root* R and denote the set of non-root leaves $L^* \setminus R$ by L. The edges of the tree will be denoted by $E(T)$ and the nodes by $V(T)$.

Let G be a finite group which acts transitively on the set of four DNA nucleotides $\{A, G, C, T\}$. Along each edge $e \in E(T)$, we will have independent G-valued random variables ξ_e. These will represent the random mutation of the DNA nucleotides along the edges of the tree. The probability mass function for ξ_e is given by the function p_e defined for each $g \in G$. That is

$$p_e(g) \quad := \quad \text{Prob}(\xi_e = G)$$

with $\sum_{g \in G} p_e(g) = 1$.

We will consider the direct product G^{n-1} to be the set of all *leaf colorations* $\sigma : L \to G$, and we denote the value of σ at a leaf $l \in L$ by σ_l. We can produce a random leaf coloration by evaluating each of the random variables ξ_e along the edges and giving each leaf the group element

obtained by multiplying all of the group elements along the unique path $R \to l$ from the root to the leaf. We will denote by f_σ the probability of seeing a leaf coloration obtained in this manner.

Suppose we now associate an irreducible representation of G, ψ_l, with each leaf $l \in L$. Then, we can define an irreducible representation for the direct product G^{n-1} by

$$\psi \quad = \quad \otimes_{l \in L} \psi_l.$$

Conversely, given an irreducible representation for the group G^{n-1}, we can obtain an irreducible representation ϕ_l of G for each of the leaves $l \in L$.

Next, we define the *leaf set* of an edge $e \in E(T)$ by

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

For an edge $e \in E(T)$ and an irreducible representation ϕ of G^{n-1}, we can define a new representation for G^{n-1} in the following manner:

$$\psi_e \quad = \quad \otimes_{l \in L} \psi_{l,e} \tag{5.1}$$

where $\psi_{l,e}$ is defined to be ψ_l if $l \in L_e$ and the trivial representation of degree d_{ψ_l} if $l \notin L_e$. Notice that for all edges $e, e' \in E(T)$, the degree of ψ_e and $\psi_{e'}$ are equal. Also note that ψ_e can be viewed as a representation for G by using the natural, injective homomorphism between G and G^{n-1}.

Now, suppose that ϕ is a representation of G and $e \in E(T)$. We can make the following two definitions:

$$l_e(\phi) \quad = \quad \sum_{g \in G} p_e(g)\phi(g) \tag{5.2}$$

$$r_\psi \quad = \quad \prod l_e(\psi_e). \tag{5.3}$$

Notice that $l_e(\phi)$ is just the Fourier transform of the probability mass function p_e along the edge e. Thus, it may be more intuitive to denote this function by $\hat{p}_e$.

The main theorem relates the Fourier transform of the probability function defined on the leaf colorations to the Fourier transforms of the probability mass functions along the edges. It is a direct application of the lemma below. The lemma will be presented in a more general setting, but we will note the connections to our previous tree-based setting.

Let $A = (a_{ij})$ be a $p \times q$ matrix with integer entries. Let

$$x = (x_1, x_2, \ldots, x_q)^T \in G^q$$

be a vector of length q with each entry $x_i \in G$. The, we define a vector $y \in G^p$ by $y = (y_1, y_2, \ldots, y_p)^T$ with

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

We will abbreviate this as $y = Ax$. This notation resonates with the notation used in the works by Evans (2004) and Evans and Speed (1993). In those works, a group-valued random variable at each of the leaves of the tree is defined to be a sum (product) of random variables along the edges. Indeed, this relationship is occasionally represented by an appropriate "design matrix" in Evans and Speed.

Thus, we will think of the vector x as representing an evaluation of all the random variables along the edges of the tree. The vector y will correspond the the leaf coloration obtained by propagating the group elements down the tree from the leaves to the root.

Let us be given functions $p_j : G \to \mathcal{C}$ for $j \in \{1, 2, \ldots, q\}$. For each $x \in G^p$, we define

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

In our original setting, the functions p_j correspond to the probability mass functions p_e along each edge. With this in mind, $F(x)$ is the probability of obtaining a particular evaluation x of group elements along the edges.

Then, for $y \in G^p$, we define

$$f(y) \quad = \quad \sum_{Ax=y} F(x).$$

Note that this sum is over all possible evaluations of the random variables along the edges which give rise to the leaf coloration y. Thus, $f(y)$ will represents the probability of seeing a leaf coloration y.

The following lemma relates the Fourier transform of f to the probability mass functions along the edges.

Lemma 5.1. *Let ψ_i be an irreducible representation of G for $1 \leq i \leq n$ and $\psi = \psi_1 \otimes \psi_2 \cdots \otimes \psi_p$ be the corresponding irreducible representation of G^p. Then,*

$$\hat{f}(\psi) \quad = \quad \prod_{j=1}^{q} \sum_{g \in G} \left(p_j(g) \bigotimes_{i=1}^{p} \psi_i \left(g^{a_{ij}} \right) \right)$$

Proof. By definition, we have that

$$\begin{aligned}\hat{f}(\psi) &= \sum_{y \in G^p} \psi(y) f(y) \\ &= \sum_{y \in G^p} \left(\psi(y) \sum_{Ax=y} F(x) \right) \\ &= \sum_{x \in G^q} F(x) \psi(Ax).\end{aligned}$$

Now, we have that

$$\begin{aligned}\psi(Ax) &= (\psi_1 \otimes \cdots \otimes \psi_p)(Ax) \\ &= \bigotimes_{i=1}^{p} \psi_i((Ax)_i) \\ &= \bigotimes_{i=1}^{p} \psi_i \left(\prod_{j=1}^{q} x_j^{a_{ij}} \right) \\ &= \prod_{j=1}^{q} \bigotimes_{i=1}^{p} \psi_i \left(x_j^{a_{ij}} \right).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{f}(\psi) &= \sum_{x \in G^q} F(x) \psi(Ax) \\ &= \sum_{x \in G^q} \left(\prod_{j=1}^{q} p_j(x_j) \right) \left(\prod_{j=1}^{q} \bigotimes_{i=1}^{p} \psi_i \left(x_j^{a_{ij}} \right) \right) \\ &= \sum_{x \in G^q} \prod_{j=1}^{q} \left(p_j(x_j) \bigotimes_{i=1}^{p} \psi_i \left(x_j^{a_{ij}} \right) \right) \\ &= \prod_{j=1}^{q} \sum_{g \in G} \left(p_j(g) \bigotimes_{i=1}^{p} \psi_i \left(g^{a_{ij}} \right) \right).\end{aligned}$$

□

Now, we can apply this lemma in the context of our tree.

Theorem 5.1. *Let $\psi = \otimes_{l \in L} \psi_l$. Then, we have the following Fourier inverse pair:*

$$r_\psi = \sum_{\sigma \in G^{n-1}} \psi(\sigma) f_\sigma \tag{5.4}$$

$$f_\sigma = \frac{1}{|G|^{n-1}} \sum_{\phi \in \hat{G}^{n-1}} d_\phi \mathrm{Tr}(\phi(\sigma^{-1}) r_(\phi)). \tag{5.5}$$

Proof. First, notice that this theorem just states that r_ψ is the Fourier transform of f, $\hat{f}(\psi)$. Thus, we know that the two equations above are equivalent. Thus, we just need to show that r_ψ as originally defined is equivalent to the one in the theorem.

To do this, we will apply the previous lemma in the following setting: $p = n - 1$, $q = |E(T)|$, and $A = (a_{le})$ with $a_{le} = 1$ if e lies on the path from R to l and zero otherwise. From the original definition of r_ψ, we have

$$\begin{aligned} r_\psi &= \prod_{e \in E(T)} l_e(\psi_e) \\ &= \prod_{e \in E(T)} \sum_{g \in G} p_e(g) \psi_e(g) \\ &= \prod_{e \in E(T)} \sum_{g \in G} p_e(g) \left(\bigotimes_{l \in L_e} \psi_l \right)(g). \end{aligned}$$

Thus, by the previous lemma, we have that $r_\psi = \hat{f}_\psi$.

Finally, we confirm that the functions f and F work in our setting. Let $\Xi = (\xi_e : e \in E(T))$ be the vector of random group elements selected independently on the edges, $p_e(g) = \mathrm{Prob}(\xi_e = g)$, and Y be the vector of the resulting leaf coloration. As we noted before, independence among the edges implies $F(x) = \mathrm{Prob}(\Xi = x)$ and $f(y) = \mathrm{Prob}(\mathrm{Y} = y)$. □

Since $l_e(\psi_e)$ is a Fourier transform of a probability mass function along an edge, this theorem allows us to express the Fourier transform of $f(\sigma)$ as a product of Fourier transforms of functions along the edges.

5.1 Free Algebras

Exploiting this product structure was the key to Sturmfels and Sullivant's work in determining the invariants the abelian group-based models. Since all the irreducible representations for abelian groups are one dimensional,

all of their Fourier transforms were just complex numbers (1×1 matrices over $\mathbb{C}$). This allowed them to define the following monomial map:

$$a_{\psi_1 \otimes \cdots \otimes \psi_{n-1}} \mapsto \prod_{e \in E(T)} b_{e,\psi_e}.$$

This can be extended to a complex polynomial ring homomorphism from the ring with an a-indeterminate for all the irreducible representations of G^{n-1} to a ring with a b-indeterminate for each pair e and $\psi_{e'}$. The invariants are the kernel of this homomorphism.

When a nonabelian group is acting on the nucleotides, some of the Fourier transforms are matrices and not numbers. In this case, the polynomial ring homomorphism defined above may not be appropriate. In general, matrices do not commute like the complex numbers or the indeterminates in a polynomial ring.

The most natural noncommutative generalization of the the polynomial ring $K[x_1, x_2, \ldots, x_n]$ is the *free (associative) algebra* $X\langle x_1, x_2, \ldots, x_n \rangle$.

Definition 5.1. *Let X be a nonempty, finite set. We think of the elements of X as letters in an alphabet. A* word *of length n is an ordered n-tuple of elements of X.*

Example 5.1. *For example, if $X = \{a, b\}$, then the following are all the words of length 2.*

$$aa$$
$$ab$$
$$ba$$
$$bb$$

Definition 5.2. *We will define the set X^* to be the set of all words of finite length from X. Suppose $x, y \in X^*$. Then, we define their product vw to be the word obtained by concatenating w to the end of v.*

Example 5.2. *Again, suppose $X = \{a, b\}$. Let $x = abba$ and $y = bbb$. Then,*

$$\begin{aligned} xy &= abbabbb \\ yx &= bbbabba. \end{aligned}$$

Notice that this multiplication is a noncommutative operation.

Now, consider the vector space V with basis X^* over some field K. We extend the product in X^* to V by distributivity:

$$\left(\sum_i \alpha_i v_i \right) \left(\sum_j \beta_j w_j \right) = \sum_{i,j} \alpha_i \beta_j v_i w_j.$$

This multiplication turns V into an associative algebra. If $|X^*| = n$, then we call V the free (associative) algebra over K on n generators. We will denote the free algebra by $K\langle X\rangle$.

A free algebra $K\langle x_1, x_2, \ldots, x_n\rangle$ behaves much like the polynomial ring $K[x_1, x_2, \ldots, x_n]$ with the exception that the indeterminates of

$$K\langle x_1, x_2, \ldots, x_n\rangle$$

do not commute while the elements of $K[x_1, x_2, \ldots, x_n]$ do. However, there are indeed some differences. For example, free associative algebras do not have a unit element while the polynomial rings do.

5.1.1 Ideals for Free Associative Algebras

Just as we consider ideals of polynomial rings, we can consider ideals of free algebras. For example, let I be the ideal in $K\langle a, b, c\rangle$ generated by the following commutivity relations:

$$ab - ba$$
$$ac - ca$$
$$bc - cb.$$

The (left, right, or double-sided) ideal I consists of all scalar multiples, sums of the elements of I, and products of the elements of I by the elements of $K\langle a, b, c\rangle$ (on the left, on the right, or on both sides). If we see were to look at the quotient of $K\langle a, b, c\rangle$ with the double-sided ideal I, we'd see

$$K\langle a, b, c\rangle / I \cong K[a, b, c].$$

Given an ideal in a free algebra, we can ask about a minimal set of generators for the ideal. The theory of Gröbner bases of commutative (polynomial) rings has been extended to the noncommutative case of free algebras. For a brief introduction, see Teo Mora's "An introduction to commutative and noncommutative Gröbner bases" (1994). Unlike the commutative case, there not exist an algorithm for producing a set of generators (a Gröbner basis) for a generic ideal in the noncommutative case. This is due to the fact that not all ideals are finitely generated.

As proved in Helton and Stankus (1999), the theory of *elimination ideals* carries over to the noncommutative case. Elimination ideals are useful in solving implicitization problems – going from a parametric form of an algebraic variety to an implicit form. This suggests a definition for a noncommutative generalization of toric ideals.

Definition 5.3. *Let X and T be finite sets. A* noncommutative toric ideal *is the kernel of a homomorphism ϕ between free algebras $K\langle X\rangle$ and $K\langle T\rangle$ which maps monomials in $K\langle X\rangle$ to monomials in $K\langle T\rangle$.*

One method by which to begin investigating these objects would be the free software package Bergman which can be found at `http://servus.math.su.se/bergman/`. It provides an easy to use interface for computing noncommutative Gröbner bases in free algebras. In addition, it has elimination orderings (for elimination ideals) built in. It would be interesting to see if these noncommutative toric ideals are as highly structured as their commutative counterparts.

Chapter 6

Statistical Properties of Algebraic Invariants for Phylogenetic Trees

In order for the reconstruction of phylogenetic trees via algebraic invariants to be applicable, it needs to be able to deal with data from real DNA sequences. Allman and Rhodes (2003) gives the following characterization of the current state of affairs:

> Then, in order to apply invariants to real data, one must decide what it means for an invariant to be "close to vanishing" on observed frequencies. A statistical understanding of the behavior of these polynomials on noisy data is highly desirable. Moreover, as there are infinitely many invariants, choosing a finite set of generators with good statistical properties is necessary. Finally, robustness of the method under violation of model assumptions is critical to applications, since models of sequence evolution are only approximate assumptions of reality. While much work remains to implement such a plan, the approach has intrigued a number of researchers.

We can see this lack of a thorough statistical understanding of the invariant polynomials when we look at, say, Chapter 15 of "Algebraic Statistics for Computational Biology" (Pachter and Sturmfels, 2005). When they evaluated the effectiveness of reconstructing tress with the method of invariants, they used an ad-hoc method which selected the tree that minimized the sums of the squares of the polynomials after evaluating them at

the observed pattern frequencies. With this in mind, I plan to investigate the statistical behavior of the polynomial invariants on noisy data due to the effects of only having finite DNA sequences.

A Multinomial Experiment

Suppose we have a m aligned sequences of length n from an alphabet of ℓ letters. Furthermore, suppose that these sequences were produced via some Markov process model for DNA sequence evolution on some tree topology T.

Given a tree and transition mechanism, we can calculate the probability p_σ of seeing a given pattern σ. Then, each column can be thought of as the result of a Benuoulli trial where there are ℓ^m possible outcomes (one for each σ) and the probability of obtaining outcome σ is p_σ.

Let X_σ be the number of "successes" of seeing σ upon n of these independent Bernoulli trials. Then, X_σ has a binomial distribution with parameters n and p_σ. The covariance between between any two distinct X_{σ_1} and X_{σ_2} is given by

$$\operatorname{cov}\left(X_{\sigma_1}, X_{\sigma_2}\right) \quad = \quad -np_{\sigma_1}p_{\sigma_2}.$$

As $n \to \infty$, we have

$$\frac{X_\sigma - np_\sigma}{\sqrt{np_\sigma(1-p_\sigma)}} \quad \to \quad N\left(\mu = 0, \sigma^2 = 1\right).$$

We are interested in the pattern frequencies $\bar{p}_\sigma = X_\sigma/n$. As $n \to \infty$, we can calculate the distribution of the pattern frequencies:

$$\begin{aligned} X - np_\sigma &\sim N\left(\mu = 0, \sigma^2 = np_\sigma(1-p_\sigma)\right) \\ X &\sim N\left(\mu = np_\sigma, \sigma^2 = np_\sigma(1-p_\sigma)\right) \\ \bar{p}_\sigma = X/n &\sim N\left(\mu = p_\sigma, \sigma^2 = p_\sigma(1-p_\sigma)/n\right). \end{aligned}$$

Note that this approximation only holds for relatively large n. A typical rule of thumb is that both np_σ and $n(1-p_\sigma)$ both must be greater than 5. This presents somewhat of a problem in the general case since p_σ is not known when we are trying to reconstruct a phylogenetic tree.

Since we do know what p_σ is, we the following worst case bound for $\bar{p}_\sigma$:

$$\bar{p}_\sigma \quad \sim \quad N\left(\mu = p_\sigma, \sigma^2 = 1/4n\right)$$

since the largest value $p_\sigma(1-p_\sigma)$ can take on is $\frac{1}{4}$.

Invariants in Probability Coordinates

Now, we want to see how the algebraic invariants of a phylogenetic tree behave when we plug in these pattern frequencies. We will first consider the case of linear invariants in the probability coordinates. As a first approximation, assume that X and Y are independent random variables. If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

and

$$aX \sim N(a\mu_X, a^2\sigma_X^2).$$

Suppose we have a linear invariant $f((x_\sigma)) = f_0 + \sum c_\sigma x_\sigma$ in the probability coordinates. Then,

$$\begin{aligned} f((X_\sigma)) &= f_0 + \sum c_\sigma X_\sigma \\ &\sim N\left(f_0 + \sum c_\sigma \mu_\sigma, \sum c_\sigma^2 \sigma_\sigma^2\right) \\ &\sim N\left(f_0 + \sum c_\sigma p_\sigma, \sum c_\sigma^2 \left(\frac{1}{4n}\right)\right) \\ &\sim N\left(0, \frac{1}{4n}\sum c_\sigma^2\right) \end{aligned}$$

in the worst case.

Now, consider a set of DNA sequences as described above. We can calculate each of the pattern frequencies $\bar{p}_\sigma$ by just going through the sequence and counting the number of times we see the pattern σ. We evaluate the linear invariant by plugging in the pattern frequencies $\bar{p}_\sigma$ to obtain

$$x = f((\bar{p}_\sigma)).$$

If the DNA sequences come from the tree topology and model of DNA sequence evolution associated with f, then x will come from the distribution $N\left(0, \frac{1}{4n}\sum c_\sigma^2\right)$. We want to calculate the probability that x does indeed come from this distribution. Approximately 95% of the samples drawn from $f((X_\sigma))$'s probability distribution will be within two standard deviations of the mean (which is 0 in this case). Thus, if

$$\begin{aligned} |x| &> 2\sqrt{\frac{1}{4n}\sum c_\sigma^2} \\ &= \sqrt{\frac{1}{n}\sum c_\sigma^2} \end{aligned}$$

then we can be 95% sure that the sequences do not come from the tree topology and model of DNA sequence evolution associated with f.

Unfortunately, the invariants in probability coordinates become extremely large and it becomes prohibitive to write them down and evaluate them. Furthermore, linear invariants are relatively rare and do not have the distinguishing power as the higher-degree invariants. Thus, it is more useful to consider the invariants in Fourier coordinates.

Invariants in Fourier Coordinates

Now, we will look how the invariants behave in Fourier coordinates. For now we will just consider the case when the model of DNA sequence evolution is an abelian group-based model. In that case, for every pattern τ we will define a new quantity q_τ that is obtained by applying the Fourier transform $\mathcal{F}$ to the p_σ's at a representation character χ_τ where χ_τ is the image of τ under the natural isomorphism between G and $\hat{G}$.

Example. Suppose we are considering 2 sequences of 0s and 1s. Let

$$\mathbf{p} = (p_{00}\, p_{01}\, p_{10}\, p_{11})^T$$

and

$$\mathbf{q} = (q_{00}\, q_{01}\, q_{10}\, q_{11})^T.$$

Then,

$$\begin{aligned} \mathbf{q} &= \mathcal{F}(\mathbf{p}) \\ &= \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{pmatrix}\mathbf{p} \end{aligned}$$

Similarly, we define $\bar{\mathbf{q}} = \mathcal{F}(\bar{\mathbf{p}})$. Since the invariants in Fourier coordinates will be polynomials of indeterminates corresponding to q_τ, we must look at the distributions of each of the q_τ's.

Multivariate Normality

Multivariate normal distributions have the convenient property that every linear combination of the individual marginal distributions is normally distributed. Because every $\bar{q}_\tau$ will be a linear combination of the $\bar{p}_\sigma$'s, we will want the p_σ's to be distributed as a multivariate normal distribution.

As each of the marginal binomial distributions converge to normal distributions, the multinomial distribution converges to the multivariate normal distribution. Because normal distributions are extremely well-studied and their statistical properties are often well-behaved, it will be useful to approximate the multinomial distribution by its associated multivariate normal distribution. Unfortunately, I was not able to find specific rates of convergence to the multivariate normal. The closest paper I was able to find was Andrew Carter's "Deficiency Distance between Multinomial and Multivariate Normal Experiments" (2002), but that appears to not quite be what we need in addition to imposing "smoothness" constraints on the multinomial distribution probabilities.

Therefore, if we want to have sense for how the $\bar{q}_\tau$'s behave, we can apply tests of multivariate normality to multinomial distributions to see how many samples need to be drawn for the approximation to hold.

Tests of Multivariate Normality

In their paper "A New Test for Multivariate Normality" Székely and Rizzo (2005), Székely and Rizzo present a test for multivariate normality based on the Euclidean distance between samples. We will call this test the $\mathcal{E}$-test. When compared against tests such as Marida's skewness and kurtosis tests, they find the *E-test* to be a powerful competitor. They proposed the following test statistic $\mathcal{E}$ for d-variate normality:

$$\mathcal{E} = n\left(\frac{2}{n}\sum_{i=1}^{n} E\left\|y_i - Z\right\| - E\left\|Z - Z'\right\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|y_i - y_j\right\|\right)$$

where $y_1, \ldots, y_n$ is the standardized sample. The random variables Z and Z' are independent standard d-variate normal variables, and $\|\cdot\|$ represents the standard Euclidean norm. The null hypothesis for the test is that the distribution from which the samples come from is the standard d-variate normal. Thus, at a significance level of $\alpha = 0.01$, we would fail to reject the null hypothesis that our (standardized) data comes from the standard multivariate normal whenever the p-value was greater than 0.01.

An R implementation of Székely and Rizzo's $\mathcal{E}$-test can be found in the *energy* package which can be obtained from any R CRAN mirror. (R Development Core Team, 2007) For each of the sequence lengths n listed below, R was used to sample 2000 times from a multinomial distribution with 16 different "patterns" whose probabilities were uniformly distributed between 0 and 1. Due to implementation issues with the E-test in R, the multinomial samples could not be used directly; otherwise, a divide by zero error

n					
< 500	$< 2 * 10^{-16}$	$\cdots$			
500	$< 2 * 10^{-16}$	$\cdots$			
1000	0.01005	0.2513	0.05025	0.005025	0.2814
1500	.3819	0.2211	0.01508	0.2613	0.6784
> 1500	> 0.05	$\cdots$			

Table 6.1: p-values for multivariate normal $\mathcal{E}$-test.

would occur. A small error term, normally distributed with mean 0 and standard deviation .3, was added to each of the values; changing the deviation did not produce significant changes in the result of the test. Each test took approximately 1 minute to complete. The average p-values for each of the sequence lengths can be seen in Table 6.1.

If we use a confidence value of $\alpha = 0.05$, then we almost always reject the null hypothesis of the data being multivariate normal for $n < 1500$. On the other hand, for $n \geq 1500$, we fail to reject the null hypothesis. While this is mere anecdotal evidence, it suggests that sequences lengths of approximately 1500 are needed to obtain approximate multivariate normality. However, a theoretical result on the rate of convergence would be much preferred and is something worth investigating.

Assuming Multivariate Normality

Thus, assuming that the multivariate normal approximation holds, we obtain

$$\begin{aligned} \bar{q}_\tau &= \sum c_\sigma \bar{p}_\sigma \\ &\sim N\left(\sum c_\sigma \mu_\sigma, \sum c_\sigma^2 \sigma_\sigma^2\right) \\ &\sim N\left(\sum c_\sigma p_\sigma, \left(\frac{1}{4n}\right) \sum c_\sigma^2\right). \end{aligned}$$

We can always choose a Fourier Transform $\mathcal{F}$ so that each row has length 1. This implies that $\sum c_\sigma^2 = 1$. Thus,

$$\begin{aligned} \bar{q}_\tau &\sim N\left(\sum c_\sigma p_\sigma, \frac{1}{4n}\right) \\ &\sim N\left(q_\tau, \frac{1}{4n}\right) \end{aligned}$$

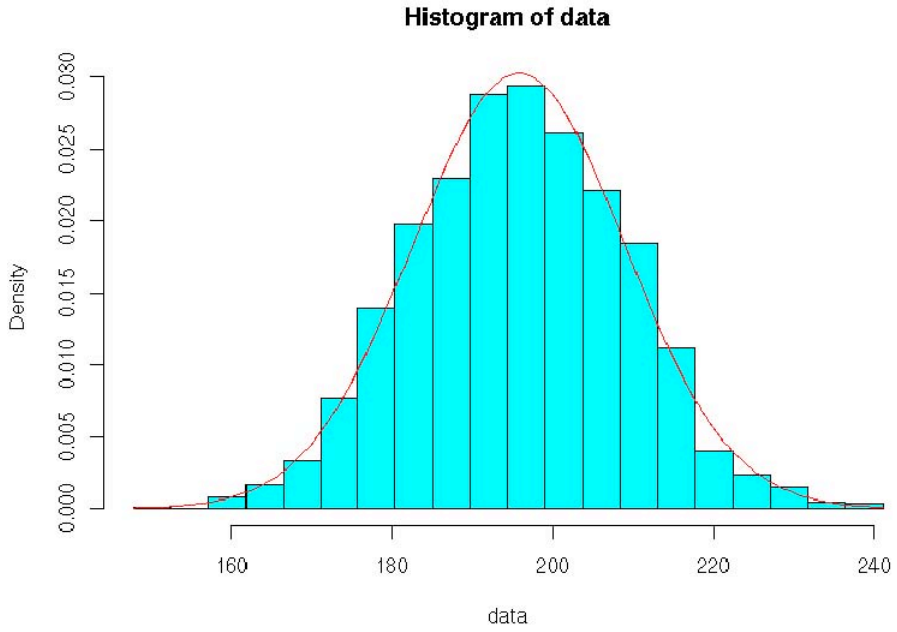


Figure 6.1: A histogram for one of the "pattern frequencies" when $n = 2500$. Plotted with the histogram is the curve for a normal distribution with the same mean and standard deviation as the data.

Notice that the variance of $\bar{q}_\tau$ is inversely proportional to the length of the sequence. This highlights the fact that the technique of using algebraic invariants for tree reconstruction improves as the length of the sequences increase.

In the Fourier coordinates, we have the nice property that all invariants will be binomials. Suppose we have a monomial r of the $\bar{q}_\tau$'s that has degree $d = \sum a_\tau$. Then, we must look at the distribution of r. First, we will compute the covariance between two arbitrary $\bar{q}_\tau$'s:

$$\begin{aligned}
\operatorname{cov}(\bar{q}_{\tau_1}, \bar{q}_{\tau_2}) &= E\left[\bar{q}_{\tau_1}\bar{q}_{\tau_2}\right] - q_{\tau_1}q_{\tau_2} \\
&= E[\left(\sum c_\sigma \bar{p}_\sigma\right)\left(\sum d_\sigma \bar{p}_\sigma\right)] \\
&= \left(\sum_{\sigma_a}\sum_{\sigma_b} c_{\sigma_a} d_{\sigma_b} E\left[\bar{p}_{\sigma_a}\bar{p}_{\sigma_b}\right]\right) - \left(\sum c_\sigma p_\sigma\right)\left(\sum d_\sigma p_\sigma\right) \\
&= \left(\sum_{\sigma} c_\sigma d_\sigma E\left[\bar{p}_\sigma \bar{p}_\sigma\right] - c_\sigma d_\sigma p_\sigma p_\sigma\right) \\
&\quad - \left(\sum_{\sigma_a \neq \sigma_b} c_{\sigma_a} d_{\sigma_b} E\left[\bar{p}_{\sigma_a}\bar{p}_{\sigma_b}\right] - c_{\sigma_a} d_{\sigma_b} p_{\sigma_a} p_{\sigma_b}\right) \\
&= \sum_\sigma c_\sigma d_\sigma \operatorname{var}(\bar{p}_\sigma) \\
&= \sum_\sigma c_\sigma d_\sigma \sigma_\sigma^2 \\
&\leq \sum_\sigma c_\sigma d_\sigma \left(\frac{1}{4n}\right) \\
&= \frac{\ell^m}{4n}\sum_\sigma c_\sigma d_\sigma.
\end{aligned}$$

Note that if $\tau = \tau_1 = \tau_2$ then we have

$$\operatorname{cov}(\bar{q}_\tau, \bar{q}_\tau) \leq \frac{|G|}{4n}$$

since $|G| = \ell^m$ and $\sum_\sigma c_\sigma^2 = 1$. In the case where $\tau_1 \neq \tau_2$, then

$$\operatorname{cov}(\bar{q}_\tau, \bar{q}_\tau) \leq 0$$

since the rows of the Fourier transform matrix are orthogonal, that is $\sum c_\sigma d_\sigma = 0$.

Now, we consider $r = \prod \bar{q}_\tau^{a_\tau}$.

$$\begin{aligned} E[r] &= E\left[\prod \bar{q}_\tau^{a_\tau}\right] \\ &= \prod E\left[\bar{q}_\tau^{a_\tau}\right] \\ &= \prod E\left[\bar{q}_\tau\right]^{a_\tau}. \end{aligned}$$

If $a_\tau > 1$ for any τ, then we are in trouble since (in general) $E\left[X^{a_\tau}\right] \neq E[X]^{a_\tau}$. Luckily, all of the invariants (at least for abelian groups) in Fourier coordinates are square-free. Thus, we will have that $E\left[\bar{q}_\tau\right]^{a_\tau} = q_\tau^{a_\tau}$ for all τ. Then, we obtain

$$E[r] = \prod q_\tau^{a_\tau}.$$

Now, suppose we have an invariant $f((x_\tau)) = r_1((x_\tau)) - r_0((x_\tau))$ where r_1 and r_0 are monomials. Then,

$$\begin{aligned} E\left[f((\bar{q}_\tau))\right] &= E\left[r_1((\bar{q}_\tau)) - r_0((\bar{q}_\tau))\right] \\ &= E\left[r_1((\bar{q}_\tau))\right] - E\left[r_0((\bar{q}_\tau))\right] \\ &= \prod q_\tau^{a_\tau} - \prod q_\tau^{b\tau} \\ &= r_1((q_\tau)) - r_0((q_\tau)) \\ &= f((q_\tau)) \\ &= 0. \end{aligned}$$

Thus, when we plug in the transformed pattern frequencies into the Fourier coordinate invariants, we should "expect" to get 0.

The variance is a bit more difficult to calculate. For each term in the binomial, we have a product of normally distributed random variables. Additionally, since the binomials are square-free, it is a product of independent normally distributed random variables. There are two promising methods I've found for understanding the distribution of a product of random variables: a 1947 result by Aroian and the use of the Mellin transform.

Approximating the Distribution for Sums of Products of Normal Variables

To begin, we will limit ourselves to quadratic invariants. Then we can rely on some previous research – "Approximating the Distribution for Sums of Products of Normal Variables" by Ware and Lad (2003). They present three different methods for constructing an approximation to the p.d.f. of

a product of two normal random variables: a numerical integration procedure in MATLAB, a Monte Carlo construction, and an approximation to the analytic result using the Normal distribution. I want to look at the approximation using a normal distribution.

Let $Y = X_1 X_2$ be the product of two normally distributed random variables with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. We can approximate of the p.d.f. $f(y)$ of Y by calculating the first two moments of Y and using a normal distribution with those moments. In Craig's 1936 paper "On the frequency function of xy" (1936), he determined the algebraic form of the moment-generating function for a product of two random variables. Aroian (1947) showed that the product is asymptotically normal as either $\delta_1 = \mu_1/\sigma_1$ or $\delta_1 = \mu_1/\sigma_1$ approach infinity. Aroian et al. (1978) considered six different cases depending on what is known about the parameters δ_1, δ_2, and ρ, the correlation coefficient between X_1 and X_2.

I will consider the case when $\rho = 0$, that is when we are looking at the product of two independent normally distributed random variables. As we have seen above, this is true for $q_{\tau_1}$and q_{τ_2} when $\tau_1 \neq \tau_2$. The moment generating function for Y is given by

$$M_Y(t) = \left(\sqrt{1-\sigma_1^2\sigma_2^2t^2}\right) e^{\frac{\mu_1\mu_2 t + \frac{1}{2}(\mu_1^2\sigma_1^2+\mu_2^2\sigma_2^2)t^2}{1-\sigma_1^2\sigma_2^2t^2}}.$$

The mean, variance, and skewness are given by

$$\begin{aligned} E(Y) &= \mu_1\mu_2 \\ V(Y) &= \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2 \\ \alpha_3(Y) &= \frac{6\mu_1\mu_2\sigma_1^2\sigma_2^2}{\left(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2\right)^{3/2}}. \end{aligned}$$

These moments can also be found by noticing $E[Y^r] = E[X_1^r]\,E[X_2^r]$ and using the moments of the normal distribution. Although Ware and Lad (2003) only consider products of two normally distributed random variables, the above technique will hold for products of any number of *independent* normally distributed random variables. For example, consider the

product $Z = X_1 X_2 \cdots X_n$. Then,

$$\begin{aligned} E[Z] &= E[X_1] E[X_1] \cdots E[X_n] \\ &= \mu_1 \mu_2 \cdots \mu_n \\ V[Z] &= E\left[Z^2\right] - E[Z]^2 \\ &= E\left[X_1^2\right] E\left[X_1^2\right] \cdots E\left[X_n^2\right] \\ &= \prod_{i=1}^{n} \left(\mu_i + \sigma_i^2\right) - \prod_{i=1}^{n} \mu_i. \end{aligned}$$

From above, we see that

$$\begin{aligned} V(Y) &= \mu_1^2 \sigma_2^2 + \mu_2^2 \sigma_1^2 + \sigma_1^2 \sigma_2^2 \\ &= \sigma_1^2 \sigma_2^2 \left(1 + \delta_1^2 + \delta_2^2\right). \end{aligned}$$

Thus, as δ_1 and δ_2 increase, the distribution of Y tends toward

$$N\left(\mu_1 \mu_2, \mu_1^2 \sigma_2^2 + \mu_2^2 \sigma_1^2\right)$$

since the normal distribution is exactly specified by its first two moments.

The quality of this approximation depends on the skewness $\alpha_3(Y)$. Because normal distributions have zero skewness, the approximation is better with small $\alpha_3(Y)$ and worse with large $\alpha_3(Y)$. The skewness of Y is largest when $\mu_1 = \sigma_1$ and $\mu_2 = \sigma_2$.

We will now look into the the values of $\delta_\tau = \mu_\tau / \sigma_\tau$ for our independent, normally distributed random variables $\bar{q}_\tau$. First off, we know that

$$\begin{aligned} \mu_\tau &= q_\tau \\ &= \sum_\sigma c_\sigma p_\sigma. \end{aligned}$$

Thus, μ_τ is a weighted average of the c_σ's since $\sum_\sigma p_\sigma = 1$. Furthermore, since $\sum_\sigma c_\sigma^2 = 1$, we have that $|\mu_\tau| \leq 1$. Also, we know that $\sigma_\tau \leq \frac{1}{4n}$ for all τ. As long as μ_τ is not arbitrarily close to zero, then we can obtain approximate normality as $n \to \infty$. For example, if $n = 250$ and $\mu_2 = \mu_{2_1} = 0.1$, then

$$\begin{aligned} \sigma_1 = \sigma_2 &\leq 0.001 \\ \delta_1 = \delta_2 &\geq 100 \\ \alpha_3(Y) &= \frac{6 \mu_1 \mu_2 \sigma_1^2 \sigma_2^2}{\left(\mu_1^2 \sigma_2^2 + \mu_2^2 \sigma_1^2 + \sigma_1^2 \sigma_2^2\right)^{3/2}} \\ &\leq 0.0212. \end{aligned}$$

With this skewness, we could relatively accurately approximate the distribution of a product of two independent normally distributed random variables as a normal distribution. This will require investigation into the values of p_σ obtained for "sane" model parameters. I suspect this will be the case.

At the end of his 1947 paper, Aroian stated that his result on the product of two normal distributions can be generalized to the product any number of random normal variables that come from a multivariate normal distribution. More concretely, if $(X_1, X_2, \ldots, X_n)$ form a joint normal distribution then,

$$\prod_{i=1}^{k} X_{a_i}$$

can be approximated by a normal distribution where $1 \leq a_1 < a_2 < \cdots < a_k \leq n$ and $k \leq n$. Aroian stated that this more general result would be proved in a later paper, but I was unable to find such a paper. A rigorous proof of this result would be beneficial in understanding the statistical behaviour of the invariants. Notice that the square-free nature of the polynomials in Fourier coordinates come into play here. If the products we're considering are not square-free, the analysis becomes much more difficult.

Mellin Transform

The following section follows closely Lomnicki's paper "On the Distribution of Products of Random Variables" (1967). Let $f(x)$ be a function. Then, then Mellin transform

$$\begin{aligned} M\{f(x)|s\} &= E\left[x^{s-1}\right] \\ &= \int_0^\infty x^{s-1} f(x) dx. \end{aligned}$$

Under suitable conditions (*need to look these up*), there in an inversion integral when we consider $M\{f(x)|s\}$ as a function of the complex variable s:

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s} M\{f(x)|s\}\, ds.$$

The Mellin convolution of two functions $f_1(x)$ and $f_2(x)$ is defined as

$$g(x) = \int_0^\infty \frac{1}{y} f_2\left(\frac{x}{y}\right) f_1(y) dy$$

which is also the p.d.f. $h_2(x)$ of the product $x = x_1 x_2$ of two independent positive, random variables with p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$. We also have

$$M\{h_2(x)|s\} = M\{f_1(x_1)|s\} M\{f_2(x_2)|s\}.$$

This can be extended for the p.d.f. $h_n(x)$ of the product $x = x_1 \cdots x_n$ of n independent, positive random variables:

$$M\{h_n(x)|s\} = \prod_{i=1}^{n} M\{f_i(x_i)|s\}.$$

We can treat the more general problem with random variables that may take on both positive and negative values by decomposing each p.d.f. $f_i(x_i)$ into two positive components with disjoint support:

$$f_i(x_i) = f_i^-(x_i) + f_i^+(x_i)$$

where $f_i^+ = f_i$ on the interval $[0, \infty)$, $f_i^- = f_i$ on the interval $(-\infty, 0]$, and both are identically 0 everywhere else. Then, we can get the p.d.f. for

$$h_n(x) = h_n^-(x) + h_n^+(x)$$

whose components are defined by

$$\begin{aligned} M\{h_n^-(-x)|s\} &= M\{f_n^+(x)|s\} \cdot M\{h_{n-1}^-(-x)|s\} \\ &\quad + M\{f_n^-(-x)|s\} \cdot M\{h_{n-1}^+(x)|s\} \\ M\{h_n^+(x)|s\} &= M\{f_n^+(x)|s\} \cdot M\{h_{n-1}^+(x)|s\} \\ &\quad + M\{f_n^-(-x)|s\} \cdot M\{h_{n-1}^-(-x)|s\}. \end{aligned}$$

Since we will be considering normal distributions, we will have that

$$f_i^-(-x_i) = f_i^+(x_i)$$

for all $i = 1, \ldots, n$. Thus, we get that

$$f_n^-(-x) = h_n^+(x).$$

In the nice case where $f_i = f$ for all $i = 1, \ldots, n$ then

$$\begin{aligned} M\{h_n^+(x)|s\} &= M\{h_n^-(-x)|s\} \\ &= 2^{n-1} M\{f^+(x)|s\}^n. \end{aligned}$$

Unfortunately, the Mellin transform of the p.d.f. of a Gaussian distribution cannot be written in closed form. In Springer and Thompson, they numerically compute values for the p.d.f. of a product of identical Gaussians. In our case, we have product of nonidentical Gaussians. There is more work left to be done to determine how best to apply the Mellin transform in our case.

Appendix A

Representations of the Symmetric Group

In this appendix, Young's orthogonal representations for the generators of the S_2, S_3, and S_4 are given. Each of the partitions whose Young diagrams are given below correspond to an irreducible representation. In order to obtain a representation for a generic element of one of the symmetric groups:

1. Write the permutation as a product of the transpositions $t_n = (n-1\ n)$.
2. Multiply the corresponding matrices together.

S_2

$$() \mapsto (1) \quad (12) \mapsto (1)$$

$$() \mapsto (1) \quad (12) \mapsto (-1)$$

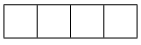
S_3

$$() \mapsto (1) \quad (12) \mapsto (1) \quad (23) \mapsto (1)$$

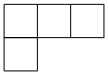
$$() \mapsto \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \quad (12) \mapsto \begin{pmatrix} -1 & \\ & 1 \end{pmatrix} \quad (23) \mapsto \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$

$$() \mapsto (1) \quad (12) \mapsto (-1) \quad (23) \mapsto (-1)$$

S_4



$$() \mapsto (1) \quad (12) \mapsto (1) \quad (23) \mapsto (1) \quad (34) \mapsto (1)$$

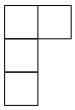


$$() \mapsto \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \qquad (12) \mapsto \begin{pmatrix} -1 & & \\ & 1 & \\ & & 1 \end{pmatrix}$$

$$(23) \mapsto \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} & \\ & & 1 \end{pmatrix} \qquad (34) \mapsto \begin{pmatrix} 1 & & \\ & \frac{1}{3} & \frac{2\sqrt{2}}{3} \\ & \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{pmatrix}$$

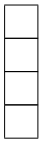
$$() \mapsto \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \qquad (12) \mapsto \begin{pmatrix} -1 & \\ & 1 \end{pmatrix}$$

$$(23) \mapsto \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad (34) \mapsto \begin{pmatrix} -1 & \\ & 1 \end{pmatrix}$$



$$() \mapsto \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \qquad (12) \mapsto \begin{pmatrix} -1 & & \\ & -1 & \\ & & 1 \end{pmatrix}$$

$$(23) \mapsto \begin{pmatrix} -1 & & \\ & \frac{1}{2} & \frac{\sqrt{3}}{2} \\ & \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \qquad (34) \mapsto \begin{pmatrix} \frac{1}{3} & \frac{2\sqrt{2}}{3} & \\ \frac{2\sqrt{2}}{3} & -\frac{1}{3} & \\ & & -1 \end{pmatrix}$$



$$() \mapsto (1) \quad (12) \mapsto (-1) \quad (23) \mapsto (-1) \quad (34) \mapsto (-1)$$

Bibliography

Allman, Elizabeth, and John Rhodes. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences* 186:113–144.

Aroian, Leo A. 1947. The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics* 18(2):265–271. URL http://links.jstor.org/sici?sici=0003-4851%28194706%2918%3A2%3C265%3ATPFOTP%3E2.0.CO%3B2-O.

Aroian, Leo A., Vidya S. Taneja, and Larry W. Cornwell. 1978. Mathematical forms of the distribution of the product of two normal variables. *Comm Statist A—Theory Methods* 7(2):165–172.

Carter, Andrew V. 2002. Deficiency distance between multinomial and multivariate normal experiments. *The Annals of Statistics* 30(3):708–730. URL http://links.jstor.org/sici?sici=0090-5364%28200206%2930%3A3%3C708%3ADDBMAM%3E2.0.CO%3B2-I.

Cavender, James, and Joseph Felsenstein. 1987. Invariants of phylogenies: a simple case with discrete states. *Journal of Classification* 4:57–51.

Craig, Cecil C. 1936. On the frequency function of xy. *The Annals of Mathematical Statistics* 7(1):1–15. URL http://links.jstor.org/sici?sici=0003-4851%28193603%297%3A1%3C1%3AOTFFO%3E2.0.CO%3B2-U.

Doyle, James A. 1998. Phylogeny of vascular plants. *AnnuRev Ecol Syst* 29:567–569.

Evans, Steven, and Terry Speed. 1993. Invariants of some probability models used in phylogenetic inference. *The Annals of Statistics* 21:355–377.

Evans, Steven N. 2004. Fourier analysis and phylogenetic trees. In *Modern Signal Processing (Lecture notes from an MSRI Summer School)*, eds. D. Healy Jr. and D. Rockmore. Cambridge University Press.

Felsenstein, Jospeh. 2003. *Inferring Phylogenies*. Sinauer Associates.

Helton, J. William, and Mark Stankus. 1999. Computer assistance for 'discovering' formulas in system engineering and operator theory. URL `citeseer.ist.psu.edu/helton99computer.html`.

James, Gordon, and Adalbert Kerber. 1984. *The Representation Theory of the Symmetric Group*. Cambridge University Press.

Kimura, Motoo. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America* 78:454–458.

Lake, James. 1987. A rate-independent technique for the analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution* 4:167–191.

Lomnicki, Z. A. 1967. On the distribution of products of random variables. *Journal of the Royal Statistical Society Series B (Methodological)* 29(3):513–524. URL `http://links.jstor.org/sici?sici=0035-9246%281967%2929%3A3%3C513%3AOTDOPO%3E2.0.CO%3B2-V`.

Mora, Teo. 1994. An introduction to commutative and noncommutative Gröbner bases. In *Selected papers of the second international colloquium on Words, languages and combinatorics*, 131–173. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V. doi: http://dx.doi.org/10.1016/0304-3975(94)90283-6.

Pachter, L., and B. Sturmfels. 2005. *Algebraic Statistics for Computational Biology*. New York, NY, USA: Cambridge University Press.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Serre, Jean-Pierre. 1977. *Linear Representations of Finite Groups*. Springer.

Sturmfels, Bernd, and Seth Sullivant. 2005. Toric ideals of phylogenetic invariants. *Journal of Computational Biology* 12:204.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. In *Molecular systematics (2nd ed.)*, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 407–514. Sunderland, Massachusetts: Sinauer Associates, Inc.

Székely, Gábor J., and Maria L. Rizzo. 2005. A new test for multivariate normality. *J Multivar Anal* 93(1):58–80. doi:http://dx.doi.org/10.1016/j.jmva.2003.12.002.

Székely, L.A., M.A. Steel, and P.L. Erdös. 1993. Fourier calculus on evolutionary trees. *Advances in Applied Mathematics* 14:200–216.

Ware, Robert, and Frank Lad. 2003. Approximating the distribution for sums of products of normal variables. URL `http://www.math.canterbury.ac.nz/php/research/abstracts/abstract(2003-15).php`.