2019

STOCHASTIC OPTIMIZATION FOR TROPICAL PRINCIPAL COMPONENT ANALYSIS OVER TREE SPACES

Page, Robert L.

Monterey, CA; Naval Postgraduate School

http://hdl.handle.net/10945/62731

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

STOCHASTIC OPTIMIZATION FOR TROPICAL PRINCIPAL COMPONENT ANALYSIS OVER TREE SPACES

by

Robert L. Page

June 2019

Thesis Advisor:	Ruriko Yoshida
Second Reader:	Michael P. Atkinson

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2019	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
STOCHASTIC OPTIMIZATION FOR TROPICAL PRINCIPAL COMPONENT ANALYSIS OVER TREE SPACES

5. FUNDING NUMBERS

6. AUTHOR(S) Robert L. Page

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

A known challenge in the rapidly growing area of phylogenomics is the lack of tools to analyze the large volume of genome data. Genomic data includes information on the evolution, structure and mapping of genomes. Phylogenetic trees are branching diagrams that show the evolutionary history of species and their genes. Gene trees show the evolutionary history of a particular gene. To analyze evolutionary history from genomic data, we reduce the dimensionality of gene trees, overcoming high dimensional analytical challenges. Through the vectorization of pairwise distances between each combination of two leaves within a phylogenetic tree, we utilize a tropical principle component analysis: a principal component analysis (PCA) in terms of a tropical metric. We project gene trees onto a two-dimensional space using a tropical PCA, a tropical convex hull that minimizes the sum of residuals between each gene tree in the dataset and its projection onto the tropical convex hull over the tree space, which is the set of all possible gene trees. Since computing a tropical PCA for the given dataset is computationally time intensive, we implement a Markov Chain Monte Carlo Metropolis-Hastings algorithm to effectively and efficiently estimate the tropical PCA. Utilizing simulation and real-world data, we implement our tropical PCA algorithm and visualize the results in two-dimensional plots, the results of which look promising and demonstrate our algorithm's strengths.

14. SUBJECT TERMS
phylogenomics, tropical principle component analysis, evolutionary trees, tree space, genomics, tropical geometry, Markov chain Monte Carlo

15. NUMBER OF PAGES
97

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

STOCHASTIC OPTIMIZATION FOR TROPICAL PRINCIPAL COMPONENT ANALYSIS OVER TREE SPACES

Robert L. Page
Major, United States Army
BS, Virginia Military Institute, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2019

Approved by: Ruriko Yoshida
Advisor

Michael P. Atkinson
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

A known challenge in the rapidly growing area of phylogenomics is the lack of tools to analyze the large volume of genome data. Genomic data includes information on the evolution, structure and mapping of genomes. Phylogenetic trees are branching diagrams that show the evolutionary history of species and their genes. Gene trees show the evolutionary history of a particular gene. To analyze evolutionary history from genomic data, we reduce the dimensionality of gene trees, overcoming high dimensional analytical challenges. Through the vectorization of pairwise distances between each combination of two leaves within a phylogenetic tree, we utilize a tropical principle component analysis: a principal component analysis (PCA) in terms of a tropical metric. We project gene trees onto a two-dimensional space using a tropical PCA, a tropical convex hull that minimizes the sum of residuals between each gene tree in the dataset and its projection onto the tropical convex hull over the tree space, which is the set of all possible gene trees. Since computing a tropical PCA for the given dataset is computationally time intensive, we implement a Markov Chain Monte Carlo Metropolis-Hastings algorithm to effectively and efficiently estimate the tropical PCA. Utilizing simulation and real-world data, we implement our tropical PCA algorithm and visualize the results in two-dimensional plots, the results of which look promising and demonstrate our algorithm’s strengths.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

BHV	Billera, Holmes and Vogtmann
DNA	Deoxyribonucleic acid
GTR	Generalized time-reversible
LFM	locus of the Fréchet mean
MCMC	Markov chain Monte Carlo
PCA	Principle component analysis
UPGMA	Unweighted pair group method using arithmetic averages

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

With advancements in technology, we can generate large volumes of genetic data relatively cheaply and quickly. The goal is to analyze the evolutionary history of species and their genes from this trove of genetic data. The high dimensionality and complex structure of this data creates significant analysis challenges. The most convenient form of this data is branching diagrams known as phylogenetic trees. The vectorization of these phylogenetic trees allows for transformations and further analysis.

Principal Component Analysis (PCA) is one of the most popular methods for dimension reduction and enables the two-dimensional visualization of gene trees. Two-dimensional visualization of gene trees facilitates quick interpretation and assessment of the results. Classical PCA is a statistical method that takes data points in a high-dimensional Euclidean space and represents them in a lower-dimensional plane in such a way that the residual sum of squares is minimized. We cannot directly apply classical PCA to a set of phylogenetic trees because the space of phylogenetic trees is not Euclidean, an assumption necessary for classical PCA.

This thesis defines tree space as the space of all phylogenetic trees containing the same number of leaves m. Tree space is the union of lower dimensional polyhedral cones in $\mathbb{R}^{\binom{m}{2}}$. Therefore we require a dimension reduction technique that works in tree space, one that we can apply to a set of gene trees.

In 2019, Ruriko Yoshida, Leon Zhang, and Xu Zhang define a tropical PCA under the tropical metric using max-plus tropical arithmetic (Yoshida et al. 2019). Utilizing the best-fit tropical polytope, having a fixed number of vertices closest to the data points, allows dimension reduction of phylogenetic trees. In the paper, (Yoshida et al. 2019) formulate a tropical PCA as the best-fit tropical polytope to an input data over the tropical projective torus and showing that this can be written as a mixed-integer programming problem. They apply the tropical PCA defined as a tropical polytope to datasets consisting of collections of phylogenetic trees.

Applying Yoshida, Zhang and Zhang's tropical PCA method to a set of phylogenetic trees requires the estimation of the optimal tropical PCA base. Enumerating all possible

combinations of s-subset of trees from the data to compute the $(s-1)$th order principal components is not effective or efficient for large volumes of genetic data. Therefore, at this moment, there is no method to compute a tropical PCA over the space of phylogenetic trees.

This thesis proposes a heuristic method to compute a tropical PCA as a tropical polytope over a space of rooted phylogenetic trees with a fixed number of leaves. More specifically, we propose a stochastic computation using a Metropolis-Hastings algorithm to estimate a tropical PCA over the space of *equidistant phylogenetic trees* with the number of leaves m.

Utilizing a Markov chain Monte Carlo (MCMC) approach and the phylogenetic software program *Mesquite*, we conduct simulation experiments to visualize and analyze the results of tropical PCA on the mixture of two (or eight) coalescent distributions on $m = 10$ leaves. With the goal of effectively visualizing gene tree clusters by distribution, we illustrate the results in easy to understand two-dimensional plots. We conduct further analysis by computing R^2 values and compare with Tom Nye's locus of the Fréchet mean PCA technique on the Billera, Holmes, and Vogtmann (BHV) tree space. Figure 1 illustrates the methodology and workflow for this thesis.

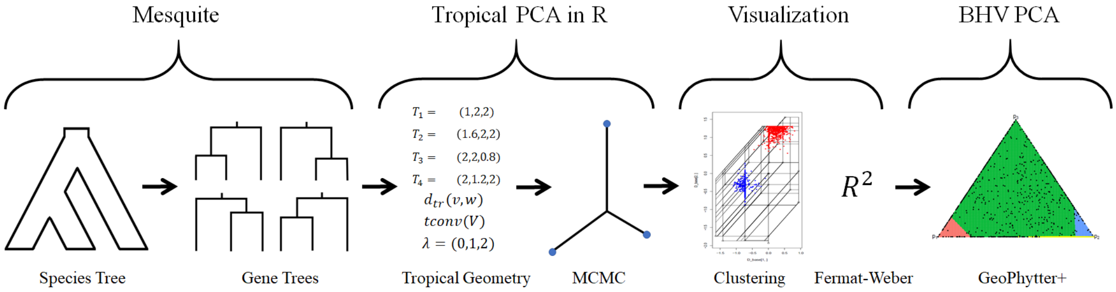


Figure 1. Thesis Methodology Diagram

Our simulation experiments confirm the ability of our Metropolis-Hastings MCMC tropical PCA approach to effectively and efficiently reduce high dimensional gene trees and visualize clustering in two-dimensional plots. We end with applications of our novel MCMC approach, which we implement on two empirical datasets, Apicomplexa and African coelacanth genomes.

Our results clearly demonstrate the effectiveness and efficiency of our MCMC tropical PCA approach. Therefore, we conclude that gene trees can be clustered by distribution (species tree) using our tropical PCA, to gain insights on the inferred species topology.

List of References

Yoshida R, Zhang L, Zhang X (2019) Tropical principal component analysis and its application to phylogenetics. *Bulletin of Mathematical Biology* 81(2):568–597.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I am grateful to Professors Ruriko (Rudy) Yoshida and Michael Atkinson for their guidance, effort, and motivation throughout my thesis journey. Professor Yoshida, thank you for sharing your work with me. Your kindness, collaboration and support throughout this process had an enormous and lasting impact. Professor Atkinson, thank you for graciously dedicating your time and effort to ensuring my ideas were well articulated.

I am forever grateful to my family for their love and support. They have always encouraged me to try my hardest and are extremely proud of me regardless of the outcome.

I am appreciative and extremely fortunate to have worked with such great people at the US Army War College. Thank you for inspiring me to attend NPS and for the continued mentorship.

I am particularly thankful to Kelsey, for much more than the late nights and early mornings helping me with my work. Thank you for everything.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Darwinism suggests that every species on Earth evolved from a common ancestor (Darwin 1859). If Darwin is correct, then all species share a common thread dating back to a single cell organism. Where a species currently is in their evolutionary history is known; however, their evolutionary path that led them to the present remains a mystery. If life on earth evolved from a common ancestor, then the comparison of a group of species reveals the similarities and differences of their evolutionary journey. Using the present knowledge of a species as a starting point, we can infer the past to uncover possible insights about a species' evolutionary history.

1.1 Introduction to Genetics

Evolutionary analysis often begins with the examination of genetic data. Genetic data is frequently shared among species with a close common ancestor. This closely related ancestor provides a link between species, which this thesis explores. All genetic data for a species is contained in its genome, which resides in most of an organism's cells. Author Matt Ridley (Figure 1.1) provides an excellent metaphor for describing the genome in his book *Genome: The Autobiography of a Species in 23 Chapters*.

> **Imagine that the genome is a book.**
> There are twenty-three chapters, called CHROMOSOMES.
> Each chapter contains several thousand stories, called GENES.
> Each story is made up of paragraphs, called EXONS, which are interrupted by advertisements called INTRONS.
> Each paragraph is made up of words, called CODONS.
> Each word is written in letters called BASES.

Figure 1.1. Genome Metaphor. (Source: Ridley 1999)

This thesis examines possible techniques for comparing books with shared stories by focusing on **genes** and **bases**. Neither the species (books) nor the genes (chapters) can be compared without examining the words comprising of bases (letters). Instead, combinations of bases (words), specifically the Deoxyribonucleic acid (DNA) are used as data.

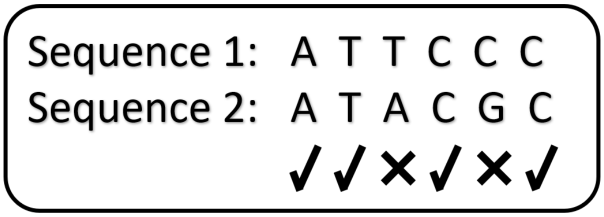
The motivation being that similar species often share genes and similar genes share bases. Comparing bases allows evolutionary insights to be gained. The challenge is that genes are made up of long chains of DNA. Ridley continues his metaphor of the genome.

> Whereas English books are written in words of variable length using twenty-six letters, genomes are written entirely in three-letter words, using only four letters: A, C, G and T (which stand for adenine, cytosine, guanine and thymine). And instead of being written on flat pages, they are written on long chains of sugar and phosphate called DNA molecules to which the bases are attached as side rungs. Each chromosome is one pair of (very) long DNA molecules. (Ridley 1999)

Our understanding of the evolutionary past may unlock insights to the future. For example, by examining the genetic data of apicomplexa, the parasitic organism which includes malaria, evolutionary inferential insights may prove invaluable for developing cures or vaccines.

1.2 Pairwise Alignment

We can examine two books (species) with a shared story (genes) and compare the words (DNA sequences) from their stories (genes). A technique for genetic comparison is to investigate the available genetic information to determine similarity. At the most basic level, we take two DNA sequences and attempt to answer if the two sequences are related (Durbin et al. 1998). Sequences may appear similar in alignment because of a commonality or simply by chance. (Durbin et al. 1998). Figure 1.2 illustrates the pairwise alignment of two example sequences.



Sequence 1: A T T C C C
Sequence 2: A T A C G C
✓ ✓ ✗ ✓ ✗ ✓

Figure 1.2. Example Sequence

Sequences 1 and 2 are aligned on top of each other in Figure 1.2, allowing us to compare the similarities and differences between the sequence characters. The check marks indicate similarities and the crosses indicate differences. Once the similarities and differences are identified, the issue is trying to determine what to do with this new information. Specifically, what DNA sequences, or portions of sequences, are being aligned and how are the alignments being compared and relationships measured. Ultimately, the relationship between sequence 1 and 2 in Figure 1.2 depends on **how** the comparison is made.

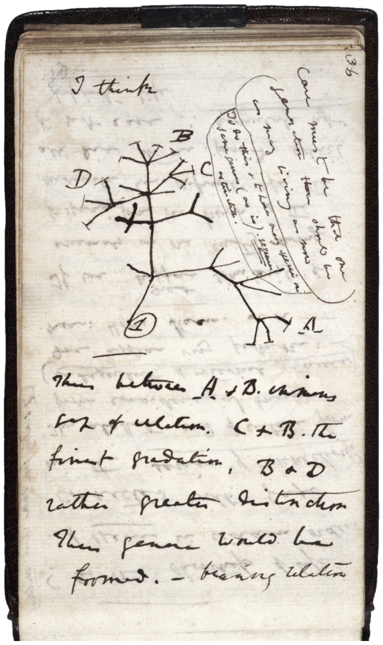
The comparisons made directly tie to the pairwise alignment assumptions and sequential analysis techniques. Chapter 2 covers a detailed look into pairwise alignments and their effects on evolutionary reconstruction. Regardless of the assumptions, the idea behind making pairwise alignments remains the same. Given DNA sequences, we want to infer how those sequences are related to each other and produce a measure (score) pertaining to that relationship, which is used as a new inferential data point.

1.3 Trees

The technique of using trees to represent the evolutionary history of a species dates back over a 180 years. In 1837, Darwin drew the first known evolutionary tree (Figure 1.3) to depict the concept of phylogeny (Darwin 1837). Written next to Darwin's evolutionary tree are key comments regarding the relationship between species, current and extinct.

Phylogenetics is the study of an organism's evolutionary relationships and changes over time. In mathematics, evolutionary relationships are typically represented by graphical structures called trees. Graphs play a fundamental role in phylogenetics (Semple and Steel 2003). Trees are graphs that have desirable structures that intuitively connect objects. The term ***tree*** is used throughout this thesis to describe different relationships among multiple organisms. A *tree* T consists of $vertices\ V$ and $edges\ E$. Vertices, or *nodes* as they are commonly referred to as, represent objects in the tree that have interrelationships we wish to visualize. Each vertex (node) in a tree represents a split or terminal end of the path. Edges, sometimes referred to as branches, represent the notion of distance. The edges link vertices together, illustrating the interrelationships of the connected objects. Figure 1.4 points out vertices and edges of an example tree.

Graphically, a tree requires the set of vertices to be nonempty and finite, and the set of edges also finite. A tree cannot have zero vertices (no objects) nor can a tree have an infinite amount of vertices. Therefore, one vertex is the smallest a tree can be, and it cannot be infinite in size. Trees are also acyclic and connected. This means that all vertices in T can be reached by taking a path of distinct vertices with no way of looping back to the start point in the graph without retracing steps. Both properties make intuitive sense and are highly desirable when representing evolutionary trees later in the thesis.


36
I think
B
C
D
1
A

Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

Figure 1.3. Darwin's 1837 Evolutionary Tree Sketch. Source: Darwin (1837).

Without labels the example tree in Figure 1.4 is not very useful. Adding appropriate labels to trees allows for further graphic proprieties to be defined and explored. Labels can be added to vertices and edges. Figure 1.5 has labeled vertices $V = \{1, 2, 3, 4, 5\}$. A vertex with *degree > 1* is an interior vertex and represents the coming together of two or more subgroups. Figure 1.5 has interior vertices $V = \{4, 5\}$. Leaves in a tree are vertices with *degree = 1* and are often referred to as tips or terminal nodes. Figure 1.5 has labeled leaves $V = \{1, 2, 3\}$. An edge is internal if both vertices are interior vertices and external if one vertex is a leaf. Edge $E = \{(3, 4)\}$, in Figure 1.5, is an external branch (edge) because vertex $V = \{3\}$ is a leaf.

Figure 1.5 is called a rooted binary tree because all the edges are splitting away from the root node (a vertex with its degree 2), $V = \{5\}$, in a specific direction (downward).

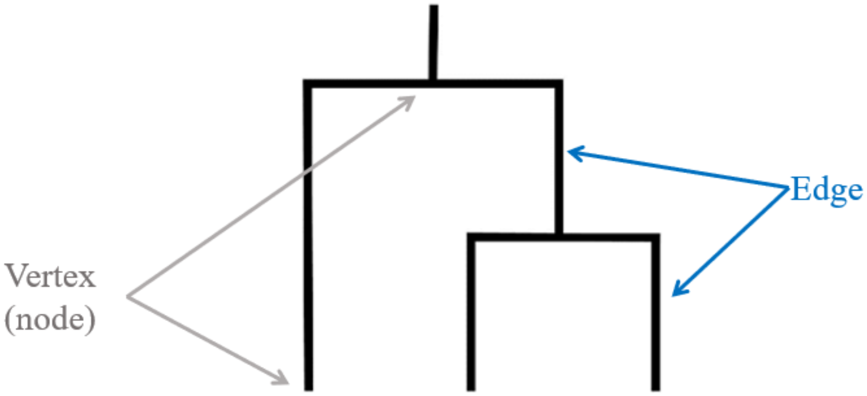


Figure 1.4. Example Tree

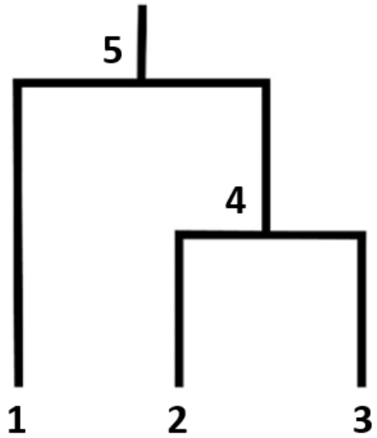


Figure 1.5. Labeled Example Tree

Additionally, all interior nodes bifurcate or split into exactly two edges. Rooted trees are a fitting representation for phylogenetic trees because the root depicts a common ancestor of the species within the tree. Trees are not always rooted or binary, but for the purposes of this thesis we will continue assuming all trees are both.

Counting nodes (vertices) and branches (edges) in a rooted tree begins with the leaves and continues upward to the root. In Figure 1.5, there are three leaves $V = \{1, 2, 3\}$ and three branches moving upward. Moving up the tree from the leaves, the branches first join at node $V = \{4\}$. Note the number of branches is now two. Every time the branches join to form a new node, the total number of branches reduces by one. As we continue upwards further, the remaining two branches join and form the root, $V = \{5\}$.

Therefore in a rooted binary tree, if there are a total of n leaves in the tree, then there are $2n - 1$ total nodes, with $n - 1$ interior nodes. A branch (edge) length represents the distance between two connected vertices. Branch lengths are labelled from vertex to vertex and referenced as d_{ij}. For example, the edge length from *node* 1 to *node 5* would be referred to as $d_{1,5}$. The following chapter discusses trees from a phylogenetic perspective.

CHAPTER 2:
Phylogenetics

Phylogenetics is the study of an organism's evolutionary journey and genetic relationships. The genomes of current species are well documented. However; reconstructing the evolutionary history of a species or organism from its genomic data is a significant challenge (Durbin et al. 1998). For example, the human genome consists of approximately 30,000 genes (Brown 2002), conversely, apicomplexa, only consists of about 250 "well-aligned" genes (Kuo et al. 2008b). Analysis is primarily conducted on inferential evolutionary (phylogenetic) trees reconstructed from known genetic data (Semple and Steel 2003).

2.1 Phylogenetic Trees

Evolutionary, or phylogenetic trees, show an organism's evolutionary relationships over time, through the use of tree diagrams. Phylogenetic trees are a subset of branching tree graphs discussed in chapter 1.3. Phylogenetic trees still consist of vertices (nodes) and edges (branches), but now the properties have specific evolutionary interpretations.

Each node (vertex) in a phylogenetic tree represents a past or present taxon or population. In Figure 2.1, an exterior vertex (leaf or tip) represents the current taxa *(*$V = \{Species1, Species2, Species3\}$). An interior vertex represents an extinct taxa ($V = \{4, 5\}$) where ancestors split into two subgroups. Vertices and edges can still be labeled; however, only the leaves or tips are labeled in a phylogenetic tree. This is due to the past taxa often being inferred and not exactly known. Vertices in phylogenetic trees can be DNA sequences, shared genes or interrelated species, depending on the context of the tree. The root ($V = \{5\}$) of the tree now represents the common ancestor of all leaves, $Species1$, $Species2$, and $Species3$.

Phylogenetic trees remain acylic and connected. Both properties are intuitive, as a species must evolve from something. Additionally, as time progresses species can only evolve forward. Edges now represent the notion of time. The distance ($d_{Species1,Species2}$) measures the dissimilarity between *Species 1* and *Species 2* with respect to time.

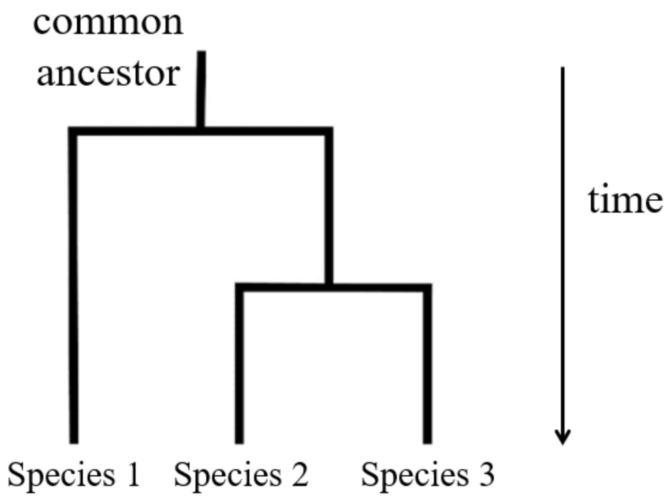


Figure 2.1. Example Rooted Binary Phylogenetic Tree

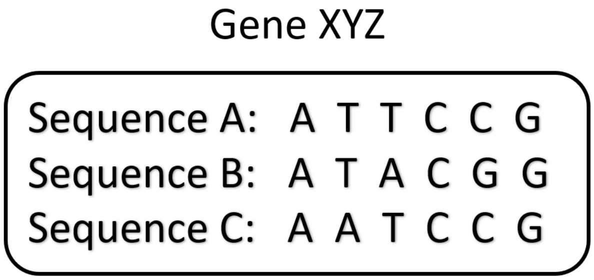
2.2 Phylogenetic Reconstruction

Phylogenetic reconstruction uses genetic data to create an inferential evolutionary (phylogenetic) tree. Cavalli and Edwards speak to the relevance of phylogenetic trees:

> Evolution can only be described in terms of the characters that are changing, and it is convenient to represent such changes in a multidimensional character-space in which each population occupies a position determined by the values of the characters it exhibits (Cavalli-Sforza and Edwards 1967).

The *changing characters* are the changes in DNA sequences. DNA sequences represent a shared gene across multiple species. Trees are excellent at representing the evolutionary changes of this shared gene through node splits and leaves.

The transformation of DNA sequences into gene trees allows shared genes across multiple species to be studied. Multiple steps and techniques are involved in the reconstruction process. There are several types of tree reconstruction methods; we will focus on pairwise distance based methods. Pairwise distance based tree reconstruction methods are intuitive and a proven technique (Cavalli-Sforza and Edwards 1967) for building phylogenetic trees. Distance-based phylogenetic tree reconstruction involves the multiple pairwise alignments of DNA sequences. These pairwise alignments examine each pair of $\{i,\ j\}$ sequences and determines the distance, d_{ij}, between each pair (Durbin et al. 1998). Hamming distance is the number of differences between two strings. The phylogenetic reconstruction example in Figure 2.2 illustrates this relatively simple distance measure.



Gene XYZ

Sequence A: A T T C C G
Sequence B: A T A C G G
Sequence C: A A T C C G

Figure 2.2. Multiple Pairwise Alignments for Gene XYZ

In Figure 2.2, DNA sequences A, B, and C represent the same shared gene (Gene XYZ) across three species (*Species A*, *Species B*, and *Species C*). Phylogenetic reconstruction uses the pairwise distances between each pair of sequences to construct a tree that infers the species structure for *Gene XYZ*. In order to reconstruct a phylogenetic tree for Gene XYZ, the Hamming distance, d_{ij}, between each sequence pair is calculated. The distances are visualized in the distance matrix Table 2.1. Distance matrices are read top to bottom and left to right. The start point, i, is represented by the vertical sequence labels (Sequence A, Sequence B, Sequence C) on the left. The endpoint, j, is the horizontal sequences labels (Sequence A, Sequence B, Sequence C) positioned on top . Distance matrices are symmetric across the zero diagonal, from top-left to bottom-right, because the distance from any object to itself is always zero. Therefore, the distance from A to B is the same as from B to A. More formally, $d_{AB} = d_{BA}$.

After the pairwise distances are measured, hierarchical clustering is required to transform the distances into a phylogenetic tree. Unweighted pair group method using arithmetic averages (UPGMA) is a tried and true hierarchical clustering procedure (Sokal and Michener 1958). The UPGMA algorithm assigns each sequence a cluster C_i. Next, the two closest clusters, according to the distance matrix, are coalesced into an ancestral node. That ancestral node is assigned a cluster and the process is iterated until only one cluster remains, the root node.

The reconstruction process is described through an example, using Gene XYZ. The first step in UPGMA is to identify the smallest (nonzero) distance. The yellow boxes, in Table 2.2, highlight the smallest (nonzero) distance.

	A	B	C
A	0	2	1
B	2	0	3
C	1	3	0

Table 2.1. Distance Matrix of Hamming Distances (d_{ij}) For Gene XYZ

	A	B	C
A	0	2	1
B	2	0	3
C	1	3	0

Table 2.2. Distance Matrix (d_{ij}) Smallest Distance

Next, we combine sequence A and C and average the distance to B (as seen in Table 2.3). The distance $d_{A,B} = 2$ and $d_{C,B} = 3$. Therefore, the distance $d_{AC,B} = 2.5$. The inferred distance from the AC immediate ancestor is 0.5, because distance $d_{A,C} = 1$, as shown in Figure 2.3. A, C is assigned a cluster and a new distance matrix is created, shown in Table 2.3.

Only two clusters remain, AC and B. They are joined at the root node, and the average distance to the root is calculated. The distance $d_{AC,B} = 2.5$ and $d_{A,C} = 1$. Therefore, the distance $d_{B,root} = 1.25$, $d_{A,root} = 1.25$, and $d_{C,root} = 1.25$, as shown in Figure 2.3. The inferred phylogenetic gene tree reconstruction for Gene XYZ is complete and depicted in Figure 2.3.

This basic example of phylogenetic reconstruction illustrates the power of trees to visualize DNA sequences. With additional genetic data we can reconstruct multiple gene trees for a species and infer evolutionary information. Mutations are changes in the current sequences of genes. Mutations of genes occur naturally, and the previous example is too basic to account for those state changes. Advanced distance based methods measure evolutionary distances and account for mutations. A classic example of mutation is natural selection Darwin (1859). Sophisticated distance based methods are necessary to overcome the challenges with reconstructing multiple sequences while accounting for mutations.

	AC	B
AC	0	2.5
B	2.5	0

Table 2.3. UPGMA (Mid-Clustering) Distance Matrix (d_{ij})

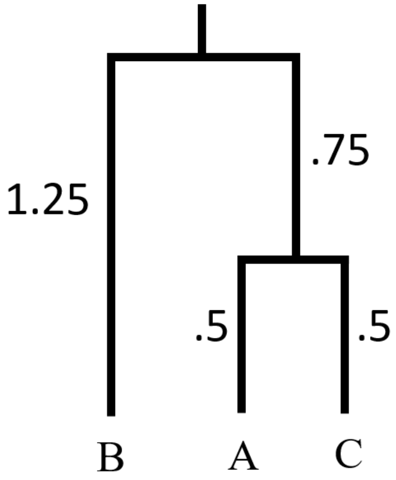


Figure 2.3. Reconstructed Equidistant Gene Tree For Gene XYZ

Substitution models are evolutionary DNA models that attempt to account for these mutations. Utilizing continuous-time Markov chains, substitution models implement a transitional rate matrix to account for the sequence mutation probabilities over time. The output, produced from the raw sequential DNA data, estimates the evolutionary distance that includes multiple mutations. The generalized time reversible model (GTR) is a widely accepted substitution model (Tavaré 1986). GTR allows for flexible parameterization of gene sequences. In Chapter 13 of *Inferring Phylogenies*, author Joseph Felsenstein does a great job detailing GTR models (Felsenstein 2004).

Phylogenetic reconstruction not only recreates a single gene tree but facilitates the reconstruction of all gene trees, for all shared genes, across multiple species, in an effort to understand the species tree. This begs the question, does possession of gene trees, for all shared genes across multiple species, allow us to infer the species trees? More formally, can we efficiently and effectively visualize gene trees clusters by distribution (species) to gain insights on species tree topology?

2.3 Tree Space

Armed with genetic reconstruction techniques, addressing the research question should be straight forward. Ideally we would leverage the totality of the reconstructed phylogenetic trees to infer a centralized location. That centralized location represents the possible species tree or topology. The issue is that phylogenetic trees exist in a geometric space that is unfamiliar. Conversely, Euclidean space is well studied and understood. A better understanding of the space where phylogenetic trees exist is necessary.

Tree space represents the geometric space where phylogenetic trees exist. Visualizing gene tree clusters by species to gain insights on phylogenomic tree topoglogies requires analysis in tree space. This thesis defines a *tree space* as the space of *ultrametrics*, where an ultrametric is defined in Section 2.4. As we will explain in Section 2.4, there is a one-to-one map between an equidistant tree and an ultrametric. Therefore, the space of all possible equidistant trees can be represented as the space of ultrametrics. Trees are equidistant if the distance from the root to each leaf is equal. This property ensures that each leaf of a tree is at the same evolutionary time.

Figure 2.4 is an equidistant gene tree for Gene123. Gene123 is a shared gene across Species A, Species B, Species C and Species D. $SeqA$ is the DNA sequence for Gene123 in Species A. The gene tree is equidistant because Sequence A, Sequence B, Sequence C and Sequence D are equal distance from the root. For example, the distance between the root of Figure 2.4 and Sequence A is one (0.25 + 0.75). The distance between the root and Sequence C is also one (0.6 + 0.4).

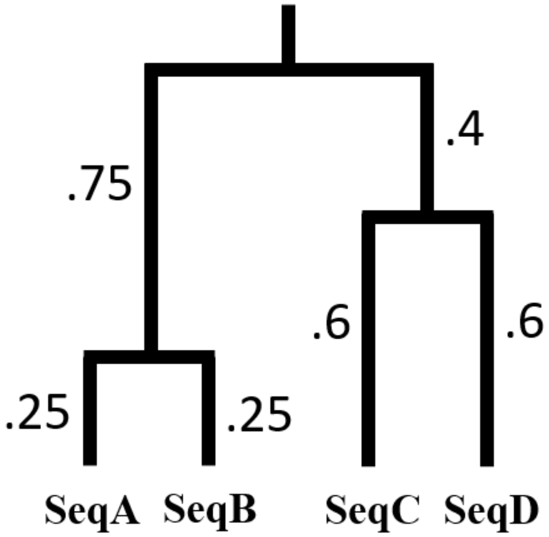


Figure 2.4. Equidistant Gene Tree For Gene123

2.4 Phylogenetic Vectorization

While tree formats are helpful for visualizing evolutionary relationships, a matrix is better for mathematical analysis. The preferred distance matrices used in phylogenetics are dissimilarity maps. A dissimilarity map (Table 2.4), shows the distance, d_{ij}, between any leaf pairs $i,\ j$. The main advantage of transforming phylogenetic trees into dissimilarity maps is that the distances d_{ij} are easily extracted.

	A	B	C	D
A	0	0.5	2	2
B	0.5	0	2	2
C	2	2	0	1.2
D	2	2	1.2	0

Table 2.4. Dissimilarity Map For Gene123 (d_{ij})

Dissimilarity maps are symmetric, all distances d_{ij} are repeated within the matrix as d_{ji}. For example, in Table 2.4 the distance from Sequence A to Sequence B, d_{AB}, is 0.5 and the distance from Sequence B to Sequence A, d_{BA}, is also 0.5. Therefore, the entire matrix is not needed. All distance data from our original phylogenetic tree, Figure 2.4, is accounted for by taking the upper (or lower) triangular of the dissimilarity map.

	A	B	C	D
A	0	0.5	2	2
B	0.5	0	2	2
C	2	2	0	1.2
D	2	2	1.2	0

Table 2.5. Upper Triangular For Gene123

In order to best utilize the information stored in the upper triangular, the distance data may be vectorized. The data is put into vector form by assigning each dissimilarity map position to a corresponding index position in the vector. To determine the length of the vector, the number of leaves (n) is multiplied by the number of leaves minus one ($n - 1$). This product is then divided by 2.

$$vectorlength = (n \cdot n - 1)/2 \tag{2.1}$$

Table 2.5 has four leaves (A, B, C, D); therefore, the length of the vector is six, $(4 \cdot 3)/2 = 6$. The order in which the data points are assigned to the vector is important and must be the same each time the process is repeated. The distances, d_{ij}, are assigned to the vector from top to bottom, left to right, starting to the right of d_{AA}. The vectorization of the dissimilarity map is shown in Figure 2.5.

$$v = \left(\underset{d_{ab}}{0.5,} \quad \underset{d_{ac}}{2,} \quad \underset{d_{ad}}{2,} \quad \underset{d_{bc}}{2,} \quad \underset{d_{bd}}{2,} \quad \underset{d_{cd}}{1.2} \right)$$

Figure 2.5. Vectorized Upper Triangular For Gene123

Once vectorized, the ultrametric condition is tested to ensure phylogenetic tree reconstruction is correct. The ultrametric condition states for all distinct i, j, k leaves, the maximum distance between any pair of leaves occurs twice (Semple and Steel 2003; Durbin et al. 1998). For example, in Figure 2.5, suppose $i = A, \ j = B, \ k = D$. Then we have $\max\{d_{AB}, d_{AD}, d_{BD}\} = \max\{0.5, 2, 2\} = 2$. We see the maximum distance occurs twice, d_{AD} and d_{BD}. Assuming the ultrametric condition holds, we successfully reconstructed phylogenetic trees from the DNA sequences. Once vectorized, the phylogenetic trees are in a format that supports further analysis.

2.5 Phylogenetic Dimensionality

Phylogenetic trees reside in a multidimensional tree space. The dimensions of the tree space are determined by the number of leaves on the trees. Visualizing a higher dimensional tree space is extremely difficult. The vector shown in Figure 2.5, is a six-dimensional vector. For context, a phylogenetic tree with only 10 leaves exists in a 45 dimensional tree space. Visualization of vectorized phylogenetic gene tree clusters residing in higher dimensions requires a dimensionality reduction method.

The most widely used dimension reduction technique is Principle Component Analysis (PCA). PCA, developed by Karl Pearson in 1901, finds the best-fitting lower dimensional plane to represent a higher dimensional space (Pearson 1901). Classical PCA utilizes multidimensional Euclidean space. In order to explore the research question, we need a method that reduces dimensionality in a tree space.

This thesis aims to explore the effects of multiple dimensional reduction techniques on gene trees in tree space. Due to the lack of analysis tools in the field of phylogenomics, visualizing the results plays a crucial role in the ability to determine the degree to which gene trees cluster. Two-dimensional visualization is relatively quick computationally and provides an easy method of assessing tree space PCA results.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Literature Review

Phylogenetics provides the basis for reconstructing as many gene trees as possible with the given data. However; a large number of reconstructed gene trees by themselves do not get us closer to inferring a species' tree. Attempting to piece together the species' tree from gene trees is only an inference because the true species tree will never be known. Phylogenomics, a newer field of study, attempts to use all shared genes among interrelated species to glean insight on the species' tree. An example of phylogenomics is the examination of shared genes among humans, gorillas and chimpanzees in order to infer their ancestral relationships. These relationships may help analysts better understand how these species evolved.

Phylogenomics utilizes various techniques which attempt to leverage reconstructed gene trees to infer species topologies. Several decades ago there were no proven techniques or tools to infer species topologies from gene trees. Many of these gaps and challenges were due to the complex structure of tree space, the high dimensionality of phylogenetic trees, few techniques for inference, and the lack of computational power. Recent research efforts have made significant progress in overcoming these gaps and challenges.

Our thesis aims to answer if gene trees cluster by distribution, with the intent of leveraging clusters to gain species topology insights. For example, given shared gene trees across multiple species, can we utilize proven techniques to analyze and visualize clustering? Through visualization, can we determine if gene trees cluster by species? Lastly, how effective are the dimension reduction techniques? This literature review highlights four significant breakthroughs in phylogenetic research that we utilize to answer our research questions.

3.1 Tree Space Geometry

The first hurdle overcome in phylogenomics was proving the existence of a tree space and a distance metric usable in that tree space. Prior to the work of authors Louis Billera, Susan Holmes, and Karen Vogtmann there were conflicting results and conclusions in the field of phylogenomcis (Billera et al. 2001). In 2001, Bilera, Holmes and Vogtmann (BHV) published *Geometry of the Space of Phylogenetic Trees* which featured a breakthrough that proved the existence of a defined tree space.

BHV proved the existence of a continuous space of trees, $\mathcal{T}_n$, containing all phylogenetic trees with the same number of leaves (n) (Billera et al. 2001). The space has unique geometric properties, most importantly, the existence of geodesic paths and centroids (Billera et al. 2001). The geodesic is a unique shortest path connecting any two trees in a tree space $\mathcal{T}_n$ (Billera et al. 2001). Centroids are the defined center for a set of trees in a tree space $\mathcal{T}_n$, and are helpful in conducting comparisons (Billera et al. 2001).

The BHV tree space $\mathcal{T}_n$ is critical to phylogenomics. At the time of publication, the research by (Billera et al. 2001) addressed the fundamental need for a geometric phylogenetic tree space and a method for measuring tree distance and centroids. This research is important because it proves that gene trees with the same number of leaves exist in the same dimension and tree space (Billera et al. 2001). These findings (Billera et al. 2001) provide an avenue to reconstruct gene trees for all shared genes across multiple species.

Additionally, the geodesic enables us to measure the distance between gene trees and locate the central location (centroid) for any cluster of gene trees. Unfortunately, numerous challenges still exist. Phylogenetic trees with numerous leaves remain in high dimensions. Also, the tree space $\mathcal{T}_n$ is not Euclidean, thus, we cannot use classical techniques to reduce its dimensionality (Billera et al. 2001). High dimensionality limits our ability to visualize, therefore a tree space dimensionality reduction technique is needed. Visualization is important because it helps us better understand the large number of phylogenetic trees in tree space.

3.2 Tree Space Principal Components Analysis

Tom Nye addresses the need to reduce dimensionality in the BHV tree space $\mathcal{T}_n$ (Nye 2011). He uses the geodesic metric, developed by (Billera et al. 2001), to represent phylogenetic trees in the tree space and reduces dimensionality through Principal Component Analysis in the tree space. Tom Nye's approach identifies the centroid of a set of phylogenetic trees, and projects the trees onto a line through the centroid. He iterates this process until he finds the line that minimizes the sum of squared distances between the trees and their projected points (Nye 2011). He proves that the successful completion of these steps reduces dimensionality of a set of trees and presents a method for identifying principal paths (Nye 2011).

Tom Nye's research proves that dimension reduction is feasible in the tree space which is critical to our ability to cluster and visualize gene trees in two dimensions. However; his approach is not without gaps and challenges. The lines through the centroid are of infinite length, requiring numerous assumptions to make the approach feasible (Nye 2011). Additionally, there is no algorithm to efficiently iterate between steps or project onto the principal path (Nye 2011). As a result, conducting PCA in the space of phylogenetic trees is computationally strenuous with no guarantee of success (Nye 2011). The greatest challenge Nye faces is that his technique cannot be generalized to a higher order PCA. His technique can only be applied to reduce the dimensionality of the space into the two dimensional line (Nye 2011). Therefore the need remains to develop an efficient method for successful dimension reduction in the tree space.

3.3 Tree Space Principal Component Analysis Utilizing the Fréchet Mean

Due to the challenges of his 2011 PCA approach (Nye 2011), Tom Nye, with Xiaoxian Tang, Grady Weyenberg, and Ruriko Yoshida develop a more effective PCA method to efficiently reduce dimensionality in the BHV tree space $\mathcal{T}_n$ (Nye et al. 2017). Their approach projects a set of trees onto the locus of the Fréchet mean. The locus of the Fréchet mean represents the position of the centroid for a set of trees (Nye et al. 2017). Projecting onto a known geometric object, the locus of the Fréchet mean resolves the challenges associated with infinite length lines and locating the centroid for a set of trees (Nye et al. 2017). They proved this approach successfully reduces dimensionality in the tree space and did so more efficiently (Nye et al. 2017).

Nye, Tang, Weyenberg and Yoshida resolve some consistency issues; however, projecting onto the locus of the Fréchet mean creates new challenges. Specifically, there is no algorithm to efficiently determine the optimal geodesic paths for projection, resulting in time intensive computations (Nye et al. 2017). Additionally, the tree projections in lower dimensions are frequently skewed and clustered towards the boundaries of the Fréchet mean, creating unexplained variance and making visualization ineffective. While the locus of the Fréchet mean approach reduces dimensionality (Nye et al. 2017), gaps remain. A need remains for a method to efficiently reduce tree space dimensionality in a manner that allows us to analyze gene trees in two dimensions.

3.4 Tropical Geometry

The research of Ruriko Yoshida, Leon Zhang, and Xu Zhang successfully reduces tree space dimensionality, facilitating species trees inference (Yoshida et al. 2019). Yoshida, Zhang and Zhang developed tropical PCA, a novel PCA approach, which uses a tropical metric in lieu of the BHV geodesic metric (Yoshida et al. 2019). The tropical metric uses tropical geometry and the structure of tropical algebra to represent a tree space. A brief review of tropical basics is necessary before discussing tropical PCA (Yoshida et al. 2019).

The term *tropical*, with regards to mathematics, originates from colleagues of Brazilian mathematician Imre Simon, who respected Imre's work and named it after Brazil's tropical climate (Pin 1998). Tropical algebraic structure is desirable for use in phylogenetics because optimization on graphs can be easier to formulate (Cuninghame-Green 1979). Figure 3.1 is a fictitious example species tree (S_1) containing three species: Humans, Chimpanzees and Gorillas. Contained within Figure 3.1 are four gene trees (T_1,T_2,T_3,T_4), shown in Figure 3.2. We use Figures 3.1 and 3.2 to illustrate tropical basics. Chapter 4.1.1 contains a deeper discussion regarding how this thesis simulates species and gene trees.

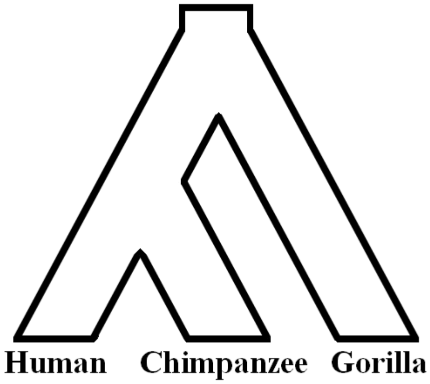
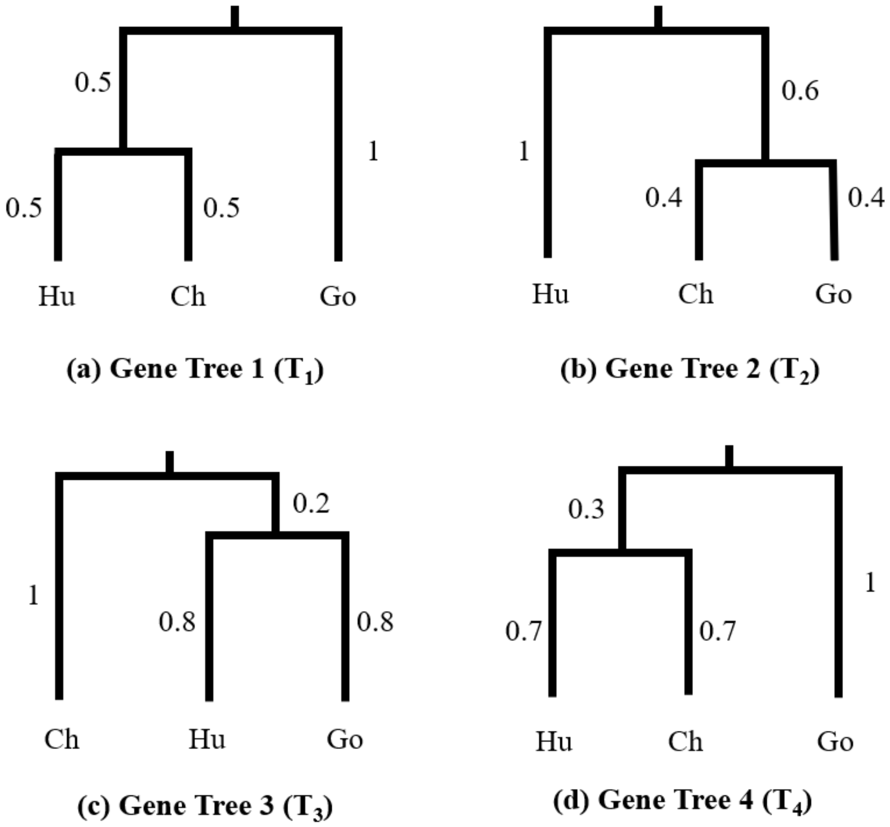


Figure 3.1. Species Tree (S_1)



(a) Gene Tree 1 (T_1)

(b) Gene Tree 2 (T_2)

(c) Gene Tree 3 (T_3)

(d) Gene Tree 4 (T_4)

Figure 3.2. Gene Trees Contained Within The Species Tree (S_1)

Tropical Semiring

The tropical geometry used in this thesis is based on a max-plus tropical semiring. The semiring is the set of all real numbers $\mathbb{R}$ and negative infinity ($-\infty$), on which tropical addition and tropical multiplication are defined (Maclagan and Sturmfels 2015). The max-plus tropical semiring replaces the algebraic addition (+) and multiplication (×) arithmetic operators with tropical addition ($\oplus$) and tropical multiplication ($\odot$) operators (Maclagan and Sturmfels 2015).

$$a \oplus b := \max\{a, b\}, \quad a \odot b := a + b \quad \text{where } a, b \in \mathbb{R}.$$

For example, the tropical addition, or sum, of 3 and 6 is 6. The tropical product of 3 and 6 is 9. More formally, $3 \oplus 6 := \max\{3, 6\} = 6$ *and* $3 \odot 6 := 3 + 6 = 9$.

Tropical Metric

Tropical PCA uses the tropical metric to define the distance between two trees (Yoshida et al. 2019). To better illustrate the tropical metric, we vectorize gene trees (T_1,T_2,T_3,T_4) in Figure 3.2. Figures 3.3 and 3.4 show the vectorization process for gene trees (T_1,T_2,T_3,T_4).

	Hu	Ch	Go
Hu	0	1	2
Ch	-	0	2
Go	-	-	0

(a) Gene Tree 1 (T_1)

	Hu	Ch	Go
Hu	0	2	2
Ch	-	0	0.8
Go	-	-	0

(b) Gene Tree 2 (T_2)

	Hu	Ch	Go
Hu	0	2	1.6
Ch	-	0	2
Go	-	-	0

(c) Gene Tree 3 (T_3)

	Hu	Ch	Go
Hu	0	1.4	2
Ch	-	0	2
Go	-	-	0

(d) Gene Tree 4 (T_4)

Figure 3.3. Gene Tree Dissimilarity Maps (d_{ij})

$$v_1 = (\underset{d_{Hu,Ch}}{1,} \quad \underset{d_{Hu,Go}}{2,} \quad \underset{d_{Ch,Go}}{2})$$

(a) Gene Tree 1 (T_1)

$$v_2 = (\underset{d_{Hu,Ch}}{2,} \quad \underset{d_{Hu,Go}}{2,} \quad \underset{d_{Ch,Go}}{0.8})$$

(b) Gene Tree 2 (T_2)

$$v_3 = (\underset{d_{Hu,Ch}}{2,} \quad \underset{d_{Hu,Go}}{1.6,} \quad \underset{d_{Ch,Go}}{2})$$

(c) Gene Tree 3 (T_3)

$$v_4 = (\underset{d_{Hu,Ch}}{1.4,} \quad \underset{d_{Hu,Go}}{2,} \quad \underset{d_{Ch,Go}}{2})$$

(d) Gene Tree 4 (T_4)

Figure 3.4. Gene Trees Vectorized Distances (d_{ij})

Figure 3.5 depicts the gene trees (T_1,T_2,T_3,T_4) in vector form. We use the tropical metric to measure the distance between gene trees. Suppose we have two gene trees $w, v \in \mathbb{R}^e$

with $w = (w_1, \dots, w_e)$ and $v = (v_1, \dots, v_e)$. Then (Yoshida et al. 2019) defines the tropical metric between two points in $\mathbb{R}^e$ as:

$$d_{\text{tr}}(v, w) \;=\; \max\big\{ |v_i - w_i - v_j + w_j| \;:\; 1 \le i < j \le e \big\}. \tag{3.1}$$

This metric is also known as the *generalized Hilbert projective metric* (Maclagan and Sturmfels 2015; Liverani and Wojtkowski 1994).

$$\begin{array}{rrrr} T_1 = & (1, & 2, & 2) \\ T_2 = & (2, & 2, & 0.8) \\ T_3 = & (2, & 1.6, & 2) \\ T_4 = & (1.4, & 2, & 2) \end{array}$$

Figure 3.5. Gene Tree (T_1,T_2,T_3,T_4) Vectors

Example 1 *For gene trees* T_1 *and* T_2*,* $e = 3$*,* $T_1 = (1, 2, 2)$*,* $T_2 = (2, 2, 0.8)$*. Then* $d_{\text{tr}}(T_1, T_2) = \max\big\{ |2 - 1 - 2 + 2|,\ |2 - 2 - 0.8 + 2|,\ |2 - 1 - 0.8 + 2| \big\} = \max\big\{ |1|,\ |1.2|,\ |2.2| \big\} = 2.2.$

Note that in order for the tropical metric to be a metric, it must satisfy the condition

$$d_{\text{tr}}(v, w) = 0 \text{ if and only if } v = w.$$

However, if $v = (0, 0, 0)$, $w = (c, c, c)$ where $c \neq 0$ and $c \in \mathbb{R}$, then $d_{\text{tr}}(v, w) = 0$. Therefore in order to force the tropical metric to be a metric, we work on the algebra over the *projective space* $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ with $\mathbf{1} = (1, \dots, 1)$, i.e., $\mathbb{R}^e$ with an additional condition where we define $(c, c, \dots, c) \in \mathbb{R}^e = (0, 0, \dots, 0)$ for any $c \in \mathbb{R}$. Also note that $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ is isometric to $\mathbb{R}^{e-1}$.

Tropical Convex Hull

Tropical convex hulls are the smallest convex combination set of points. Tropical convex hulls are required for the projection of gene trees. A subset $S \subset \mathbb{R}^e$ is called *tropically convex* if it contains the point $a \odot x \oplus b \odot y$ for any $x, y \in S$ and any $a, b \in \mathbb{R}$ (Yoshida et al. 2019). The *tropical convex hull* or *tropical polytope* $\text{tconv}(V)$ generated by a finite subset $V \subset \mathbb{R}^e$ is the smallest tropically convex subset containing $V \subset \mathbb{R}^e$.

The tropical convex hull of $V = \{v_1, \ldots, v_s\}$ can also be written as the set of all possible tropical linear combinations of vertices $v_1, \ldots, v_s$ such that:

$$\text{tconv}(V) = \{a_1 \odot v_1 \oplus a_2 \odot v_2 \oplus \cdots \oplus a_s \odot v_s : v_1, \ldots, v_s \in V \text{ and } a_1, \ldots, a_s \in \mathbb{R}\}.$$

Example 2 *Gene trees T_1, T_2 and T_3 are in $e = 3$ with $T_1 = (1, 2, 2)$, $T_2 = (2, 2, 0.8)$, $T_3 = (2, 1.6, 2) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$.* The tropical polytope of these three trees (i.e., tropical triangle) can be found in Figure 3.6.

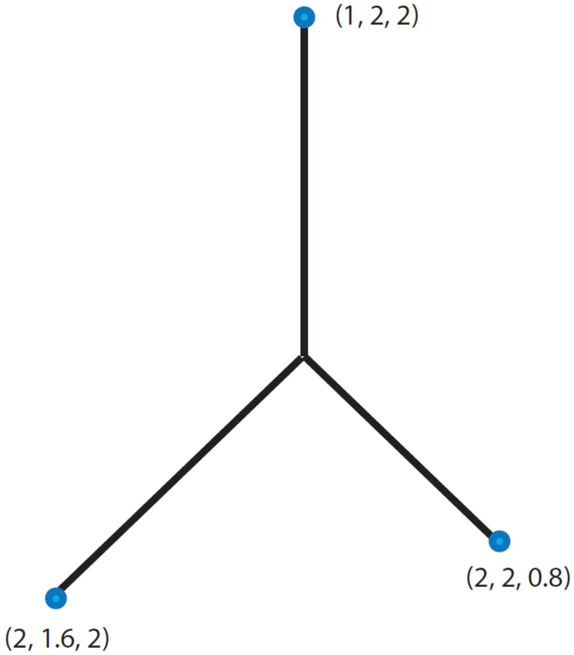


Figure 3.6. Tropical Polytope

Note that if all vertices $V = \{v_1, \ldots, v_s\}$ are ultrametrics, then the tropical convex hull tconv(V) is a subset in the space of ultrametrics (Yoshida et al. 2019). This means that all vectors $v_1, \ldots, v_s$ are ultrametrics computed from equidistant trees and the tropical polytope tconv(V) is in the tree space.

3.5 Tropical Principal Component Analysis

Yoshida, Zhang and Zhang prove the dimensionality of a set of trees in a tree space can easily be reduced to any lower dimension utilizing Tropical PCA (Yoshida et al. 2019).

Implementing tropical PCA to reduce dimensionality in a tree space is a multiple step process that leverages tropical geometry (Yoshida et al. 2019). We are interested in using Tropical PCA for this thesis because previous research shows that the tropical metric works best for the space of ultrametrics as a tree space (Yoshida et al. 2019).

Tropical Line Segment

(Yoshida et al. 2019) define a tropical geodesic, the shortest path, between two points $v_1,\ v_2 \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ as a tropical line segment. A tropical line segment is a tropical polytope generated by $v_1,\ v_2 \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. A tropical line segment between two points can be computed in linear time with known dimensionality (Maclagan and Sturmfels 2015).

Algorithm 3 (Tropical line segment in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$)

```
• Input: A vector $u = (u_1, \ldots, u_e)$ and a vector $w = (w_1, \ldots, w_e)$.
• Output: A tropical line segment between u and w in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$.
• Algorithm:
    1. Compute $\lambda = w - u = ((w_1 - u_1), \ldots, (w_e - u_e))$.
    2. Set $L = \emptyset$.
    3. For $i = 1, \ldots, e$, do:
        (a) Find the ith smallest coordinate $(i_1, \ldots, i_e)$ in $\lambda$.
        (b) Set $y = \big(\max\{\lambda_{i_1 \ldots\ i_e} + u_1, w_1\}, \ldots, \max\{\lambda_{i_1 \ldots\ i_e} + u_e, w_e\}\big)$.
        (c) Set $L = L \cup \{y\}$.
    4. Return the line segments connecting the ultrametrics in L.
```

Note: λ will be a vector of length = e.

Example 4 *The gene trees in Figure 3.2 are in $e = 3$ space with $T_1 = (1, 2, 2)$, $T_2 = (2, 2, 0.8) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. Following Algorithm 3, we have:*

$$\begin{aligned} T_1 &= (1, 2, 2) \\ T_2 &= (2, 2, 0.8) \\ T_2 - T_1 &= (1, 0, -1.2) \end{aligned}$$

Set

$$\lambda = -1.2,\ 0,\ 1$$

$$\begin{aligned}
T_1 + (-1.2, -1.2, -1.2) &= (-0.2, 0.8, 0.8) \\
\max\{(-0.2, 0.8, 0.8), T_2\} &= \max\{(-0.2, 0.8, 0.8), (2, 2, 0.8)\} \\
\max\{(-0.2, 0.8, 0.8), (2, 2, 0.8)\} &= (2, 2, 0.8) \\
&= T_2
\end{aligned}$$

$$\begin{aligned}
T_1 + (0, 0, 0) &= T_1 \\
\max\{T_1, T_2\} &= (2, 2, 2) \\
(2, 2, 2) - (c, c, c) &= (0, 0, 0), \;\; for \;\; c = 2 \\
&= (0, 0, 0)
\end{aligned}$$

$$\begin{aligned}
T_1 + (1, 1, 1) &= (2, 3, 3) \\
\max\{(2, 3, 3), T_2\} &= (2, 3, 3) \\
(2, 3, 3) - (c, c, c) &= (1, 2, 2), \;\; for \;\; c = 1 \\
&= (1, 2, 2) \\
&= T_1
\end{aligned}$$

Note: Line segments can be adjusted by a different c because we are using the tropical projection D onto the tropical polytope $\mathcal{P}$.

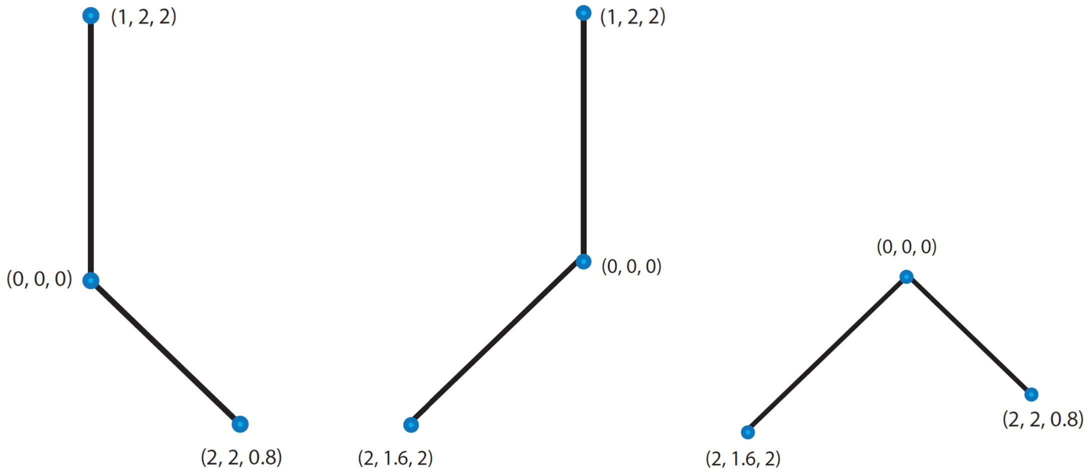
Therefore, the tropical line segment between gene trees T_1 and T_2 is a line segment from $(1, 2, 2)$ to $(0, 0, 0) = (2, 2, 2)$, and then going to $(2, 2, 0.8)$ shown in Figure 3.7a. Similarly, we can show that the tropical line segment between T_2 and T_3 is a line segment from $(2, 2, 0.8)$ to $(0, 0, 0) = (2, 2, 2)$, and then going to $(2, 1.6, 2)$ shown in Figure 3.7b. Also we can show that the tropical line segment between T_1 and T_3 is a line segment from $(1, 2, 2)$ to $(0, 0, 0) = (2, 2, 2)$, and then going to $(2, 1.6, 2)$ shown in Figure 3.7c.

Tropical Projection onto a Tropical Convex Hull

Consider a tropical polytope $\mathcal{P} = \text{tconv}(D^{(1)}, D^{(2)}, \ldots, D^{(s)})$ where $D^{(i)}$ are points in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$. Suppose $D \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. (Maclagan and Sturmfels 2015) show that the tropical projection of D onto the tropical polytope $\mathcal{P}$ is defined as:

$$\pi_{\mathcal{P}}(D) = \lambda_1 \odot D^{(1)} \oplus \lambda_2 \odot D^{(2)} \oplus \cdots \oplus \lambda_s \odot D^{(s)}, \quad \text{where } \lambda_k = \min\{D - D^{(k)}\} \quad (3.2)$$

Example 5 *The gene trees in Figure 3.2 are in $e = 3$ with $D^{(1)} = T_1 = (1, 2, 2)$, $D^{(2)} = T_2 =$*



(a) $\mathbf{T_1}$ to $\mathbf{T_2}$

(b) $\mathbf{T_1}$ to $\mathbf{T_3}$

(c) $\mathbf{T_2}$ to $\mathbf{T_3}$

Figure 3.7. Tropical Line Segments

$(2, 2, 0.8)$, $D^{(3)} = T_3 = (2, 1.6, 2) \in \mathbb{R}^3/\mathbb{R}\mathbf{1}$. *Also let* $D = T_4 = (1.4, 2, 2)$.

$$\begin{aligned}\lambda_1 &= \min\{T_4 - T_1\}\\ &= \min\{(1.4, 2, 2) - (1, 2, 2)\}\\ &= \min\{0.4, 0, 0\}\\ &= 0\end{aligned}$$

$$\begin{aligned}\lambda_2 &= \min(T_4 - T_2)\\ &= \min\{(1.4, 2, 2) - (2, 2, 0.8)\}\\ &= \min\{-0.6, 0, 1.2\}\\ &= -0.6\end{aligned}$$

$$\begin{aligned}\lambda_3 &= \min\{T_4 - T_3\}\\ &= \min\{(1.4, 2, 2) - (2, 1.6, 2)\}\\ &= \min\{-0.6, 0.4, 0\}\\ &= -0.6\end{aligned}$$

$$\begin{aligned}\pi_{\mathcal{P}}(D) &= \lambda_1 \odot T_1 \oplus \lambda_2 \odot T_2 \oplus \lambda_3 \odot T_3 \\ &= \max\{(1, 2, 2), (1.4, 1, 1.4), (1.4, 1, 1.4)\} \\ &= (1.4, 2, 2) \\ &= T_4\end{aligned}$$

Then the projection of T_4 onto the tropical polytope is $(1.4, 2, 2)$ illustrated in Figure 3.8. Therefore, T_4 is on the tropical line segments between T_1 and T_2.

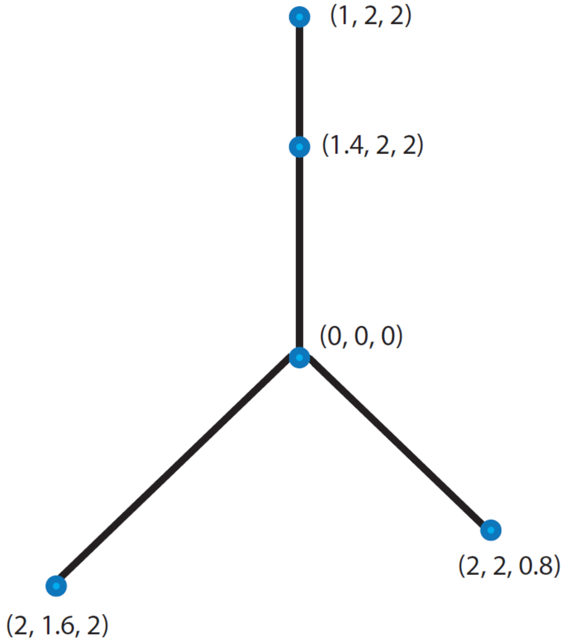


Figure 3.8. Tropical Projection onto a Tropical Polytope

In this thesis all vertices of the tropical polytopes $D^{(1)}, \ldots, D^{(s)}$ and D are ultrametrics representing equidistant trees.

Second Order Tropical PCA

(Yoshida et al. 2019) define a $(s-1)$th order tropical PCA as a tropical convex hull generated by s many vectors in $\in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ such that it minimizes the sum of tropical distances defined by the tropical metric between each data point and its tropical projection onto the tropical polytope. If all vertices of the polytope are ultrametrics then all points in the polytope are ultrametrics. Thus, a tropical PCA defined by (Yoshida et al. 2019) can be applied to a space of equidistant trees with a fixed number of leaves.

In order to describe the tropical PCA over $\in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, we will describe a classical PCA over an Euclidean space such as $\mathbb{R}^e$. A classical PCA can be described as a linear plane in $\mathbb{R}^e$ which minimizes the sum of distances between each data point in the sample and its orthogonal projection onto the linear plane. In Figure 3.9, $\{X_1, \ldots, X_7\}$ are the data points (blue circles) and their projections (red circles) onto the linear plane L. Then the classical PCA is the linear plane L such that it minimizes the sum of the squared distances between each blue circle and its red circle (the distance of each green line).

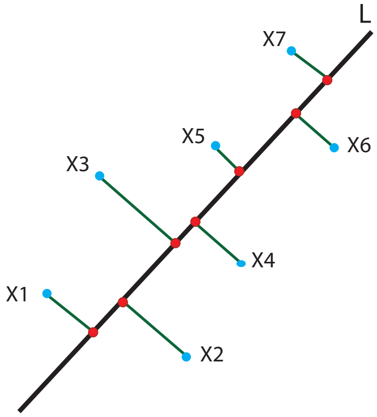


Figure 3.9. Classical PCA Example

A second order Tropical PCA of the gene trees in Figure 3.2 first requires selection of the tropical PCA base. We desire a second order tropical PCA ($s - 1 = 2$), therefore we set $s = 3$ and select three gene trees to act as our base. In this example, we select gene trees T_1, T_2, and T_3 as our tropical PCA base. Since T_4 is on the tropical line segments between T_1 and T_2 and between T_1 and T_3, as seen in Figure 3.8, T_4 is in the tropical triangle between T_1, T_2, T_3. Thus, the second order tropical PCA is the tropical triangle generated by T_1, T_2, T_3 shown in Figure 3.10. T_4 is in the tropical triangle, thus the sum of tropical distances between T_i and its projection onto the tropical triangle is 0.

Utilizing tropical PCA to reduce the dimensionality of a set of trees in a tree space is

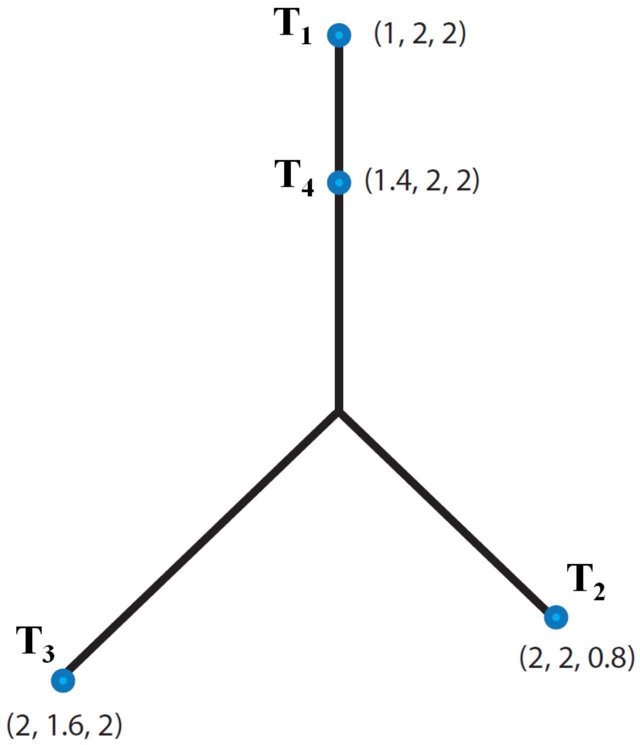


Figure 3.10. Tropical PCA Of T_1, T_2, T_3, T_4

incredibly reliable and easy to apply. Additionally, second order tropical PCA produces tree projections that are successfully visualized in two dimensions. Despite its advantages, two primary gaps remain in Tropical PCA application. The research lacked analysis leveraging simulated phylogenetic trees and tropical PCA application. Such analysis could illuminate how effectively Tropical PCA captures changes in phylogenetic data. Additionally, it is not clear the computational time complexity of a tropical PCA over a tree space. They have shown that a computing a tropical PCA over a tropical projective space can be formulated as a mixed integer programming (Yoshida et al. 2019). A more efficient computational technique to approximate the optimal solution is necessary to determine the optimal PCA base in an effort to improve computation time and accuracy.

3.6 Thesis Application

This thesis explores the relationships between gene trees and their projections in a lower dimensional space. Tropical PCA (Yoshida et al. 2019) provides a proven method to reduce the dimensionality of phylogenetic trees. Utilizing gene trees, the goal is to determine if projected gene trees cluster or group together. The intuition being that current species that share genes are closely related. Closely related species come from the same ancestor. By visualizing all projected gene trees, we verify that our tropical PCA implementation is effectively reducing gene tree dimensionality. Then we use the clustering location of gene trees to infer the ancestral (species) tree.

This thesis leverages the findings in this literature review to perform analysis on phylogenetic trees in a tropical tree space. The objective of this analysis is to answer our research question, to determine if gene trees cluster by distribution to infer the species tree. Our **hypothesis** is through the development of our *tropical PCA* (Yoshida et al. 2019) algorithm, we more efficiently and effectively visualize gene tree clusters by distribution (species), facilitating species' topology inference.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4: Methodology

Leveraging Tropical PCA (Yoshida et al. 2019), we explore the effect of varying gene tree properties on clustering and species' trees inference. Species tree inference is predicated on gene trees clustering, therefore we need to examine the conditions in which genes do and do not cluster. Additionally, our approach aims to analyze the effectiveness of second order tropical PCA in visualizing groups of projected gene trees. Chapter 4 discusses this thesis' methodology purely in terms of the approach. Chapter 5 presents the numeric results from our approach. Figure 4.1 illustrates our approach outlined in this chapter.

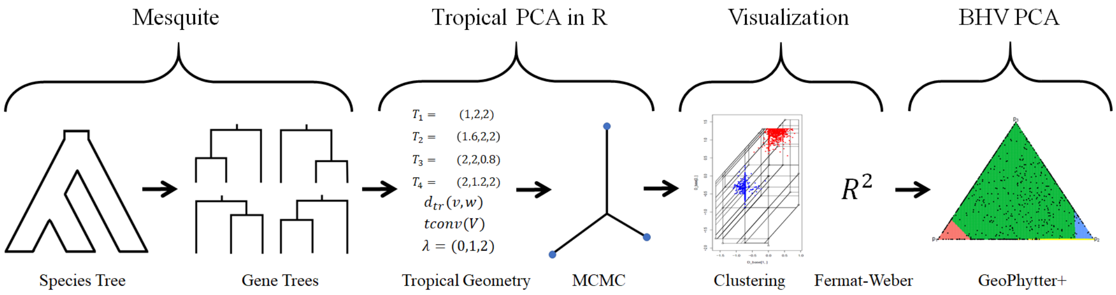


Figure 4.1. Thesis Approach Diagram

This thesis' approach is broken down into four primary steps shown in Figure 4.1: data simulation in Mequite, computing a second order tropical PCA in R, visualization, and comparison to BHV PCA.

4.1 Phylogenetic Data

Determining the effectiveness of tropical PCA (Yoshida et al. 2019) on gene tree clusters requires lots of gene tree data. While there is an abundance of phylogenetic data, such data is in the form of DNA sequences. Due to the multiple assumptions and various techniques necessary to reconstruct DNA sequences, we prefer data in the form of species and gene trees. In this thesis, we use software to create species and gene trees to allow true tree topology to be known.

4.1.1 Mesquite

Mesquite, developed by brothers Wayne P. Maddison and David R. Maddison (Maddison and Maddison 1997), is a powerful phylogenetic software program. Mesquite allows us to generate phylogenetic trees to almost any specification imaginable. In a perfect world we want to know (not infer) the true species' topology, and *Mesquite* provides us that opportunity. We use Mesquite to generate gene trees using a simple coalescent model. Coalescent models assume that genes are equally likely to be passed from one generation to the next. Coalescent models are an accepted technique and described in detail in this article (Maddison and Knowles 2006).

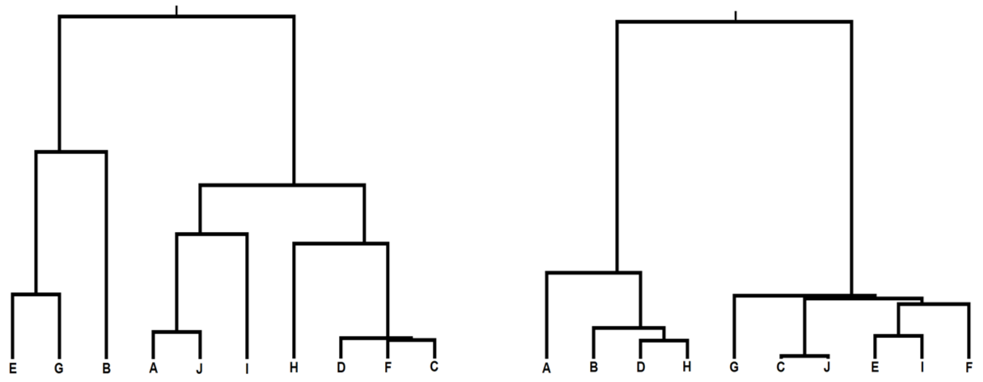
Before generating any trees, we set the conditions. The first step is to determine the number and names of the desired species. For complexity, we choose 10 species, making our trees 45 dimensional. For simplicity, we name the species: A, B, C, D, E, F, G, H, I and J. Next, we create 10 genes: a, b, c, d, e, f, g, h, i and j.

Species Trees

Species trees are simulated in Mesquite using a Yule uniform speciation process. The Yule uniform speciation process is a pure birth evolutionary process that favors rooted binary phylogenetic trees (Semple and Steel 2003). A pure birth process prevents the early termination, or death, of any species. This is important because we want the species trees to only contain current species and be equidistant.

There are two input parameters necessary to simulate species trees, the number of generated species trees and tree depth. Tree depth or species depth represents the total number of generations of the species. In order to analyze the impact of gene tree clustering we vary the species depth (Maddison and Knowles 2006). This thesis conducts two sets of analyses on generated species trees. The first analysis examines gene trees generated from two different species trees across six different species depths. The second analysis examines gene trees generated from eight different species trees across the same six species depths.

We vary the total tree depth as follows: 25,000, 50,000, 100,000, 250,000, 500,000, and 1,000,000. The unit for species (tree) depth is time in number of generations (Maddison and Maddison 1997). Figure 4.2 shows the two species trees, from our first analysis, at a species depth of 25,000.



(a) Species Tree 1

(b) Species Tree 2

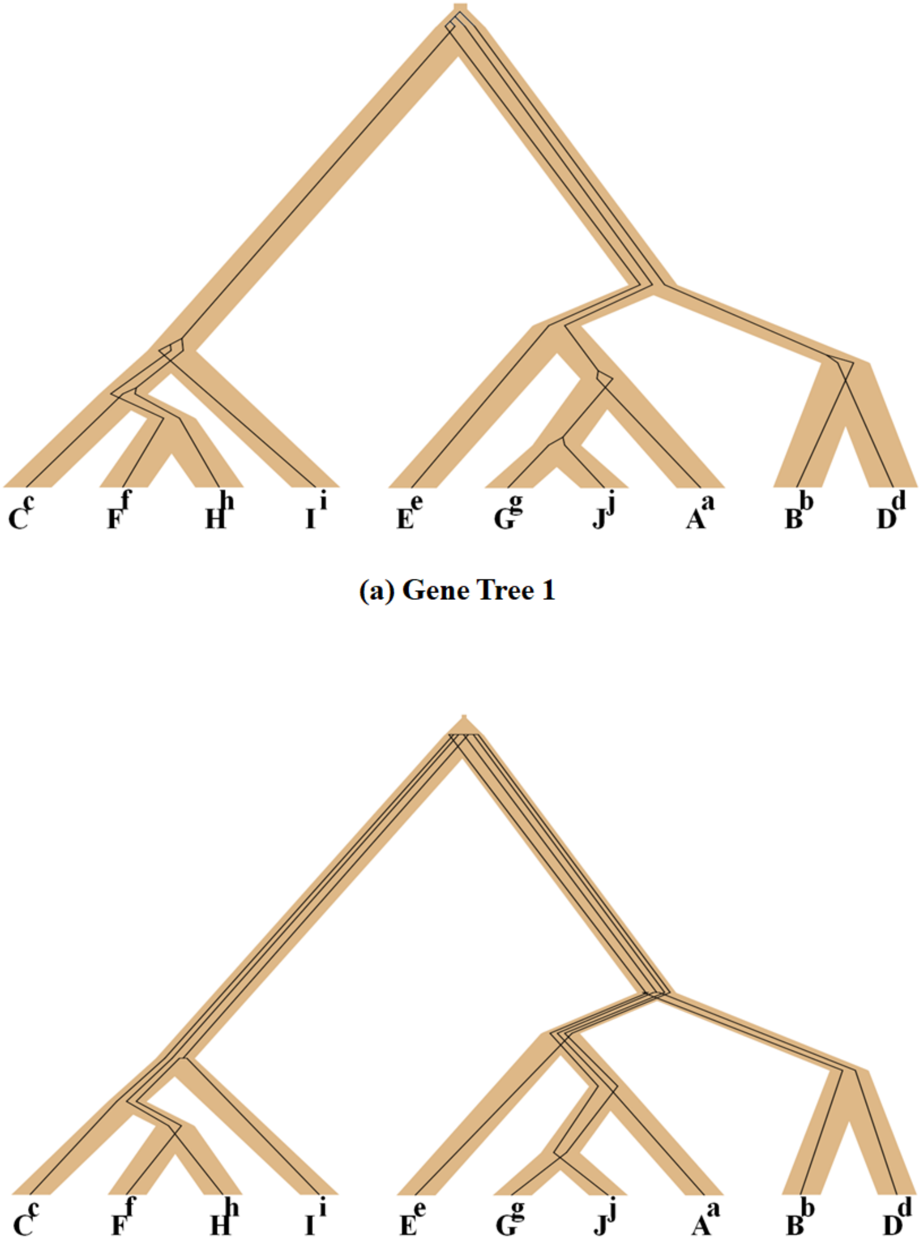
Figure 4.2. Two Simulated Species Trees (Species Depth = 25,000 Generations)

Gene Trees

We simulate 1000 gene trees from each species tree. Therefore, we simulate 2000 gene trees at each species depth in our first analysis and 8000 gene trees in our second analysis. We simulate gene trees in Mesquite using a coalescent contained within the current tree model. A coalescent model constructs gene trees by coalescing genes together at common ancestors (Aldous 2001). Gene trees are contained within our simulated species trees. Figure 4.3 shows two gene trees contained within the same species tree from Mesquite.

Creation of an association between each species and its corresponding gene enables coalescent models to simulate multiple different gene trees contained within the same species tree. The association represents the idea of containment, where each gene is contained in a species. The same species tree contains all 1000 simulated gene trees; however, each gene tree coalesces differently within the species tree. In Figure 4.3, the gene trees, represented by the black lines, are contained within the same species tree, represented by the wider tan lines.

In each gene tree (black lines), the genes (lower case letters) start at current time (bottom of the tree) with an association to a specific species (upper case letters). The coalescent model we use, merges or joins the genes (black lines) with the same gene (other black lines) found in other species within the species tree (tan color). Where the genes (black lines) merge and who (other black lines associated with a different species) the genes merges with is different for each gene tree. Note the genes for both gene trees in Figure 4.3 have identical starting locations, the bottom of the tree, yet Gene Tree 1 and Gene Tree 2 have completely different gene pathing.

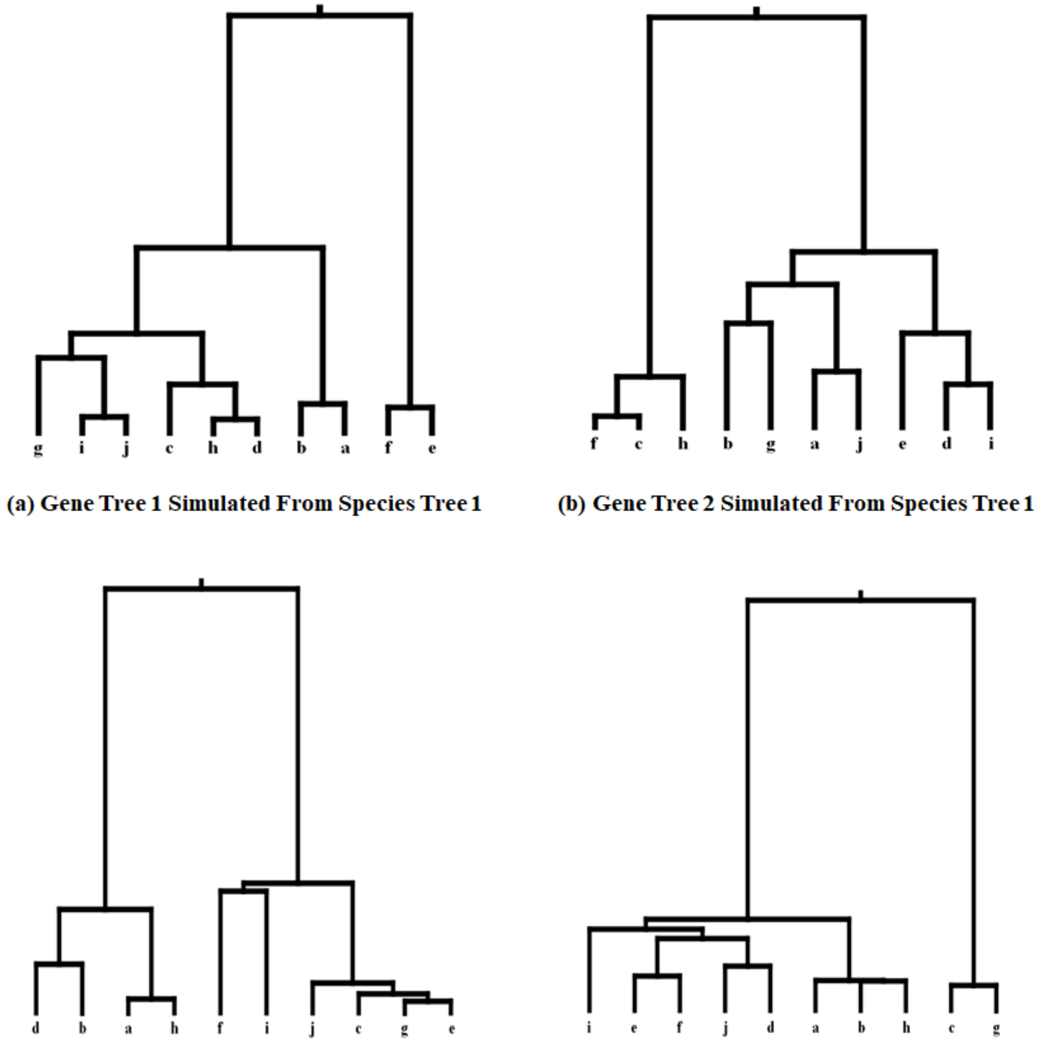


(a) Gene Tree 1

(b) Gene Tree 2

Figure 4.3. Simulated Gene Trees Within The Same Species Tree

We simulate 1000 gene trees using one parameter, the effective population size. We hold constant the effective population size of 100,000 throughout our analysis. Research by Maddison and Knowles demonstrates that an effective population size of 100,000 is reasonable and robust (Maddison and Knowles 2006). We hold the effective population size constant to allow us to explore the relationship between species depth and effective population size. Figure 4.4 illustrates gene trees from two different species trees at a species depth of 25,000.



(a) Gene Tree 1 Simulated From Species Tree 1

(b) Gene Tree 2 Simulated From Species Tree 1

(c) Gene Tree 1 Simulated From Species Tree 2

(d) Gene Tree 2 Simulated From Species Tree 2

Figure 4.4. Gene Trees Contained Within Different Species Trees

This thesis analyzes the ratio (r) between species depth and effective population size.

$$\text{ratio (r)} = \frac{\text{Species Depth}}{\text{Eff Pop Size}}$$

This ratio is vital in measuring the effectiveness of tropical PCA with regards to clustering. The intuition is that as the ratio (r) becomes larger, the clusters become more pronounced. If our tropical PCA approach is functioning properly, the gene trees will noticeably cluster. In this thesis, only species (tree) depth affects the ratio since we hold effective population size constant. As a result, the greater the species depth, the more generations of species in the tree.

4.2 Tropical Principle Component Analysis in R

This thesis leverages the powerful statistical programming language R to conduct tropical PCA applications (Ihaka and Gentleman 1993). With thousands of phylogenetic gene trees at our disposal, the goal of our approach is to leverage Tropical PCA (Yoshida et al. 2019) to visualize the gene trees by species. R allows us to explore high dimensional tree space in an effort to locate the optimal tropical PCA base. Locating the optimal tropical PCA base is a significant challenge (Yoshida et al. 2019).

4.2.1 Stochastic Optimization

As trees become more complex, locating the optimal PCA base becomes more difficult. Increased difficulty requires significantly more time. The exact relationship between tree complexity and the time required to conduct tropical PCA over a tree space is not known. Yoshida, Zhang and Zhang explore all possible combinations of gene trees in order to approximate the tropical PCA base (Yoshida et al. 2019). However; exploring all possible combinations is not feasible when using large quantities of gene trees. For context, the total number of combinations of 3 gene trees selected from 2000 gene trees, is equal to 1,331,334,000 possible PCA base combinations.

Using enumeration to search in high dimensional tree space for the best fitting plane takes a lifetime (Hastings 1970). Additionally, this enumeration approach assumes the optimal PCA base consists of the available gene trees. However; the optimal PCA base does not necessarily include any of the gene trees we create. The effective implementation of tropical PCA requires a more efficient method to find the optimal base.

Markov Chain Monte Carlo

We implement a Markov chain Monte Carlo (MCMC) technique to reduce computational time and more efficiently explore planes in high dimensional tree space. MCMC consists of two basic principles, Monte Carlo methods and Markov Chains. Monte Carlo methods leverage repeated random sampling from a distribution to determine the probability of occurrences. Using Monte Carlo methods to find the best fitting tropical PCA base is not feasible because the probability distribution of the gene trees is unknown.

However; the addition of Markov chains to a Monte Carlo method allows for probabilistic repeated random sampling despite an unknown multi-dimensional probability distribution. MCMC methods create states (Markov chains) where the corresponding probability of moving from one state to another acts as the unknown probability distribution we desire (Felsenstein 2004). MCMC works by implementing a random walk in the state space, essentially mimicking a large amount of time or number of state transitions (Felsenstein 2004). After a long enough walk, the Markov chain is considered in steady-state, therefore, the equilibrium distribution π_i is equal to the unknown probability distribution (Felsenstein 2004).

This thesis uses the MCMC Metropolis-Hastings algorithm to locate the best fitting tropical PCA plane in an efficient amount of time. The Metropolis-Hastings algorithm iteratively selects a potential new state, determines the probability of acceptance, and accepts or rejects the proposed new state (Hastings 1970). We implement our MCMC Metropolis-Hastings algorithm in R using two primary functions. The first function generates proposals and the second function accepts or rejects the proposed new state.

We leverage MCMC to explore tree space for us. First, we select a random set of trees to serve as our tropical base. Then, we randomly walk the tree space distribution by modifying the leaves and branches of the current state. This modification consists of randomly rearranging leaves on the base tree and altering internal edge lengths. This technique, similar to (Mau and Newton 1997), leverages the posterior distribution to assist in smarter searching patterns. We stop to measure the sum of all tropical distances of the gene trees to their projections. We use the current state as the set of trees for the tropical base. Next, we accept or reject the proposed set of trees for our tropical base. That acceptance is determined by the comparison of the sum of all tropical distances at the current state to the best tropical distance measurement to date.

This proposal process is more formally written as: Let $\Pi_\Phi(S) := \sum_{i=1}^{|S|} d_{tr}(d_i, d_i')$, where $S = \{d_1, \ldots d_n\}$ with $d_i \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ are ultrametrics and d_i' is the tropical projection onto a tropical triangle Φ. Let $[m] := \{1, \ldots, m\}$.

We want to accept proposals, new sets of trees, because they could potentially lead us to the best-fitting tropical base. Yoshida, Zhang and Zhang prove the optimal tropical triangle (base) minimizes the sum of tropical distances between gene trees and their tropical projections onto that tropical triangle (Yoshida et al. 2019). $\Pi_\Phi(S)$ allows us to smartly explore tropical tree space, accepting and rejecting proposals of new sets of trees on the basis of how they compare to the current best tropical triangle (base). The following algorithm computes a proposal state, i.e., a set of proposed trees.

Algorithm 6 (Finding the proposed set of trees)

- *Input: Set of equidistant trees (current PCA base)* $\{T_1, T_2, T_3\}$, $k \in [m] := \{1, \ldots, m\}$.
- *Output: Next set of equidistant trees (proposed PCA base)* $\{T_1', T_2', T_3'\}$.

```
for <$i = 1, \ldots, 3$> do
    Set $T_i' = T_i$.
    Pick random numbers $(i_1, \ldots, i_k) \subset [m]$ without replacement.
    Permute the tree leaf labels $(i_1, \ldots, i_k) \subset [m]$ of $T_i'$ with a random permutation $\sigma$ in the symmetric group on $\{i_1, \ldots, i_k\}$.
    Pick a random internal branch $b_1$ in $T_i'$ with branch length $l_i$.
    Update $l_i := l_i + \epsilon \cdot c$ where $\epsilon \sim Unif\{\pm 1\}$, and $c \sim Unif[0, l_i/m]$.
    Pick another branch $b_2$ with branch length $l$ on the path from the root to a leaf where the branch $b_1$ is also on the path.
    If $l - \epsilon \cdot c < 0$ then set $l := 0$ and $l_i := l_i + l - \epsilon \cdot c$. If not then set $l := l - \epsilon \cdot c$.
end for
Return $\{T_1', T_2', T_3'\}$.
```

We use the Metropolis algorithm to decide whether the proposal state is accepted or rejected. Let $\Phi_{(w_1,w_2,w_3)}$ be the tropical triangle defined by ultrametrics w_1, w_2, w_3.

Algorithm 7 (Metropolis algorithm)

- *Input: Current set of equidistant trees $\{T_1, T_2, T_3\}$ and the proposal state, $\{T_1', T_2', T_3'\}$. The sample of ultrametrics $S = \{d_1, \ldots d_n\}$.*
- *Output: Decision to accept or reject the proposal.*

```
Compute ultrametrics $u_1, u_2, u_3$, from $T_1, T_2, T_3$, respectively.
Compute ultrametrics $v_1, v_2, v_3$, from $T_1', T_2', T_3'$, respectively.
Compute $\Pi_{\Phi_{u_1,u_2,u_3}}(S)$ and $\Pi_{\Phi_{v_1,v_2,v_3}}(S)$.
Set $p = \min\{1, \Pi_{\Phi_{u_1,u_2,u_3}}(S)/\Pi_{\Phi_{v_1,v_2,v_3}}(S)\}$.
Accept a proposal $\{T_1', T_2', T_3'\}$ with probability $p$.
```

With Algorithms 6 and 7, we have the following MCMC algorithm.

Algorithm 8 (MCMC algorithm to estimate the second order principal components)

- *Input: Sample of equidistant trees $\{T_1, \ldots, T_n\}$. Constant positive integer $C > 0$.*
- *Output: Second order principal components $\{T_1^*, T_2^*, T_3^*\}$.*

```
Set $S := \{d_1, \ldots, d_n\}$ where $d_i$ is the ultrametrics computed from a tree $T_i$, for $i = 1, \ldots, n$.
Pick random trees $\{T_0^1, T_0^2, T_0^3\} \subset \{T_1, \ldots, T_n\}$ and compute ultrametrics $u_1^*, u_2^*, u_3^*$ respectively.
Set $i = 1$, $k = m$, where $m$ is the number of leaves.
repeat
    if <$i$ mod $C$ equals zero and $k > 0$> then
        set $k = k - 1$.
    end if
    Compute the proposal $\{T_1^1, T_1^2, T_1^3\}$ via Algorithm 6 with $\{T_0^1, T_0^2, T_0^3\}$ and $k$.
    Compute ultrametrics $u_1, u_2, u_3$, from $T_1^1, T_1^2, T_1^3$, respectively.
    if <Algorithm 7 returns "accept"> then
        Set $T_0^1 = T_1^1$, $T_0^2 = T_1^2$, and $T_0^3 = T_1^3$.
    end if
```

```
        if <$\Pi_{\Phi_{u_1,u_2,u_3}}(S) < \Pi_{\Phi_{u_1^*,u_2^*,u_3^*}}(S)$> then
            set $u_1^* := u_1$, $u_2^* := u_2$, $u_3^* := u_3$.
        end if
        Set $i = i + 1$.
    until <Converges>
    Return the ultrametrics $u_1^*, u_2^*, u_3^*$.
```

This thesis desires the best fitting second order tropical PCA base in 45 dimensions. We iterate the MCMC approach in Algorithm 8 10,000 times and yield the PCA base that produces the lowest sum of all tropical distances.

Heating and Cooling

Due to the complexity of high dimensional space, the modification process within our MCMC algorithm occasionally gets stuck searching. Getting stuck refers to not being able to improve our current location or state. To alleviate this problem, we implement hot and cold chains. Cooling down occurs as we slowly diminish the severity of the modification process. This allows the trees to settle into a natural state. We heat the search by resampling the set of gene trees in our tropical base. Our MCMC function has adjustable heating and cooling parameters. Through heating and cooling we find the best-fitting tropical PCA base that minimizes the sum of all tropical distances.

4.3 Visualization and Analysis

Mesquite provides the simulated species and gene trees, while the MCMC algorithm provides the best-fitting tropical base. We hypothesize that gene trees projections will cluster by species as the ratio (r) increases. There is a ratio r for each set of species and gene tree simulations across both sets of simulations. We use two methods to measure the ability of tropical PCA to represent gene tree clusters by species. The first method is to visualize the gene tree clusters, at each ratio (r), through two dimensional plots. The second method is to measure the tropical proportion of variance, or r-squared, for each second order tropical PCA.

4.3.1 Two Dimensional Visualization

The easiest way to determine the effectiveness of tropical PCA is to project and plot the gene trees. If tropical PCA is successful, the gene trees will visibly group by species. We expect that gene tree clusters are more pronounced as the ratio (r) increases. A larger ratio (r) results from an increase in total species (tree) depths, because we are fixing the effective population size. An increase in total species (tree) depths provides the species more time (generations) to evolve from their common ancestor. Allowing more time for species to evolve facilitates more time for gene mutations, thus larger differences in shared genes between species in the same species tree.

In this thesis, we visualize two sets of simulations with known species trees. This allows us to plot each gene tree and attribute it to a specific species tree. We plot all available gene trees at each ratio (r) across both sets of simulations. For example; in the first set of simulations, we anticipate seeing a jumble of the 2000 gene trees at ratio r = 0.25. However; at a ratio of r = 10, we expect to see two distinct gene tree clusters. In the second set of simulations, we anticipate seeing eight distinct gene tree clusters at ratio r = 10.

Visualization of tropical space is not as intuitive as Euclidean space. Plotting in tropical space requires the construction of connected tropical line segments to define the tropical PCA base prior to plotting the gene trees. In both sets of simulations, the tropical PCA base is a 2-dimensional representation of a 45-dimensional space. As a result, the number of tropical line segments is significantly more than Figure 3.7 in Chapter 3. The resulting visualization has an indiscernible number of line segments.

4.3.2 Tropical Proportion of Variance

To analyze the fit of tropical PCA on the observed data we use the fraction of variance of unexplained and the coefficient of determination. The tropical geometry of the space requires us to define a non-Euclidean fraction of variance of unexplained. We define the fraction of variance of unexplained as:

$$\frac{\Pi_{\Phi}(S)}{\Pi_{\Phi}(S) + SS_{reg}}$$

SS_{reg} is the ”explained sum of squares” and is defined as:

$$SS_{reg} = \sum_{i=1}^{n} d_{tr}(\hat{u}_i, \bar{u})$$

Where $\hat{u}_i$ is the tropical projection of an ultrametric u_i for a tree T_i in the input sample onto a tropical polytope and $\bar{u}$ is a Fermat Weber point of $\{\hat{u}_i, \ldots, \hat{u}_n\}$. A Fermat Weber point $\bar{u}$ of $\{\hat{u}_i, \ldots, \hat{u}_n\}$ is defined as:

$$\bar{u} = \arg\max_{u} \sum_{i=1}^{n} d_{tr}(u, \bar{u}_i)$$

Additional details on Fermat Weber points may be found at (Lin and Yoshida 2018).

We use the unexplained variance to calculate the coefficient of determination, or R^2, and define it as:

$$R^2 = 1 - \frac{\Pi_\Phi(S)}{\Pi_\Phi(S) + SS_{reg}} = \frac{SS_{reg}}{\Pi_\Phi(S) + SS_{reg}}$$

This thesis uses the Fermat-Weber distance described in (Yoshida et al. 2019) as the R^2 measure. A larger R^2 value, between 0-1, indicates that the tropical PCA base is better at reducing dimensionality. It is important to remember that multiple factors, in addition to the fit of the tropical PCA base, such as the number of gene trees, affect the tropical R^2 value.

4.3.3 Empirical Data

This thesis also uses real world data to determine the effectiveness of tropical PCA in visualizing gene tree clusters to gain inferential species topology insights. Using tropical PCA on real world data is an unsupervised learning problem because we will never know the true species topology. Species topology inference works by associating each gene tree projection with its location on the two-dimensional tropical PCA triangle (base). The location directly corresponds to a plausible species tree topology. We record the location of each gene tree projection and the corresponding species tree topology, the sum of which provides insights for the true species topology.

We have two empirical data sets, African coelacanth (Liang et al. 2013) and Apicomplexa genome data (Kuo et al. 2008a). The African coelacanth genome data consists of 1193 gene trees, each with 10 leaves. The Apicomplexa data set consists of 252 gene trees, each with 8 leaves. After projecting the datasets on to their best fitting tropical PCA base, we record the gene tree projection locations and infer plausible species topologies.

4.4 BHV PCA

This thesis compares tropical PCA results to BHV PCA results from chapter 3 (Nye et al. 2017). We apply both sets of simulation data to `GeoPhytter+`, a java software program created by Tom Nye (Nye 2016). `GeoPhytter+` conducts higher order principal component analysis in BHV tree space utilizing the locus of the Fréchet mean (Nye et al. 2017). `GeoPhytter+` results are also visualized in R (Nye 2016).

The purpose of utilizing BHV PCA as a basis of comparison is not to prove which technique is better, tropical or BHV PCA. Rather BHV PCA is a proven technique, and makes for a good measuring stick for which we wish to gauge the effectiveness of our MCMC tropical PCA algorithm.

Locus of the Fréchet Mean PCA Base

In order to conduct BHV PCA we first determine the locus of the Fréchet mean PCA base (Nye et al. 2017). We input the simulated gene trees for each ratio (r) level, and use the `FitLFMTriangle` function within `GeoPhytter+` to search for the best-fitting base (Nye 2016). `GeoPhytter+` outputs the best locus of the Fréchet mean base, consisting of three trees. Additionally, `GeoPhytter+` outputs gene tree projections onto the locus of the Fréchet mean and produces a R^2 statistic (Nye 2016).

Locus of the Fréchet Mean Visualization

Once we have the best-fitting locus of the Fréchet mean PCA base, the next step is to compute the gene tree topologies for plotting. We input the Locus of the Fréchet Mean PCA Base into the `DecomposeLFMTriangle` function within `GeoPhytter+` (Nye 2016). This function attempts to identify the topologies within the PCA base and outputs a summary of all possibilities (Nye 2016). With the gene tree projections and topologies in hand, we use R (Ihaka and Gentleman 1993) to visualize the gene trees and their species topologies.

CHAPTER 5:
Results

This thesis aims to answer the question, can we visualize gene tree clusters by distribution to infer species topologies? We hypothesize that through the development of our tropical PCA algorithm, we more efficiently and effectively visualize gene tree clusters by distribution (species). Prior to inferring species topologies, we examine the effectiveness of our tropical PCA algorithm on reducing gene tree dimensionality.

Before discussing results, we want to highlight two important concepts for the reader to keep at the forefront. First, multiple factors in biology affect the degree to which gene trees cluster, most of which are outside the scope of this thesis. The only clustering factor this thesis explores is the ratio (r) between the total species (tree) depths and the species' effective population size.

$$\text{ratio (r)} = \frac{\text{Total Species (Tree) Depths}}{\text{Eff Pop Size}}$$

Additionally, tropical PCA is a dimension reduction technique and does not assist with clustering in any way. This thesis implements a second order tropical PCA to reduce gene tree dimensionality for the purpose of easy visualization and intuitive analysis.

5.1 Simulation Experiments

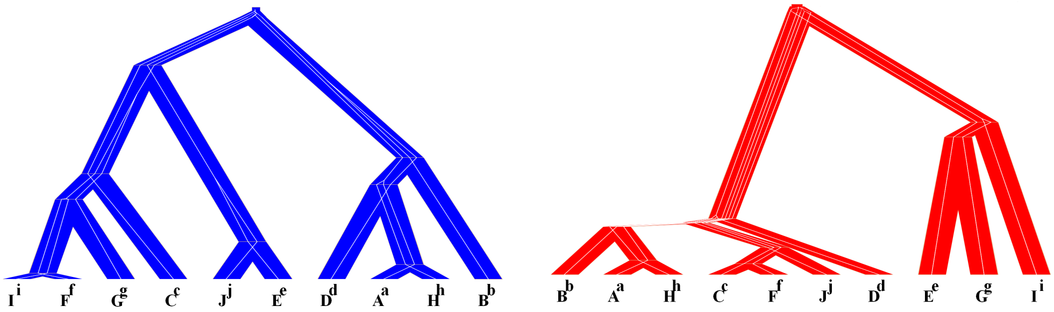
This thesis conducts two sets of simulation experiments to measure the effectiveness of our Tropical PCA algorithm. In both simulation experiments, we explore gene tree clustering at various species (tree) depths and fix effective population size. As a result, both simulation experiments consist of six mini experiments, one mini experiment for each ratio $r = 0.25, 0.5,\ 1,\ 2,\ 5,\ 10$.

The goal of the simulation experiments is visually and analytically gauge how well our tropical PCA is performing. We expect to see gene tree clusters becoming more distinct as the ratio (r) increases. We visualize the results from each experiment with a series of plots to determine effectiveness. We also calculate a R^2 value at each ratio (r). As a basis for comparison, we also implement a BHV PCA onto the locus of the Fréchet mean (Nye et al. 2017), using the same dataset. We compare the BHV PCA results to those from our tropical PCA application.

5.1.1 First Simulation Experiment

The first simulation experiment consists of using two thousand gene trees contained within two different species trees, a thousand per species tree. Two different species trees and completely different gene trees are used for each mini experiment we conduct. Each mini experiment implements our tropical PCA and a BHV PCA on the two thousand gene trees corresponding to that ratio (r) value. After the best-fitting tropical base or locus of the Fréchet mean is determined, each mini experiment plots the gene tree projections by color and calculates the R^2 value.

Since we are plotting two thousand projections onto an unfamiliar two-dimensional tree space, either tropical or BHV, we color code the gene tree projections by the species tree it is contained within. Figure 5.1 shows the two species trees we use in the tropical PCA mini experiment for ratio $r = 1$. For all six tropical PCA mini experiments, we represent Species Tree 1 with the color blue and Species Tree 2 with the color red, as depicted in Figure 5.1. Both species trees and all gene trees (contained within each species tree) have the same leaf labels, $A - J$ and $a - j$. However, the trees themselves are all completely different.



(a) Species Tree 1

(b) Species Tree 2

Figure 5.1. First Simulation Experiment Species Trees At Ratio $r = 1$

We use Algorithm 9 to compute a tropical PCA on the mixture of two thousand gene trees from the two distributions (species trees) for each ratio (r). We implement Algorithm 9 for each tropical PCA mini experiment in the first simulation experiment.

Algorithm 9 (PCA with two distributions of gene trees from different species trees)

- *Input:* $S_1 := \{T_1, \ldots, T_{1000}\}$*, a sample of 1000 gene trees generated with a species tree* T_{S_1}*. And* $S_2 := \{T'_1, \ldots, T'_{1000}\}$*, a sample of 1000 gene trees generated with a species tree* T_{S_2} *where* $T_{S_1} \neq T_{S_2}$*.*
- *Output: PCA (Second Order Principle Components) with a sample* $\{T_1, \ldots, T_{1000}, T'_1, \ldots, T'_{1000}\}$*.*

Apply a tropical PCA and compute the second order principal components with $\{T_1, \ldots, T_{1000}, T'_1, \ldots, T'_{1000}\}$*.*
Project the points of $\{T_1, \ldots, T_{1000}\}$ *onto the tropical PCA and plot the points in blue for easy visualization.*
Project the points of $\{T'_1, \ldots, T'_{1000}\}$ *onto the tropical PCA and plot the points in red.*

Understanding Tropical PCA Plots

Figure 5.2 shows the tropical PCA results for the first simulation experiment. Before discussing results we are going to describe tropical PCA plots. The tropical PCA plots, in Figure 5.2, consist of tropical line segments and tropical gene tree projections. The tropical line segments are the black lines and dots, comprising the tropical PCA triangle base. The original gene trees are 45-dimensional, hence there are a significant number of tropical line segments throughout the plots. Think of the tropical line segments as providing the two-dimensional structure for the tropical tree space. Each box created within tropical line segment

The axes on all of our tropical plots are labeled D_base[1,] and D_base[2,]. These axes represent the $\pi_{\mathcal{P}}(D)$ projection positions on the tropical polytope base in two-dimensional tree space. Each of the $\pi_{\mathcal{P}}(D)$'s are vectors of length three, when conducting a second order tropical PCA. In example 5, from Chapter 3, the projection $\pi_{\mathcal{P}}(D)$ was equal to $(1.4, 2, 2)$.

Issues arise when projecting $(1.4, 2, 2)$, which is a three-dimensional vector, onto a two-dimensional tree space. To remedy this, we subtract (c, c, c) from all the gene trees projections prior to plotting onto our two-dimensional tropical polytope base. The goal is to subtract the specific (c, c, c) so the first index position in the $\pi_{\mathcal{P}}(D)$ vector is zero. If the first index position in every $\pi_{\mathcal{P}}(D)$ vector is always zero, then all of the gene tree projections become two-dimensional, because the first dimension is zero for all of the projections.

Example 10 *Let $\pi_{\mathcal{P}}(D)$ = the three dimensional vector (1.4,2,2).*

$$\begin{aligned}
\pi_{\mathcal{P}}(D) &= (1.4, 2, 2) \\
(1.4, 2, 2) - (c, c, c) &= (0, 0.6, 0.6),\ when\ c = 1.4 \\
\pi_{\mathcal{P}}(D) &= (0, 0.6, 0.6)
\end{aligned}$$

$$\begin{aligned}
\pi_{\mathcal{P}}(D) &= (0, \mathit{0.6}, \mathit{0.6}) \\
D\_base[1,] &= \mathit{0.6} \\
D\_base[2,] &= \mathit{0.6} \\
\pi_{\mathcal{P}}(D) &= (0.6, 0.6)
\end{aligned}$$

Now the tropical projection $\pi_{\mathcal{P}}(D)$ is two-dimensional.

The tropical plots have varying axes values for D_base[1,] and D_base[2,], ranging from $[-2, 2]$ and $[-2, 2]$, depending on the plot. This is because the tropical plot function is only creating and connecting tropical line segments for the specific ranges of the tropical triangle base, not all of tropical tree space.

Within each tropical PCA plot is a mixture of two thousand gene trees from two different species distributions as blue or red dots. For the purposes of this thesis, we are not comparing or analyzing the location of the gene tree clusters within each plot or from plot to plot. Our only interest, when examining the plots, is how effective our MCMC tropical PCA approach is at capturing the gene tree clustering, at each ratio (r). Less formally, how well do the blue and red dots cluster in Figure 5.2. Chapter 6 addresses follow on research opportunities that could explore the relationship between gene trees and projection locations in tropical tree space.

Understanding BHV PCA Plots

Figure 5.3 shows the BHV PCA results for the first simulation experiment. Before discussing result specifics we are going to discuss interpretation of BHV PCA plots. The BHV PCA plots, in Figure 5.3, use a triangular background shape to serve as the locus of the Fréchet mean, which all gene trees project onto. The color coded regions, of various shapes and sizes within each triangle, correspond to different inferential species topologies associated with the gene tree projections specific to each plot (Nye 2016) (Nye et al. 2017).

The black and yellow dots represent the BHV PCA gene tree projections. Ideally for all six BHV PCA mini experiments, the two thousand BHV PCA gene tree projections would be blue and red to depict their respective species trees; instead, they are unfortunately black and yellow. The color discrepancies are due to *GeoPhytter+*, the BHV PCA implementation software, limitations (Nye 2016). For the purposes of this thesis, we are only focusing on how well the black and yellow dots separate into clusters. For the first simulation experiment BHV PCA plots, $Blue = Black$ and $Red = Yellow$.

We want to emphasize the purpose of implementing the BHV PCA is to serve as a basis of comparison, and not to prove which technique is better. Rather BHV PCA is a proven technique that effectively captures gene tree clusters, thus makes for a good measuring stick to gauge the effectiveness of our tropical PCA algorithm. When viewing the BHV PCA plots, the question we ask ourselves is how well do the blue and red dots cluster in Figure 5.2 when compared to the black and yellow dots in Figure 5.3 at a similar ratio (r)?

First Simulation Experiment Results

The second order tropical PCA results in Figure 5.2 look great. Each two-dimensional plot successfully visualizes two thousand gene trees, which exist in 45-dimensions, as blue or red dots. As the ratio (r) increases the blue and red dots clusters in the tropical PCA plots become more pronounced, which is exactly what we are looking for. At $r = 0.25$ and $r = 0.5$, the blue and red dots are completely spread out. At the ratio $r = 1$ and $r = 2$, the blue and red dots are in distinguishable groups with few outliers. At $r = 5$ and $r = 10$, the tropical PCA plots show complete separation of the blue and red dots into tight clusters, which represent distinct species distributions. The results in Figure 5.2 showcase the effectiveness of our tropical PCA application.

The BHV PCA results in Figure 5.3 look less pronounced than the tropical PCA results in Figure 5.2. To reiterate, we are focusing solely on the clustering of the black and yellow dots within the larger triangle. Likewise, we are not focusing on the location of black and yellow dots within the BHV PCA smaller color coded regions. At $r = 0.25$ and $r = 0.5$, the BHV PCA plots in Figure 5.3 show fairly spread out black and yellow dots, which is similar to our tropical PCA results and what we expect. The black and yellow dots for $r = 1$, $r = 2$ and $r = 5$ have a degree of clustering; however, there is a nontrivial quantity of black and yellow scattering in each plot. The BHV PCA clustering for ratio $r = 10$ looks intermittent.

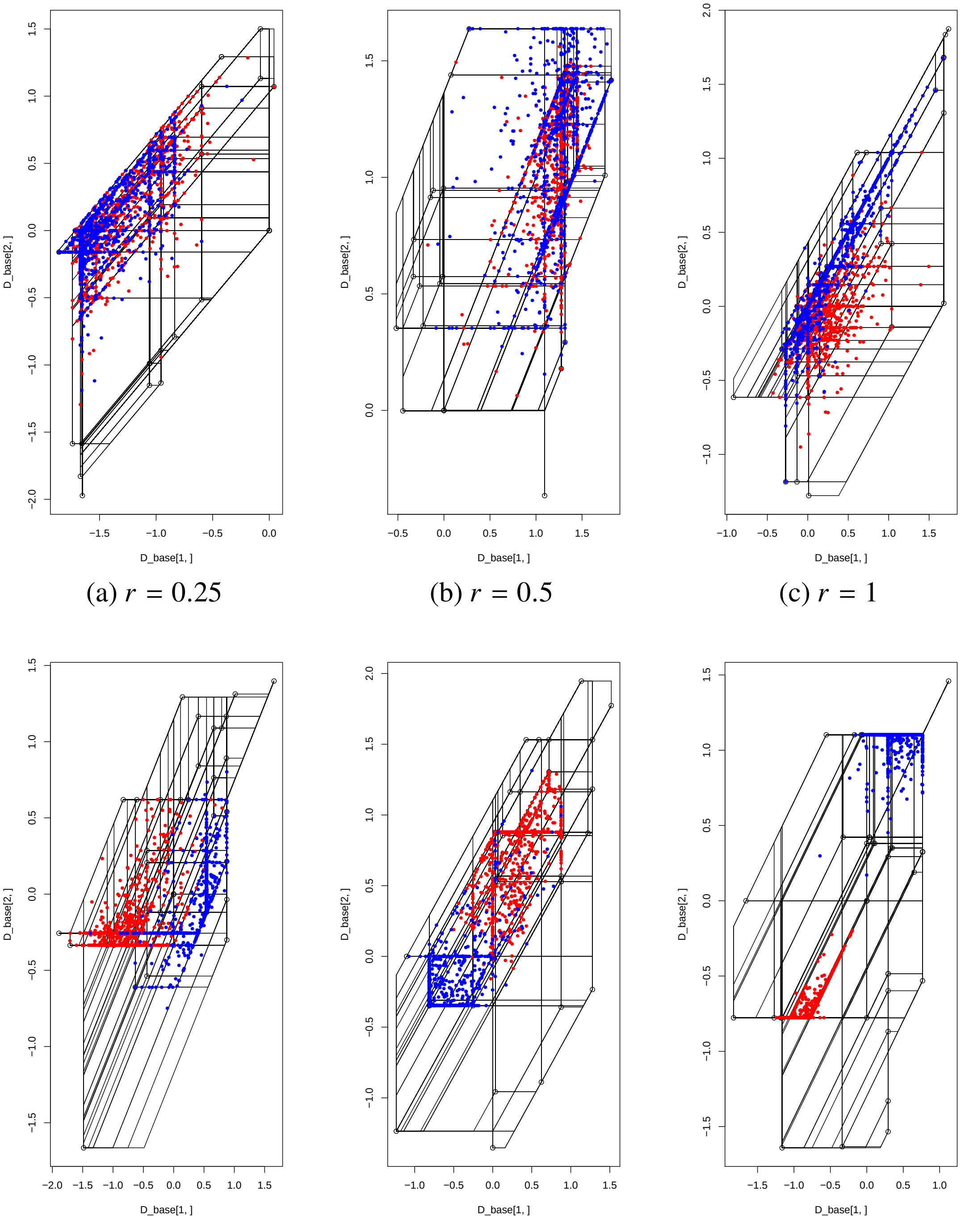
r	Tropical PCA	BHV PCA
0.25	0.316	0.009
0.5	0.297	0.009
1	0.186	0.042
2	0.319	0.034
5	0.278	0.009
10	0.396	0.009

Table 5.1. First Simulation Experiment R^2

The yellow dots do form a loose clustering at the bottom of the BHV PCA plot for ratio $r = 10$; however, the black dots are extremely spread out.

Comparing the tropical PCA and BHV PCA results in Figure 5.2 and Figure 5.3 is challenging. The results from the tropical PCA plots are not easily transferable to the BHV PCA plots for comparison. Which is why, this thesis also analyzes R^2 values for both PCA applications. Table 5.1 shows the tropical PCA and BHV PCA R^2 values for the first simulation experiment. Tropical PCA R^2 is higher than BHV PCA at every ratio (r).

The quality of the tropical PCA plots and R^2 values across the first simulation experiment support our hypothesis. While not definitive at this point, the tropical PCA plots and R^2 values speak to the effectiveness of our tropical PCA algorithm. To further examine the effectiveness of our tropical PCA approach on gene tree clustering we conduct a second, more extensive, set of simulation experiments.



(a) $r = 0.25$

(b) $r = 0.5$

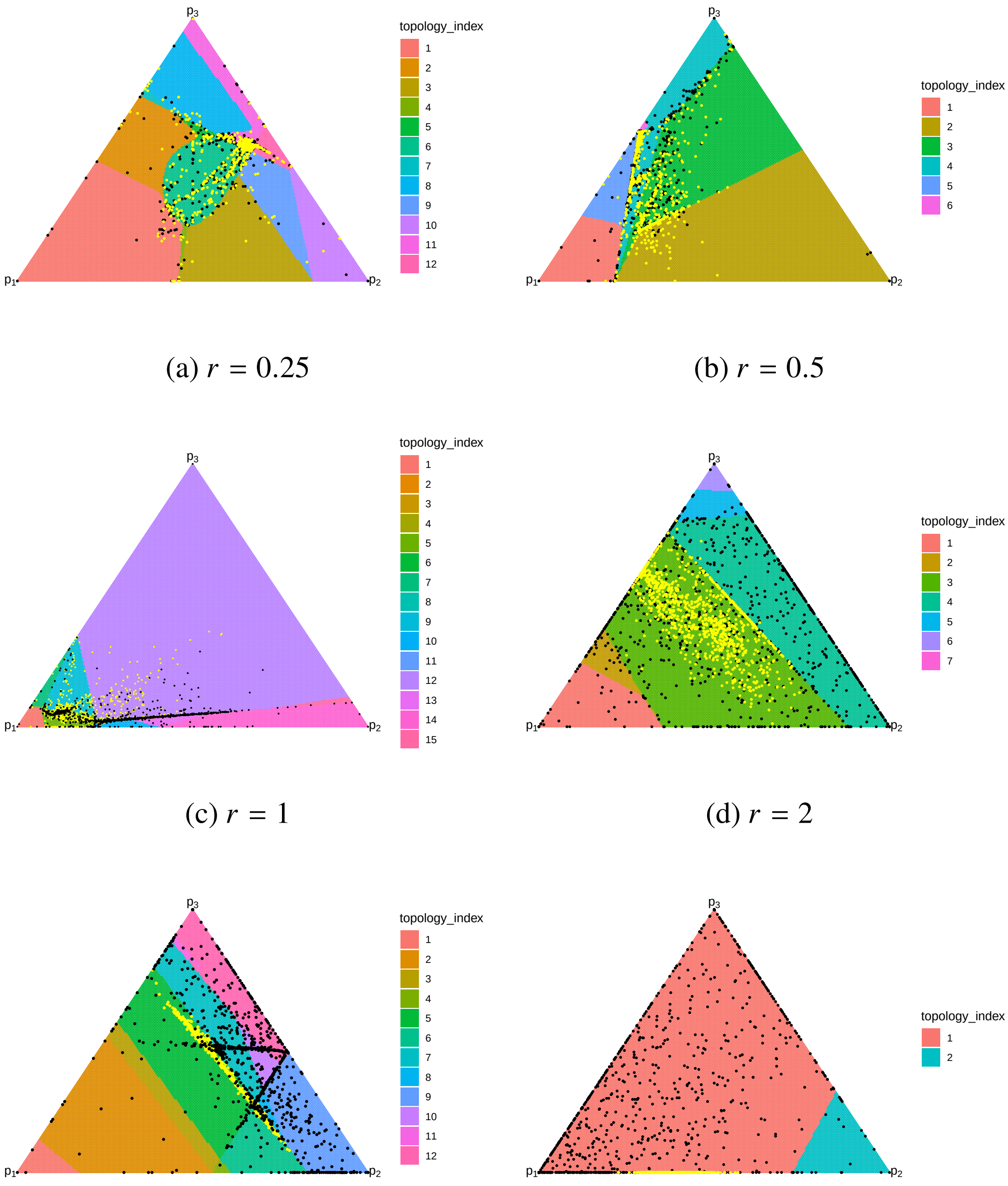
(c) $r = 1$

(d) $r = 2$

(e) $r = 5$

(f) $r = 10$

Figure 5.2. First Simulation Experiment Tropical PCA Results



(a) $r = 0.25$

(b) $r = 0.5$

(c) $r = 1$

(d) $r = 2$

(e) $r = 5$

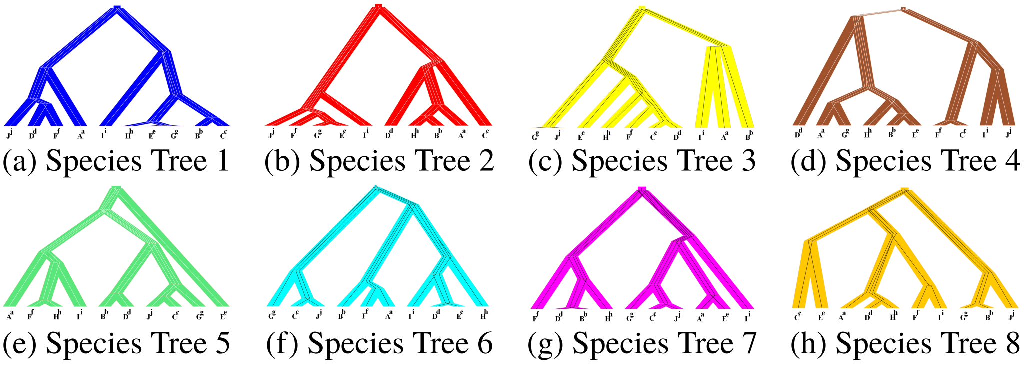
(f) $r = 10$

Figure 5.3. First Simulation Experiment BHV PCA Results

5.1.2 Second Simulation Experiment

We expand the second simulation experiment to consist of eight thousand gene trees contained within eight different species trees, a thousand per species tree. Similar to the first simulation experiment, each of the six mini second simulation experiments implements our tropical PCA and a BHV PCA on the eight thousand gene trees corresponding to the given ratio (r). We generate completely new species and gene trees for each mini experiment we conduct for our second simulation experiment. After the best-fitting tropical base or locus of the Fréchet mean is determined, each mini experiment plots the gene tree projections by color and calculates the R^2 value.

Since the second simulation experiment plots are crowded with eight thousand gene tree projections (dots), we color code the gene tree projections (dots) by the species tree it is contained within. Unfortunately, the second simulation experiment tropical PCA gene tree projections (dots) are different colors than the BHV PCA projections (dots). Figure 5.4 shows the eight species trees we use in the mini experiment for ratio $r = 2$. Although each mini experiment's species trees are different due to the varying ratio (r), the color of the species trees and corresponding gene tree projections (dots) does not change. In our tropical PCA plots the color coding is $SpeciesTree1 = Blue$, $SpeciesTree2 = Red$, $SpeciesTree3 = Yellow$, $SpeciesTree4 = LightBrown$, $SpeciesTree5 = LightGreen$, $SpeciesTree6 = Cyan$, $SpeciesTree7 = Magenta$ and $SpeciesTree8 = Orange$, as depicted in Figure 5.4.


(a) Species Tree 1 (b) Species Tree 2 (c) Species Tree 3 (d) Species Tree 4

(e) Species Tree 5 (f) Species Tree 6 (g) Species Tree 7 (h) Species Tree 8

Figure 5.4. Second Simulation Experiment Species Trees At Ratio $r = 2$

We use Algorithm 11 to compute a tropical PCA on the mixture of gene trees from the eight distributions.

Algorithm 11 (PCA with eight distributions of gene trees from difference species trees)

- *Input:* $S_i := \{T_1^i, \ldots, T_{1000}^i\}$*, a sample of gene trees generated with a species tree* T_{S_i} *for* $i = 1, \ldots, 8$*, where* $T_{S_1}, \ldots, T_{S_8}$ *have all distinct tree topologies.*
- *Output: PCA (second PCs) with a sample* $\{T_1^1, \ldots, T_{1000}^1, \ldots, T_1^8, \ldots, T_{1000}^8\}$.

Apply a tropical PCA and compute the second order principal components with $\{T_1^1, \ldots, T_{1000}^1, \ldots, T_1^8, \ldots, T_{1000}^8\}$.
Project the points of $\{T_1^i, \ldots, T_{1000}^i\}$ *onto the tropical PCA and plot the points in a distinct color for ease of visualization.*

Second Simulation Experiment Results

The second order tropical PCA results in Figure 5.5 look great. Each two-dimensional plot successfully visualizes eight thousand gene trees, which exist in 45-dimensions, as a color coded dot. The tropical PCA plots, in Figure 5.5, clearly illustrate as the ratio (r) increases, the clustering by color becomes more pronounced. At $r = 0.25$ and $r = 0.5$, the gene tree projections, or dots, are completely spread out and clusters are indistinguishable. At the ratios $r = 1$ and $r = 2$, the color coded dots form loose groupings with some outliers. At $r = 5$ and $r = 10$, the tropical PCA plots show complete separation of the color coded dots into tight clusters, which represent the eight distinct species distributions. The results in Figure 5.5 are further evidence of the effectiveness of our tropical PCA application.

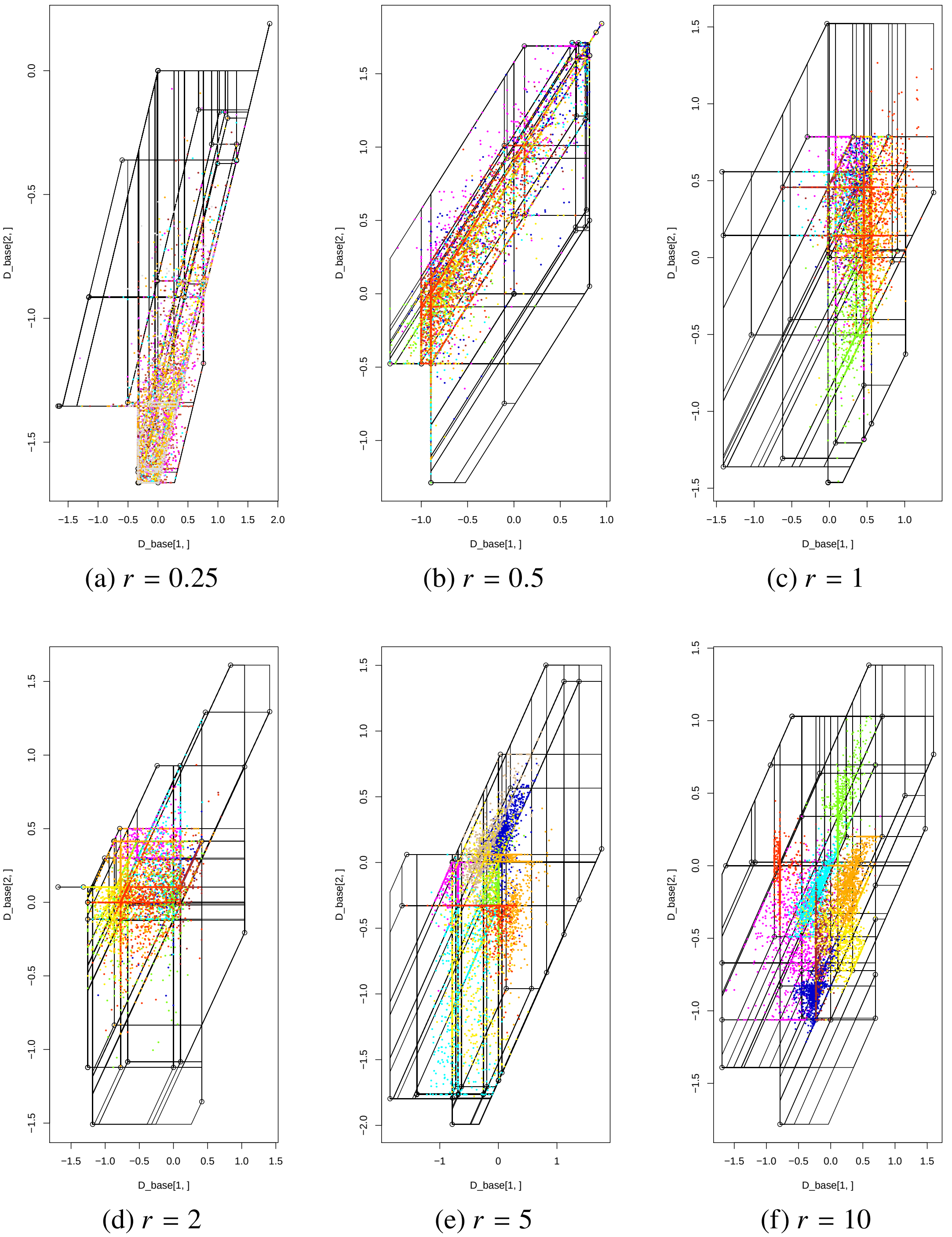
The BHV PCA results in Figure 5.6 look equally as impressive as the tropical PCA results in Figure 5.5. The second simulation experiment BHV PCA plot color coding is almost identical to the tropical PCA plot color coding, $Blue = White$, $Red = Red$, $Yellow = Yellow$, $Pink = LightBrown$, $LightGreen = LightGreen$, $Cyan = Cyan$, $Magenta = Magenta$ and $Orange = Orange$. At $r = 0.25$ and $r = 0.5$, the BHV PCA plots in Figure 5.6 indicate minimal clustering, which is similar to our tropical PCA results and what we expect. While the BHV PCA color coded gene tree projections (dots) at $r = 1$ and $r = 2$ display slight clustering, a large amount of dispersion exists within each color. The BHV PCA clustering for ratios $r = 5$ and $r = 10$, in Figure 5.6, looks as good if not better than our tropical PCA results. The BHV PCA clustering for ratio $r = 10$ specifically looks phenomenal.

r	Tropical PCA	BHV PCA
0.25	0.321	0.031
0.5	0.262	0.004
1	0.205	0.023
2	0.228	0.061
5	0.220	0.424
10	0.241	0.569

Table 5.2. Second Simulation Experiment R^2

Table 5.2 shows the R^2 values for each tropical PCA and BHV PCA mini second simulation experiment. Tropical PCA produces strong R^2 values for all ratio (r) values. BHV PCA appears to struggle at lower ratio (r) values, however at $r = 5$ and $r = 10$ the BHV PCA R^2 is notably high. The high BHV PCA R^2 values, for ratios $r = 5$ and $r = 10$, were evident by the tight clustering in Figure 5.6. Unfortunately, we do not have specific insights explaining why BHV PCA performs well at ratios $r = 5$ and $r = 10$.

The quality of the second simulation experiment results our tropical PCA algorithm produces, shown in Figure 5.3 and Table 5.2, clearly supports our hypothesis. The ability of our second order tropical PCA approach to effectively reduce high dimensional gene trees and visualize clustering in two-dimensional plots is evident by our results.



(a) $r = 0.25$

(b) $r = 0.5$

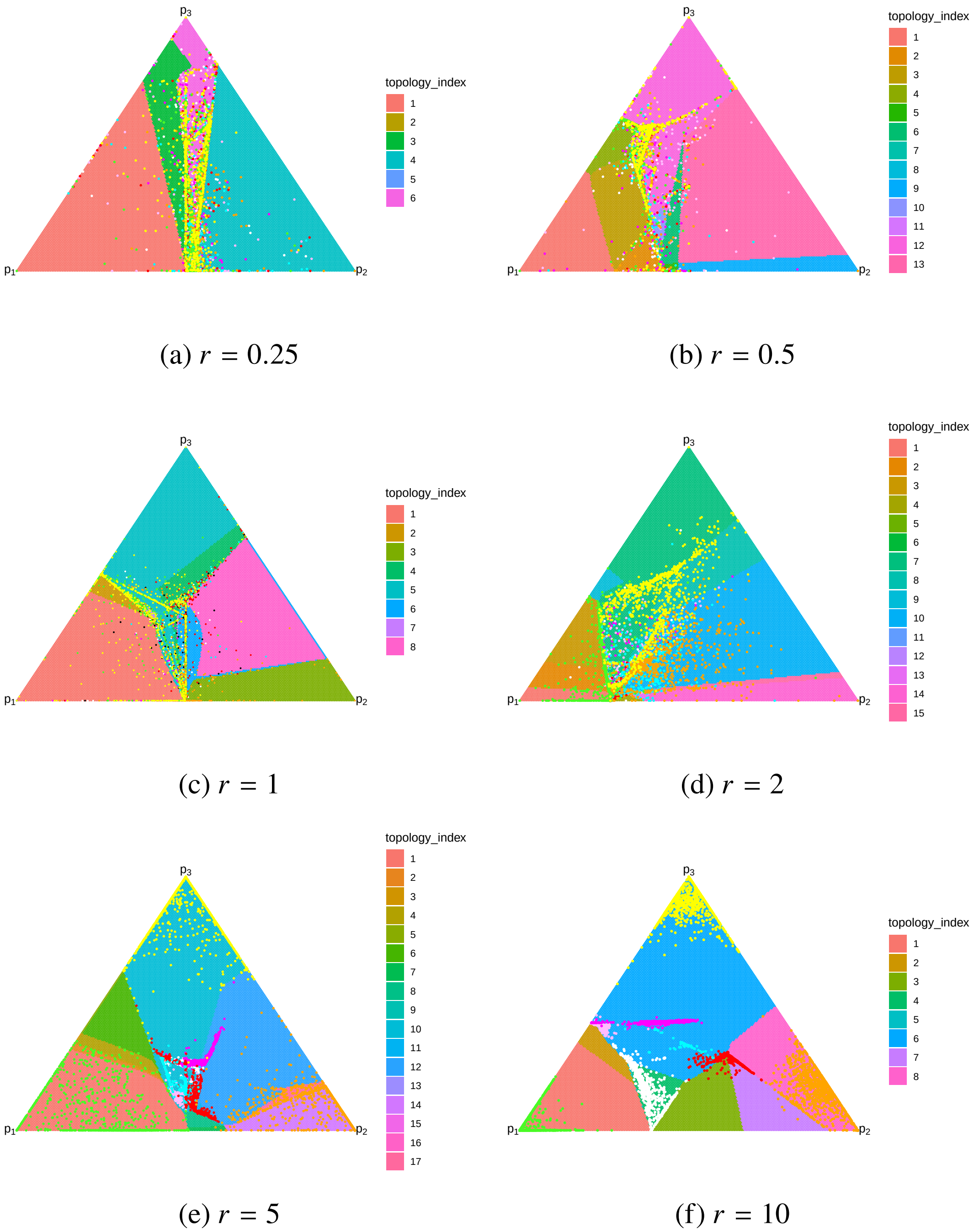
(c) $r = 1$

(d) $r = 2$

(e) $r = 5$

(f) $r = 10$

Figure 5.5. Second Simulation Experiment Tropical PCA Results



(a) $r = 0.25$

(b) $r = 0.5$

(c) $r = 1$

(d) $r = 2$

(e) $r = 5$

(f) $r = 10$

Figure 5.6. Second Simulation Experiment BHV PCA Results

5.2 Empirical Data

In addition to the simulation experiments, we apply our tropical PCA approach to two empirical datasets: Apicomplexan gene trees (Kuo et al. 2008a) and the African coelacanth genome (Liang et al. 2013). We select these empirical datasets because of their real world relevancy. Studying the African coelacanth dataset may uncover insights regarding the evolutionary journey of tetrapods (Liang et al. 2013). Studying the evolutionary path of Apicomplexa may further our understanding of disease mutations. Both empirical datasets are relatively small in comparison to the eight thousand gene tree mixture in our second simulated experiment. These datasets provide us the opportunity to determine the effectiveness of our tropical PCA approach in clustering gene trees and pose an unsupervised learning problem for species topology inference.

5.2.1 Understanding Empirical Plots

A clear distinction must be made between the simulation experiment plots and the empirical plots. Each empirical dataset contains exactly one species tree. Additionally, we do not know either empirical species tree, that which we believe contains the Apicomplexan or African coelacanth gene trees. Therefore, color coding the empirical gene tree projections (dots) by species tree (distributions), as we did for the simulation experiments, would be ineffective and pointless. The empirical dataset color coding would consist of one color for every dot in the plot, and would provide no inferential insights about the unknown species tree. This thesis does not color code empirical gene tree projections (dots) by their species distribution (tree).

The aim of this thesis is to use our tropical PCA approach on both empirical datasets to test how effectively it captures the clustering of the gene tree projections. Because there is only one species tree for each empirical dataset, albeit unknown, we expect to see only one cluster in our tropical PCA plots. If our tropical PCA approach works perfectly on an empirical dataset, and there is zero error, we would see one extremely tight cluster. The tight cluster represents the inferred species topology containing all of the gene trees in the empirical dataset. We could calculate the inferred species topology from the location of the cluster on the two-dimensional tropical base and map it back to 45-dimensions, where the inferred species tree exists

We did not focus on the projection locations for the simulation experiment plots, only clustering. Since we can not leverage a priori species tree knowledge to assist in coloring empirical projections, we need a new technique. This thesis color codes the empirical gene tree projections (dots) by their location on the two-dimensional tropical triangle base. We measure the distance between all gene tree projections and color code groupings of projections that are zero distance apart.

We use the *Robinson-Foulds Distance* between two gene tree projections as the distance measure. The *Robinson-Foulds Distance* measures the differences in branch lengths between two trees and sums the absolute values of those differences (Felsenstein 2004). Therefore, dots of the same color have a Robinson-Foulds distance of zero and dots of different colors have a Robinson-Foulds distance of greater than zero. For example, in Figure 5.7, the blue dots are all zero distance apart and blue and red dots are not zero distance apart.

The different dot colors in our tropical PCA plots indicate differences in the species tree topologies we infer.

The quantity of dots belonging to that particular color are written above each species tree diagram, in the appropriate color. For example, the yellow dots in Figure 5.7 all have the same tree topology when we map those projections back to 45-dimensions. The corresponding species tree topology inference plot, in Figure 5.8, has a yellow (68) above the tree topology. That topology is the inferred species tree topology for all of the yellow dots, for which there are sixty eight total gene tree projections.

It is important to note the appearance of black dots in the tropical PCA plots. Those black dots represent gene tree projections that we deem outliers, because of their proximity to all other gene tree projections. Outliers are gene tree projections which have a Robinson-Foulds distance of zero to a small portion of other gene tree projections. Outliers consist of atmost five percent of all gene tree projections. The outliers, or black dots, do not contribute to any species topology inferences.

Lastly, we inform the reader that we zoom in on the empirical tropical PCA plots, creating a smaller plot area. There is no change to the techniques and implementation for how we create empirical tropical plots. However, we expect only one cluster in our plots; therefore, we reduce the plot area and zoom into the cluster for better visibility. We inform the reader because zooming in on the cluster creates the illusion that the gene tree projections (dots) are more dispersed than they truly are. The axes for the empirical tropical PCA plots remain the same to aid orientation to the plot.

5.2.2 African Coelacanth Genome Data

African coelacanth or lungfish; from which did humans evolve? Insights from the African coelacanth and lungfish help biologists gain understanding on the evolution of vertebrates to land (Liang et al. 2013). As the availability of this data is relatively new, researchers are looking for methods through which they can connect the evolutionary dots to other species. Strong evidence supports that the lungfish is closer in relation to humans than the coelacanth (Amemiya et al. 2013).

The name of this dataset can be misleading, because the African coelacanth is only one of the ten species within the dataset. This dataset is sometimes referred to as the lungfish dataset. The intent of this dataset is to determine which marine vertebrate is the closest relative to land vertebrates. This data set consists of 1290 gene trees obtained from the genomic data of ten species (Liang et al. 2013).

This dataset has gene trees reconstructed from the sequences from the following ten species: *Human: Homo Sapien* (Homo), *Chicken: Gallus gallus* (Gallus), *Frog: Xenopus tropicalis* (Xenopus), *Zebrafish: Danio rerio* (Danio), *Fugu: Takifugu rubripes* (Takifugu), *African Coelacanth: Latimeria chalumnae* (Latimeria), *African lungfish: Protopterus annectens* (Lungfish), and three Cartilaginous fishes as the outgroup: *Callorhinchus milii* (Callorhinc), *Leucoraja erinacea* (Leucoraja), and *Scyliorhinus canicula* (Scyliorhin) (Liang et al. 2013).

Coelacanth Tropical Results

We apply our MCMC technique to conduct a second order tropical PCA on the African coelacanth genome dataset and infer a species topology. Our tropical PCA results are shown in Figure 5.7 and 5.8. The tropical PCA plot, in Figure 5.7, clearly visualizes the gene tree projections (dots) in a dense clustering. The gene tree projections (dots), in Figure 5.7, are part of one large clustering that primarily spans over eight tropical line segments.

Figure 5.8 illustrates the inferred species topologies for all of the gene tree projections with color coding. The inferred species tree topologies in Figure 5.8 look promising.

At first glance, having eight different inferred species tree topologies appears substandard. However, upon further inspection of Figure 5.8, the eight inferred species tree topologies are not very different. Both the leaves and the structure of all eight inferred species topologies look eerily similar.

In Figure 5.8, the three land vertebrates (humans, chickens and frogs) are closest in relation for all eight inferred topologies. Similarly, the three cartilaginous fishes are closest in relation for all eight inferred species topologies. Interestingly, our tropical PCA results are mixed for which the closest relative to land vertebrates. The fact that all eight inferred species topologies look similar is a good result. We will never know the true species topology; however, it is easier to gain evolutionary insights when all inferred species topologies are near identical.

We calculate the R^2 value for our tropical PCA approach on the African coelacanth dataset. The $R^2 = 0.325$ is favorable. The results of our empirical coelacanth tropical PCA, shown in Figure 5.7 and Figure 5.8, clearly supports our hypothesis. The ability of our second order tropical PCA approach to effectively reduce high dimensional gene trees and visualize clustering in two-dimensional plots is again evident by our results.

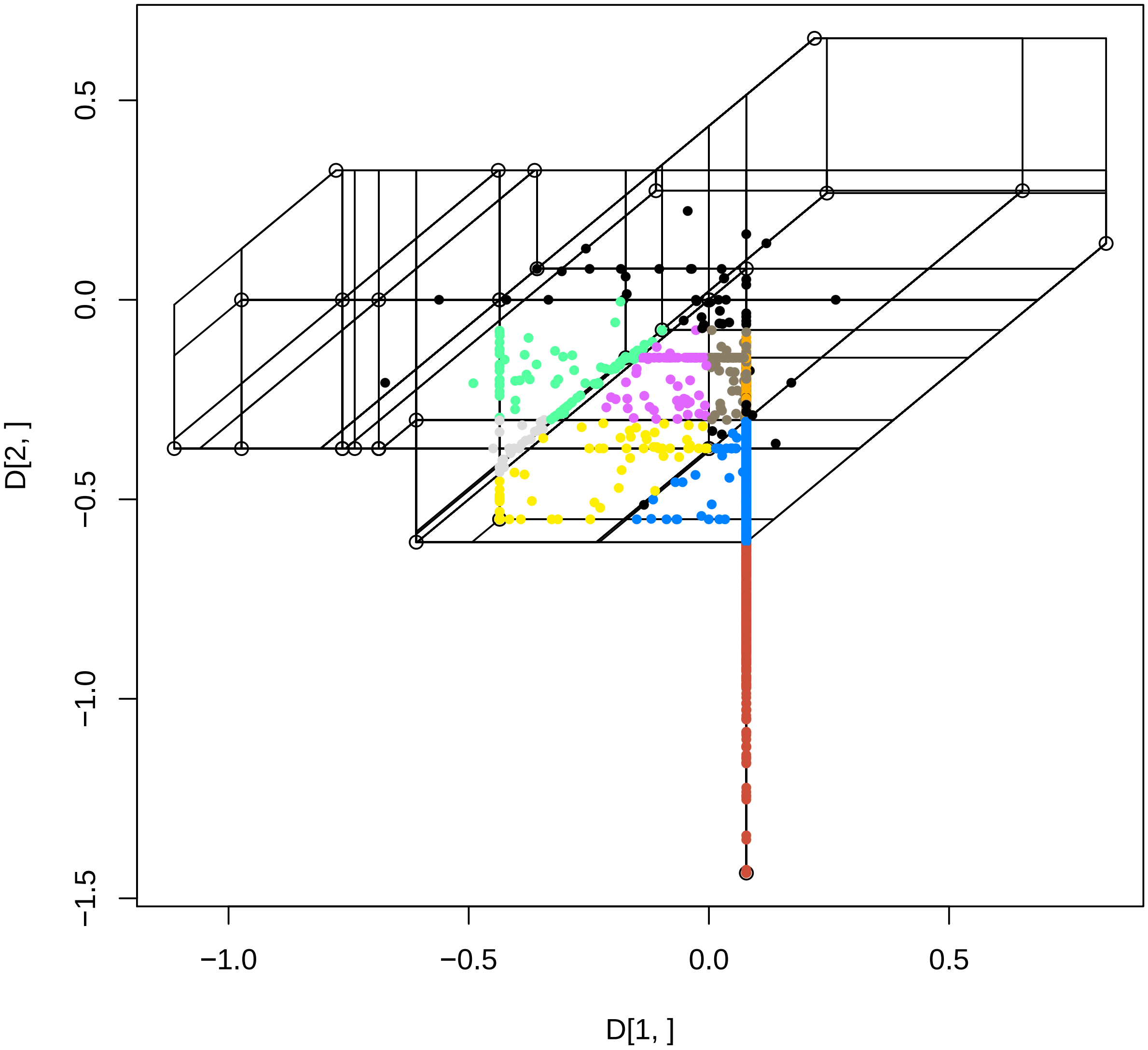


Figure 5.7. Tropical PCA for African Coelacanth

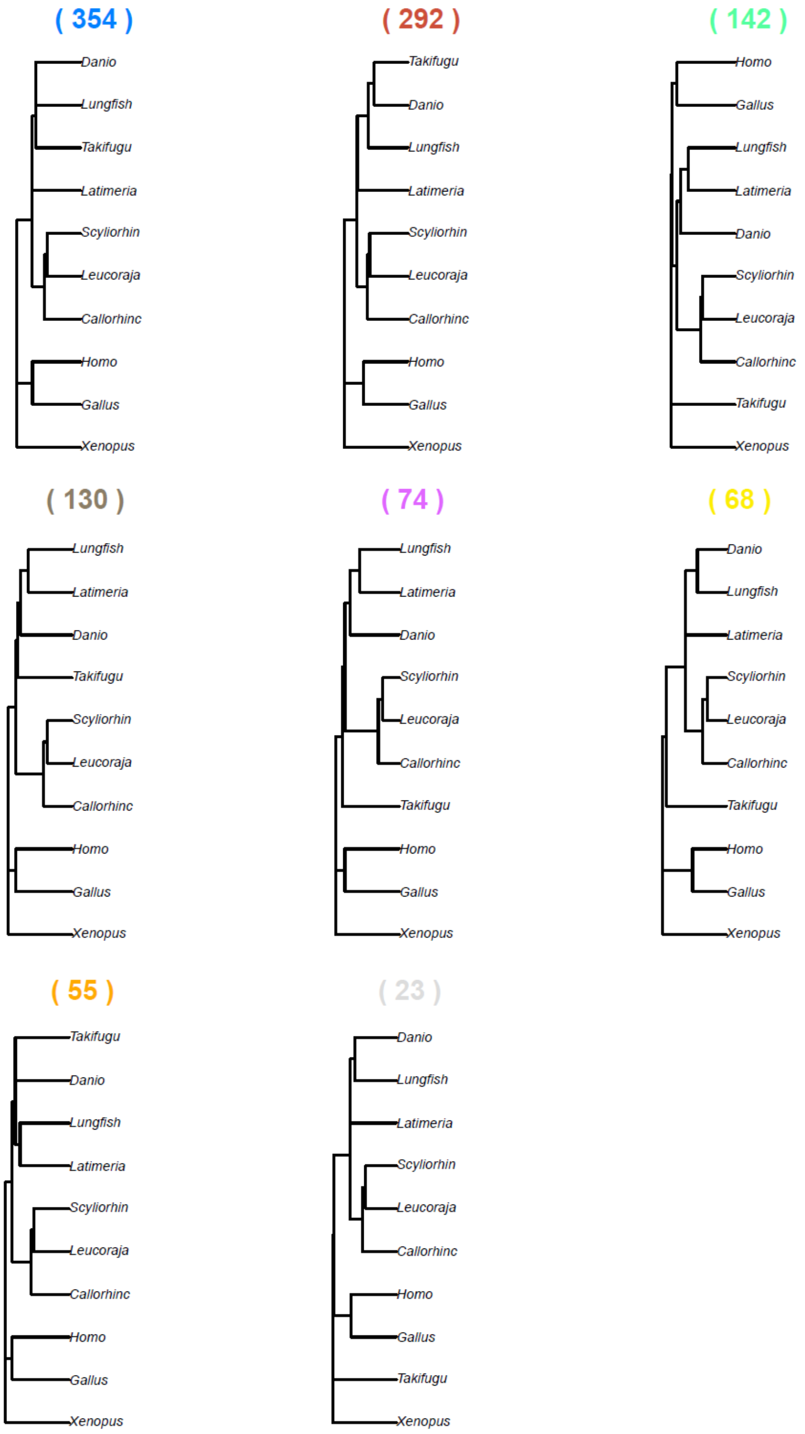


Figure 5.8. Inferred African Coelacanth Species Topologies

5.2.3 Apicomplexa Genome Dataset

The Apicomplexa phylum, contains multiple parasitic pathogens worthy of our understanding (Kuo et al. 2008a). The Apicomplexa genome dataset is our smallest dataset. This dataset consists of 268 gene trees with eight species of protoza present in (Kuo et al. 2008a). The 268 gene trees reside in 28-dimensional tree space.

This data set contains gene trees reconstructed from the following eight species: *Babesia bovis* (Bb), *Cryptosporidium parvum* (Cp), *Eimeria tenella* (Et), *Plasmodium falciparum* (Pf), *Plasmodium vivax* (Pv), *Theileria annulata* (Ta), *Toxoplasma gondii* (Tg) and an outgroup *Tetrahymena thermophila* (Tt) (Kuo et al. 2008a). We apply our MCMC technique to conduct a second order tropical PCA on the Apicomplexa dataset.

Apicomplexa Tropical Results

The second order tropical PCA results in Figure 5.9, from the Apicomplexa dataset, look great. Due to the small nature of the Apicomplexa dataset, we did not identify any outliers and instead provide color coding for all 268 gene tree projections. Figure 5.9 shows six distinct colors, all reasonably close to each other on the two-dimensional tropical base.

Figure 5.10 illustrates the inferred species topologies for all of the color coded gene tree projections. The Apicomplexa inferred species tree topologies look astounding. The inferred species topological branches and leaves depict a close relation between the inferential species topologies in Figure 5.10. In fact, the trees are so similar in Figure 5.10, that the differences are hard to spot. The inferred similarities between species trees support the true species tree being somewhere near one of the trees in Figure 5.10.

The $R^2 = 0.605$ is extremely favorable. The results of our empirical Apicomplexa tropical PCA, shown in Figure 5.9 and Figure 5.10, positively supports our hypothesis. The ability of our second order tropical PCA approach to effectively reduce high dimensional gene trees and visualize clustering in two-dimensional plots is clearly supported by our results.

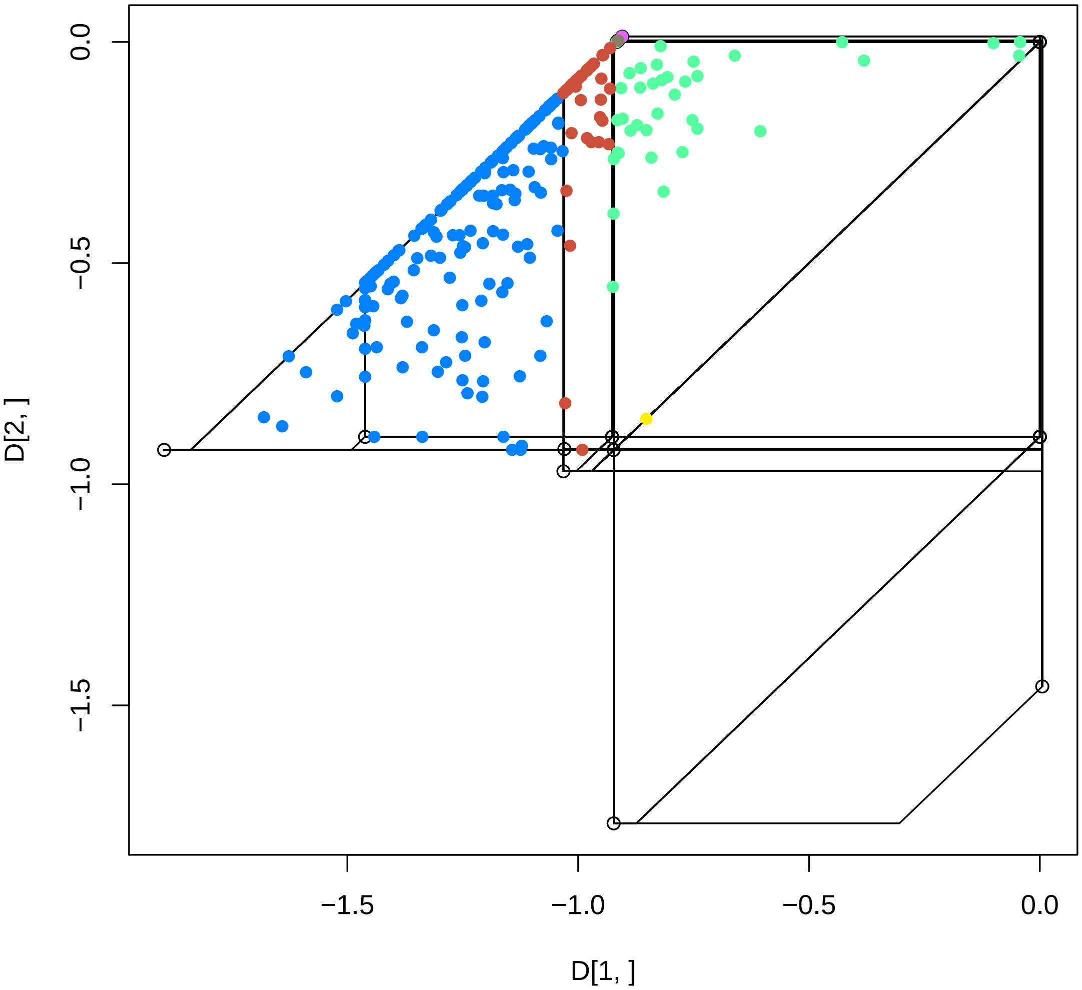


Figure 5.9. Tropical PCA For Apicomplexa

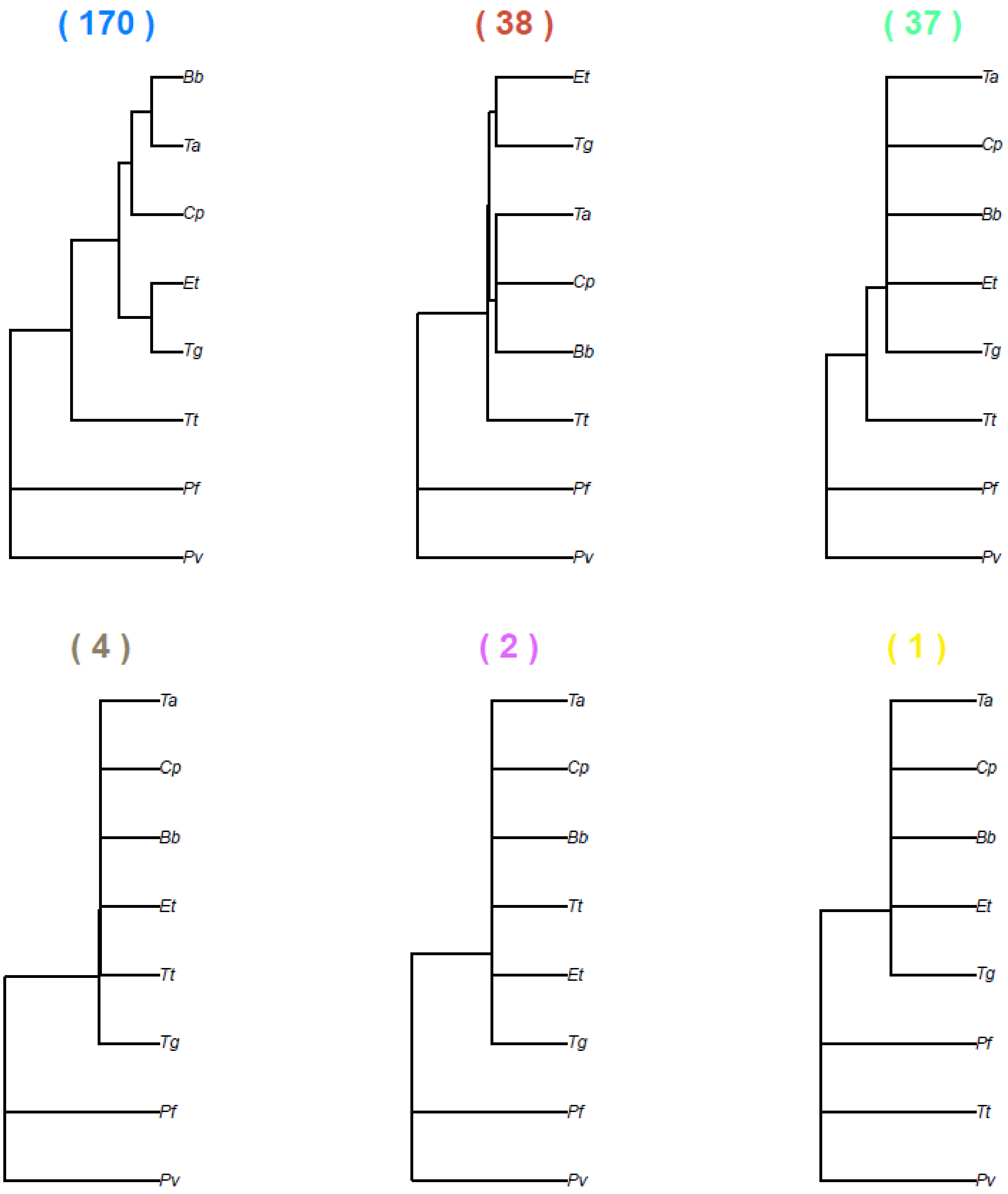


Figure 5.10. Inferred Apicomplexan Species Tree Topologies

CHAPTER 6:
Discussion

This thesis contributes to the study of phylogenomics with new techniques for efficiently clustering gene trees by distribution. We successfully implement our novel stochastic computation using a Metropolis-Hastings algorithm to search phylogenetic tree space. Using a Markov chain Monte Carlo approach and the phylogenetic software program Mesquite, we conduct two simulation experiments. We visualize the results in two-dimensional plots and measure the explainable variance using R^2 (Lin and Yoshida 2018)(Yoshida et al. 2019). We conduct further analysis by comparing our tropical PCA results with Tom Nye's locus of the Fréchet mean PCA technique on BHV tree space (Nye et al. 2017) (Billera et al. 2001). Last, we test the quality of our tropical PCA approach using empirical Apicomplexa and African coelacanth datasets.

Contributions

Through Mesquite, this thesis leverages the opportunity to create endless data to demonstrate the power of tropical PCA. We conduct two simulation experiments and confirm our Metropolis-Hastings algorithms perform well. This allows for an efficient and effective exploration of tropical tree space at a new depth and breadth. Most importantly, we prove our hypothesis that tropical PCA clusters gene trees by distribution.

Our MCMC tropical PCA approach has a relatively fast computation time. Parallel processing allows our approach to perform multiple walks through tropical tree space simultaneously. Depending on computing power, this simultaneity significantly reduces computation time. For context, conducting 10,000 iterations of our tropical PCA technique, under parallel processing, for 2000 gene trees takes approximately 10 minutes. Comparatively, enumerating through all 1,331,334,000 combinations of 2000 gene trees without parallel processing takes years and produces substandard results.

We successfully implement our innovative MCMC tropical PCA approach on two empirical datasets, Apicomplexa and African coelacanth genomes. Our results for both Apicomplexa and African coelacanth are notable. Each plot shows a tight cluster, implying that the use of gene trees to infer species topology is valid. The strong r^2 values suggest our second order tropical PCA approach is effectively reducing gene tree dimensions. The inferential topologies for the species tree are nearly identical, strengthening the proof of our hypothesis.

Future Research

We are currently working on the implementation of our MCMC tropical PCA approach on New York influenza data. Preliminary results show high r^2 values and tight gene tree clustering. Potentially noteworthy results could lead to species topology breakthroughs and prove useful to medical researchers. Understanding the past is important, but ultimately we strive to leverage the past to better inform the future.

Additional research should consider model and visualization improvements. Streamlining the tropical MCMC implementation in R, specifically the coding, will improve efficiency and usability. There is room for improvements in the heating and cooling of the MCMC function, which will increase effectiveness of the tropical PCA algorithm. Further exploration is necessary to understand the best starting set of trees for the tropical triangle (base).

Clearer and more meaningful visualization improves interpretability of tropical PCA results. For example, the tropical line segments in our plots correspond to tree topologies; however, identifying those tree topologies takes significant effort. Additionally, producing three-dimensional plots would greatly improve quality. A better understanding of tropical tree space and the mapping of the space to tree topologies may enable further application of tropical PCA techniques to phylogenetics.

List of References

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Stat. Sci.* 23:No.1, 23–34.

Amemiya C, Alföldi J, Lee A, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, et al (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496(7445):311–316, 10.1038/nature12027.

Billera L, Holmes S, Vogtmann K (2001) Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27:733–7677.

Brown TA (2002) *Genomes 2nd ed* (Wiley-Liss, Oxford).

Cavalli-Sforza L, Edwards A (1967) Phylogenetic analysis. models and estimation procedures. *Am J. Hum Genet.* 19:233–57.

Cuninghame-Green RA (1979) Minimax algebra. Beckmann M, Kunzi HP, eds., *Lecture Notes in Economics and Mathematical Systems*, 166, 1–258 (Springer-Verlag, Berlin).

Darwin C (1837) Notebook B: Transmutation of Species (1837-1838), http://darwin-online.org.uk/.

Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London).

Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press).

Felsenstein J (2004) *Inferring Phylogenetics* (Sinauer Associates Inc., Massachusetts).

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.

Ihaka R, Gentleman R (1993) R, version 3.5. Accessed May 28, 2019, https://www.r-project.org/.

Kuo C, Wares JP, Kissinger JC (2008a) The Apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees. *Mol. Biol. Evol.* 25:2689–98.

Kuo C, Wares JP, Kissinger JC (2008b) The Apicomplexan whole-genome phylogeny:an analysis of incongruence among gene trees. *Mol. Biol. Evol.* 25:2689–2698.

Liang D, Shen XX, Zhang P (2013) One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Mol. Biol. Evol.* 30 8(8):1803–7.

Lin B, Yoshida R (2018) Tropical Fermat-Weber points. *SIAM J. Discrete Math* 32(2):1229–1245.

Liverani C, Wojtkowski MP (1994) Generalization of the Hilbert metric to the space of positive definite matrices. *Pac. J. Math.* 166(2):339–355.

Maclagan D, Sturmfels B (2015) Introduction to tropical geometry. Mazzeo R, ed., *Graduate Studies in Mathematics*, 161, 1–361 (American Mathematical Society).

Maddison D, Maddison W (1997) Mesquite, version 3.6. Accessed January 2019, http://www.mesquiteproject.org/.

Maddison W, Knowles L (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55:21–30.

Mau B, Newton M (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6:122–131.

Nye T (2011) Principle component analysis in the space of phylogenetic trees. *Ann. Stat.* 39(5):2716–2739.

Nye T (2016) Geophytter+. Accessed March 2019, http://www.mas.ncl.ac.uk/ ntmwn/geophytterplus/index.html.

Nye T, Tang X, Weyenberg G, Yoshida R (2017) Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. *Biometrika* 104(4):901–922.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2(11):559–572.

Pin JE (1998) Tropical semirings. Gunawardena J, ed., *Idempotency*, 50—69 (Cambridge University Press, Cambridge).

Ridley M (1999) *Genome: The Autobiography of a Species in 23 Chapters* (Harper-Collins, New York).

Semple C, Steel M (2003) *Phylogenetics* (Oxford University Press).

Sokal R, Michener C (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 28:1409–1438.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures Math. Life Sci.* 17:57–86.

Yoshida R, Zhang L, Zhang X (2019) Tropical principal component analysis and its application to phylogenetics. *Bull. Math. Biol.* 81(2):568–597.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California