INFERENCE ON PHYLOGENETIC TREES BASED ON VECTORIZATION

by

YAN DU

(Under the Direction of Liang Liu)

Abstract

Phylogenetic trees are fundamental tools for studies in Evolutionary Biology, Population Genetics, and Comparative Genomics. However, the algebraic and topological properties of spaces of phylogenetic trees are, by and large, unexplored. The majority of contemporary works are built under the infrastructure of BHV tree space, in which each phylogenetic tree is represented as a point and the branch lengths as its coordinates. Despite the fundamental role of the BHV space in phylogenetic inference, computational complexity of the geodesic metric for the BHV space significantly limits its applications for studying algebraic and topological properties of tree spaces. In this dissertation we propose a novel mathematical framework for phylogenetic inference and demonstrate its applications in statistical inference of phylogenetic trees.

This thesis includes two major parts. First, we develop a topological vector space $\mathcal{V}$ in which the topology of a phylogenetic tree is defined as a linear map of $\mathcal{V}$. We further map phylogenetic trees with branch lengths to spaces of graphical-path vectors. We show that there exists an isomorphism between phylogenetic trees and a polyhedral complex in Euclidean space by this vectorization mapping. In addition, the topological vector space can be metricized by the $L2$ norm.

In the second part, statistical properties of phylogenetic trees are studied in the context of the corresponding metric space. Based on the vectorization of trees, we define a centroid and variability measure for phylogenetic tree and propose an estimation method for the mean tree. The estimator is inferred algebraically and demonstrated by a simulation study as asymptotically normal distributed.

INDEX WORDS: phylogenetic tree, BHV tree space, path distance, statistical inference

INFERENCE ON PHYLOGENETIC TREES BASED ON VECTORIZATION

by

YAN DU

B.S., Nankai University (China), 2011

M.S., Wuhan University (China), 2013

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2020

INFERENCE ON PHYLOGENETIC TREES BASED ON VECTORIZATION

by

YAN DU

Major Professor: Liang Liu

Committee: Jonathan Arnold
Paul Schliekelman
Pengsheng Ji
Shuyang Bai

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

Acknowledgments

I would like to express my deepest gratitude and appreciation to my major advisor Dr. Liang Liu. I have greatly benefited from his patient guidance, insightful comments and suggestions through my research. Without his guidance and persistent help, this dissertation would not have been possible. I would like to thank him for all his tremendous help and support.

I would also like to thank my committee members, Dr. Jonathan Arnold, Dr. Paul Scheliekelman, Dr. Pengsheng Ji and Dr. Shuyang Bai for their valuable suggestions and their time spending in reviewing this dissertation. I would also like to thank the faculties and staffs in Department of Statistics, and my friends from this community, for the knowledge they taught me and the help they gave me.

Lastly but most importantly, I want to thank my parents and sisters for their unconditional and endless love.

Table of Contents

List of Figures

Chapter 1

Introduction

Phylogenetic trees are tree-like mathematical graphs describing the evolutionary relationships of a given collection of organisms. Phylogenetic trees have been fundamental tools for understanding the evolution of research subjects of interest in various domains, such as the spread of pathogen (Hillis and Huelsenbeck, 1994), speciation of mammals (Wu et al., 2018), and even the development of a literary genre (Liu and Yu, 2020). As genetic data become increasingly available, it is of great interest to develop novel phylogenetic models for analyzing enormous amount of phylogenetic data generated by next-generation sequencing techniques.

Contemporary phylogenetic analysis uses molecular sequencing data collected from existing organism to infer the historical evolutionary trace of species. The basic idea is to model nucleotide substitution by a continuous time Markov chain with a homogeneous rate matrix. The uncertainty in some critical variables in reconstruction process gives rise to several types of error in reconstructed phylogenetic trees. For example, it is widely realized that different genetic sites may imply conflicting gene trees (Maddison 1997; Reid, 2014). This makes the phylogenetic tree substantially a random variable. In addition, the choice of substitution models may have misleading effect on resulting trees (Buckley and Cunningham, 2002; Hoff et al., 2016). Moreover, the reconstruction methods exploited may have effect on the inferred trees (Liu et al., 2015; Weyenberg 2015). Such systematic uncertainty as well as the estimation error results in the phenomenon that researchers are often faced with a set of phylogenies. The potential incongruency in a set of phylogenetic trees motivates a systematic approach for analysis on tree sets, especially for comparison of trees. Conventional

approaches such as summarizing a set of trees by a consensus tree discourage the statistical inference on the set of trees.

The majority work on the analysis of a set of trees is constructed upon the Billera-Holmes-Vogtmann (BHV) tree space (Billera et al., 2001). This tree space gives us a geometry view on the non-Euclidean space of all rooted trees built on a given set of taxa. By setting the geodesic distance in the BHV space, a series of classical statistical problems have been studied. For example, Nye (2011) proposed a geometrical approach to PCA in BHV tree space by constructing a geodesic principal path. Barden et al. derived the limiting behaviors of sample Fréchet means in BHV tree space (Barden et al., 2014; Barden et al., 2018). Willis (2017) developed a procedure to construct a confidence set based on the log-map function (Barden et al., 2014) in the BHV space. However, the geodesic metric used in BHV tree space brings the complexity on computation of trees.

Figure 1.1 depicts the BHV space of 4-taxon phylogenetic trees. In BHV tree space, each tree is represented as a point in an orthant. The coordinates of the tree are characterized by the lengths of its internal edges. The geodesic metric (Gromov,1987) exploited in this space is the shortest path connecting trees in this space. For computation in this space, the preliminary work is to determine the geodesic connecting the trees. Though each "orthant" can be viewed as part of the real vector space, the geodesic connecting two trees in distinct orthants steps over the orthant boundaries, which makes the computation difficult. The search for the shortest path between trees is an optimization problem. In a tree space with large trees, the optimization algorithms will be quite complicated, given that the determination of direction of each segment in the geodesic involves an optimization procedure. Several algorithms have been developed to search for the geodesic efficiently (Amenta et al., 2007; Kupczok et al., 2008; Owen and Provan, 2011). In these algorithms, the first step is to advance multiple paths joining the two trees. The construction of such paths is especially complicated for trees residing in non-neighbour orthants (Monod, 2019). Most efficient as it is, Owen and

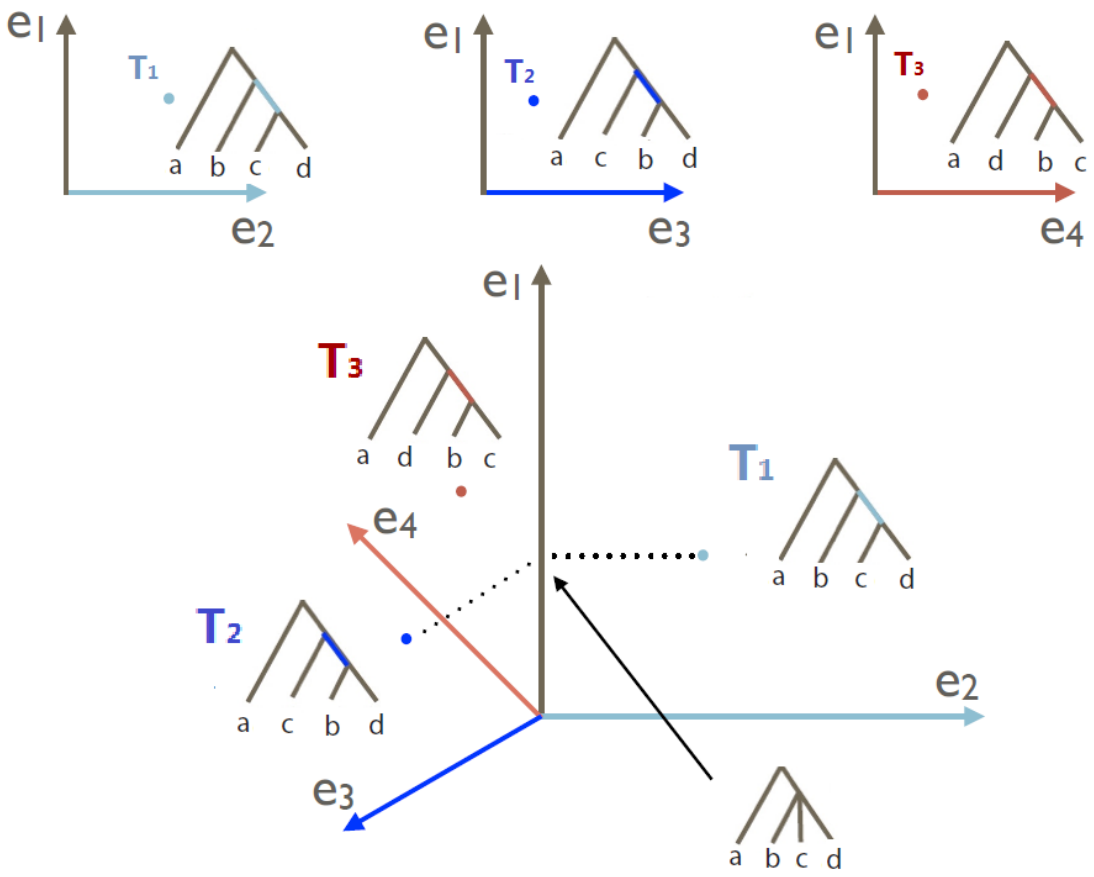

Figure 1.1: A portion of Petersen graph exhibiting how orthants associated with three adjacent 4-taxa trees graft.

Reprinted from *Confidence Sets for Phylogenetic Trees*, by A. Willis, 2017.

Provan's algorithm is shown (Lin et al., 2017) that the dimension of the geodesics in the algorithm is unbounded.

The difficulty in the computation of geodesics in BHV tree space is caused by the complex topological structure of phylogenetic trees. To alleviate this difficulty, we propose an alternative approach by mapping phylogenetic trees to a real vector space. We show that there exists an isomorphism between phylogenetic trees and the topological space of pairwise path vectors. Based on this representation, a tree metric is defined by the $L2$ norm in vector

space. Following this representation, we map the phylogenetic trees to the corresponding vectors. By exploring the structure of the corresponding vectors, we show that the image of φ is also a polyhedral complex, or in a tropical geometry view, a tropical variety (Speyer and Sturmfels, 2004). Under this infrastructure, we construct a classical statistical framework on trees based on the vectors. Specifically, we define the mean tree and tree variance as the measure of centroid and variability in trees. Based on the correspondence between trees and vectors, we propose a sample mean tree whose corresponding vector is closet to the vector corresponding to the mean tree as an estimation of mean tree. In addition, we show that our sample mean tree has satisfying limiting properties, which makes the sample mean tree defined under this vectorization infrastructure as a good estimator. This statistical application supports the significance of vectorization of trees.

Comparing with other studies on this topic, the significance of our work is that we vectorize the tree at the very first step. So, instead of using the geometry metric in BHV tree space to construct the statistical work, we conduct analysis on trees by manipulating the corresponding vectors. The analytical advantage of this approach is quite clear, as the topological ambiguity is addressed quantitatively in real vector space. In addition, this construction of tree space has potentials to develop efficient algorithms for moving phylogenetic trees in the tree space.

The remainder of this thesis is organized as follows. In Chapter 2, we provide a literature overview on phylogenetic trees and BHV tree space to explain our research motivation. In Chapter 3, we construct a mathematical infrastructure for phylogenetic trees. We start with presenting the basic concepts of trees, then introducing the mapping from tree to vector and the induced tree metric, followed by an exploration of structure of the topological vector spaces. Chapter 4 is the statistical application of our constructed settings. We establish the basic statistical infrastructure in tree space based on the corresponding vectors. The main part is that we propose a method to estimate the mean tree. In addition, we show the multivariate central limit theorem of the vector corresponding to sample mean tree. At end,

we close this thesis by Chapter 5, in which a summary and a perspective on future work are provided.

Chapter 2

Literature Review

This part provides a literature review of phylogenetic trees and BHV tree space. Aiming at explaining our research motivation, we will cover basic concepts in phylogenetics and the literature closely related to our work.

Section 2.1 is a brief introduction to the field of phylogenetics, with emphasis on the molecular evolution process. This part explains the estimation error in phylogenetic trees, as well as other factors in the reconstruction process that may affect the phylogenetic inference. Thus, the necessity of handling a set of trees is induced. In addition, we introduce the traditional methods to deal with incongruent phylogenetic trees. In Section 2.2, we provide a literature overview on the BHV tree space, which is the primary topic covering the exploration of all phylogenetic trees built on a given collection of taxa, and point out its drawback. Section 2.3 introduces some previous algebraic representation of trees and discusses the merits of representing trees by matrices/vectors.

2.1 Phylogenetic Trees

In this part, we will start from the basic introduction to phylogenetic trees, then explain the estimation and systematic errors in phylogeny reconstruction, which result in a set of incongruent trees.

2.1.1 Introduction to Phylogenetic Inference

The idea of describing evolution by a branching pattern graph has appeared in Darwin's theory, which was inspired by the phenotypic variations of finches in the Galapagos Islands

(Salemi et al., 2009). The phenotypic variation is observed to be a result of the genetic information carried by the organisms. This stimulates the contemporary phylogenetic analysis (Cavalli-Sforza and Edwards, 1967), which is to investigate the evolutionary history based on the sequencing data collected from the existing organisms.

Phylogenetics is an inter-discipline of biology and statistics. Two essential elements contribute to evolution: inheriting and mutation. As the DNA of a parent is copied to descendants, mutation sometimes occurs. The accumulation of small mutations over generations introduces the genetic variation in the organisms. In the science of speciation, if several species arise from a common ancestor, they are expected to have similar DNA sequences. Reversely, the differences in the gene sequences imply the evolutionary divergence. In a word, genomes contain traces of history. Thus, this field is inspired to use the gene sequence data from several organisms to infer their evolutionary history.

Given a sequence alignment, the variation in the alignment comes from the mutation of the nucleotides over the evolution. The most common and fundamental mutation is base substitution, which is the replacement of one base for another. That is the reason why we can compare the alignment to trace the evolution in genes. An example of such a reproduction process is shown in Figure 2.1. Such reproduction process enables us to infer a historic branching process based on the sequence data collected from contemporary organisms.

To quantify the distance among multiple sequences based on the nucleotide substitution, the straightforward method is subtraction distance, which is the proportion of the dissimilarity. However, this method neglects the possibility of multiple hits, such as the hidden mutation like $T \to A \to G$, or the back mutation like $G \to A \to G$. To model such molecular evolution, the most common approach is to treat the evolution of each site as a continuous-time Markov chain, which has the memoryless property. Taking the process in Figure 2.1 as an example, the distinction between observed mutations and the actual mutations indicates that using frequency to quantify mutation will underestimate the number of mutations. However, if we use Markov chain to model the substitutions at any particular site, with the

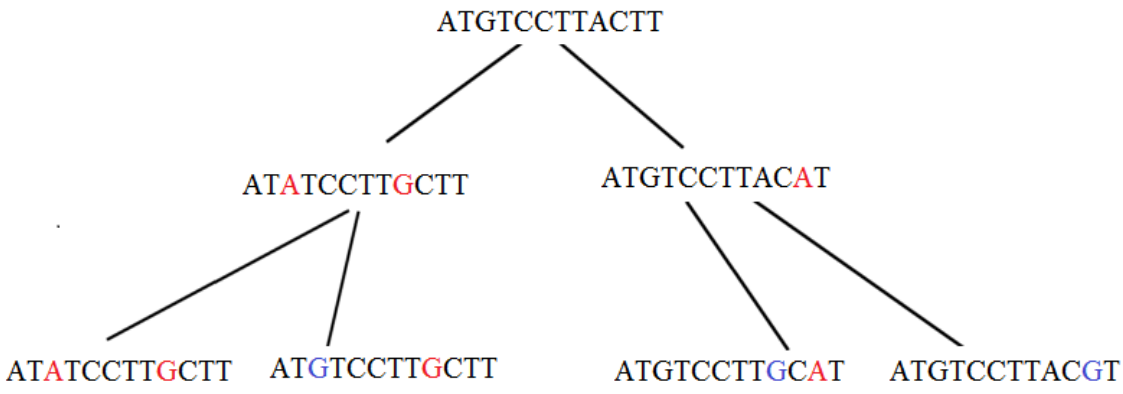

Figure 2.1: An example of reproduction with mutation

four nucleotides being set as the states of the chain, the transition probabilities for a Markov chain satisfy the Chapman-Kolmogorov theorem

$$p_{ij}(t_1 + t_2) = \sum_k (p_{ik}(t_1) + p_{kj}(t_2))$$

. This is how the Markov chain model corrects for multiple hits.

The behavior of a continuous-time Markov chain is characterized by a 4×4 transition rate matrix Q. Its corresponding transition probability matrix over time t is referred to as $P(t)$. Based on the Markov property, $P(t)$ is related to Q by the matrix exponential,

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

Besides the basic assumption of Markov chain, we can place further constraints on the substitution rate between nucleotides. Different constraints lead to different models of Markov substitution (Arenas, 2015). For example, the simplest model is Jukes-Cantor (Jukes and Cantor, 1969), which assumes all $\binom{4}{2}$ types of nucleotide substitutions have the same transition rate. A model that only constraints the time-reversibility between bases is the GTR model (Tavare, 1986), which allows for all types of substitutions to occur at a distinct rate.

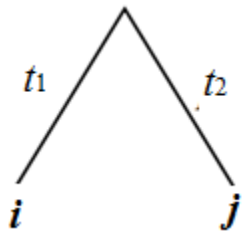

Figure 2.2: The simplest process of molecular evolution modelling by continuous-time Markov process

Example 1. We use the example of Jukes-Cantor model to illustrate how to calculate the transition probability as a function of substitution rate parameter and the evolution time parameter.

The Jukes-Cantor model assumes the transition rate matrix Q as

$$Q = \begin{pmatrix} q_{AA} & q_{AG} & q_{AC} & q_{AT} \\ q_{GA} & q_{GG} & q_{GC} & q_{GT} \\ q_{CA} & q_{CG} & q_{CC} & q_{CT} \\ q_{TA} & q_{TG} & q_{TC} & q_{TT} \end{pmatrix} = \begin{pmatrix} -\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & -\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & -\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & -\alpha \end{pmatrix} = S\Lambda S^{-1}$$

The transition matrix corresponding to this Q is

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AG}(t) & p_{AC}(t) & p_{AT}(t) \\ p_{GA}(t) & p_{GG}(t) & p_{GC}(t) & p_{GT}(t) \\ p_{CA}(t) & p_{CG}(t) & p_{CC}(t) & p_{CT}(t) \\ p_{TA}(t) & p_{TG}(t) & p_{TC}(t) & p_{TT}(t) \end{pmatrix} = \exp(Qt) = Se^{\Lambda t}S^{-1}$$

This gives the transition probability as

$$P_{i,j}(t_1 + t_2) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\alpha(t_1+t_2)}\right), i, j \in \{A, G, C, T\}$$

in the simplest process as in Figure 2.2.

Given a sequence alignment as material and a substitution model as Markov parameter, the genetic distance, or say $p-$distance, which is measured by the expected number of substitutions (like $\hat{t}_1 + \hat{t}_2$ between i and j in tree shown by Figure 2.2), can be estimated from the mutation in the alignment ($\hat{P_{i,j}}$). A more complicated approach is that, given the alignment as an observed sample and a Markov substitution to model the stochastic process, a likelihood function can be built as a function of t and substitution model parameter α. These two approaches are the basis for the tree reconstruction methods as distance-based methods and probabilistic methods, respectively.

Given the alignment and substitution model, we can infer the evolutionary relationship in a tree diagram as shown in Figure 2.3. The tree reconstruction methodology has been studied by numerous researches. Based on the materials and criteria, the methods can be grouped into the following categories.

Maximum parsimony methods (e.g., Fitch, 1971; Nixon, 1999; Goloboff, 1999) take alignment as input and aim to find the tree topology such that the given sequences can be explained with the smallest number of changes.

Distance-based takes the distance matrix as input. The distance commonly refers to genetic distance. One natural way to construct a tree from the distance matrix is the clustering algorithm. The most commonly used methods are UPGMA (Sneath and Sokal, 1973) and NJ (Bruno et al., 2000). The alternative way is to set a tree score based on the distance matrix, and then search for the tree who optimizes the tree score. The most commonly implemented score is the least square. Many algorithms have been developed to minimize the least square score (e.g., Fitch and Margoliash 1967; Kuhner and Felsenstein 1994, Gascuel 2000).

Besides the maximum parsimony and distance-based methods, the reconstruction based on probabilistic models of sequence evolution has been in great popularity since the likelihood approach is used to model the evolution on the alignment (Felsenstein, 1973a). The basic idea of the maximum likelihood method is to maximize the likelihood tree score. A frequently

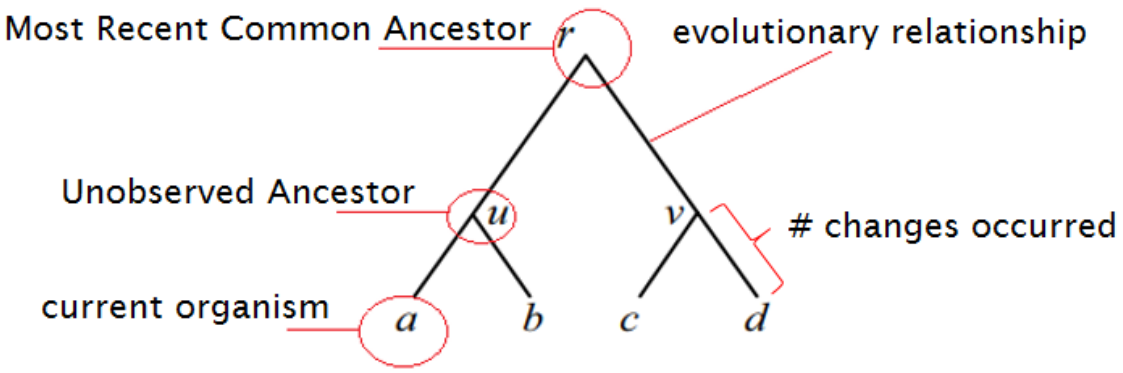

Figure 2.3: An example of estimated phylogeny with 4 taxa

used tool to build the maximum likelihood tree from the alignment is RAxML (Stamatakis, 2014). A similar alternative approach to the maximum likelihood method is the Bayesian methods (Rannala and Yang, 1996), which is also based on the probabilistic model and aims to search for the tree with the maximum posterior probability. A popular Bayesian phylogenetic analysis tool is MrBayes (Huelsenbeck and Ronquist, 2011).

After introducing how the phylogenetic trees are reconstructed from the alignment according to the substitution process, we will explain several types of errors that may appear in the phylogeny reconstruction.

2.1.2 Uncertainty in Phylogenetic Inference

Like other inferential problems, the phylogenetic inference can be impacted by estimation errors as well as systematic errors. Estimation errors arise from the random sampling of genes and taxa, and thus can be reduced by increasing the sample size. Systematic error, however, comes from the deficiency such as data artifact or model misspecification. With systematic errors, increasing the sample size of the genetic data will not help reduce the bias in estimated phylogenetic trees.

In general, the phylogenetic inferences based on different genetic sites may result in different estimated trees (Maddison 1997; Rosenberg and Nordborg, 2002; Galtier and Daubin, 2008; Reid, 2014). Because the locus evolves independently to some extent and may imply different gene trees from the underlying species tree, therefore, no matter which reconstruction approach is used, the phylogenetic trees may differ with respect to topology for genetic data collected from multiple locus. This is a profound source of estimation error in phylogeny reconstruction. In this sense, the phylogenetic tree from the underlying species tree can be substantially viewed as a random variable.

Besides the estimation errors from the sampling genes, many other factors in the reconstruction process may impact the phylogenetic inference.

The uncertainty in alignment is a critical variable when building the phylogenetic tree. Alignment is the fundamental material for the phylogenomic studies. In most organisms, the genetic information is carried by DNA or RNA, which are polynucleotides with four bases. In the gene duplication process, errors may cause mutations like deletion or insertion of the nucleotides. To make the homologous sites which are assumed to diverge from a common ancestral state comparable, a preliminary work of alignment should be done to achieve the positional homology of the genetic sites. The alignment procedure arranges the homologous sites in columns by implying "gaps" in some positions. However, the generation of optimal alignment can be difficult and controversial(Salemi et al., 2009). Because of the large size of genetic data, the alignment is often generated by an automatic program. For example, a frequently used tool to align the sequence is MAFFT (Standley, 2013). Numerous alignment programs have been developed. For example, the DCA (Stoye et al., 1997; Stoye, 1998 program is a heuristic algorithm to sum-of-pairs optimal alignment. CONTRAALIGN Do et al., 2006) implements discriminative learning techniques to align the sequences. MAFFT (Standley, 2013), which is a frequently used program to perform the alignment, offers various alignment methods, such as iterative refinement and consistency-based scoring approaches. Nevertheless, there is no uniform standard to measure the per-

formance of the aligned sequence. The choice of the alignment procedure protocol may be controversial. With this altercation, the effect of alignment is of interest. The effect of uncertainty in alignment on the resulting phylogeny has been studied in several reconstructions (e.g., Morrison and Ellis, 1997. Mugridge et al., 2000; Wu et al., 2012). Whether the application of different alignment will alter the phylogeny has no explicit answer yet. For example, there are studies indicating that changes in gene sequences may result in contentious phylogenies (Shen et al., 2017). Sometimes the choice of alignment does not impact the estimated species tree (Du et al., 2019). Furthermore, some studies (Talavera and Castresana, 2007; Wong et al., 2008) show that alignment uncertainty can be influential, while its misleading effects depend on the shape of the true phylogeny. Overall, the uncertainty in alignment may affect the reconstruction of phylogenetic trees.

The choice of substitution models is another critical variable in tree reconstruction. Several studies (Posada and Crandall, 2001; Minin et al., 2003; Lemmon and Moriarty, 2004; Hoff et al., 2016) demonstrate that the improper substitution model may mislead the phylogenetic tree inference, and the under-parameterization of the model can be more influential than the over-parameterization (Lemmon and Moriarty, 2004). From the data-analytical view, there are studies (Abadi et al., 2019; Du et al., 2019) suggest that the choice of substitution model may not impact the inference. The gene tree estimation may be robust to the choice of substitution model. In another aspect, applying different current model selection strategies to choose the most suitable substitution model leads to similar phylogenetic inference. However, it is noted that different choice of substitution models frequently results in incongruent phylogenies (Abadi et al., 2019).

Given the fixed materials and methods such as alignment, substitution model, and reconstruction approach, the phylogenetic tree can be viewed as randomly distributed in a probability space. The statistical inference on a "random" phylogenetic tree stimulates a method to represent the phylogenetic tree quantitatively.

In another aspect, the evaluation of phylogenetic trees resulting from different materials or reconstruction procedures needs further exploration. There are preliminary work to select the optimal input variables, such as optimal alignment (Thompson et al., 1999; Raghava et al., 2003) and substitution model (Minin et al., 2003; Abdo et al., 2005) according to some tree scorings and tests developed upon the scorings. However, the overall quality and accuracy of optimal alignment and substitution models cannot be guaranteed (Salemi, M. et al., 2009). Thus, the "optimal" tree is often controversial. The analysis on a set of estimated trees can help with this problem.

2.1.3 Dealing with Incongruent Trees

We have discussed why incongruent phylogenetic trees appear. In this part, we will introduce the traditional tools to deal with a set of incongruent trees. We will cover the traditional tree distances, including tree rearrangement distances and inner product distances, and the consensus method.

Within systematics, a fundamental problem is how to deal with the incongruent phylogenetic trees. If all the estimated phylogenies share the same branching pattern, then the calculation on the trees can be conducted as in Euclidean space. However, if the trees have conflicting topologies, which refers to its branching pattern without the information on branch lengths, things become complicated. In real problems, because of reasons mentioned relevant to either estimation or systematic errors, it is often necessary to compare trees with different topologies.

Consider a rooted tree built on m organisms, it has $(2m-3)!!$ (Schroder, 1870) possible topologies. There are two essential problems to investigate trees with more than one possible topology. The first one is to measure how different two trees are. If two trees are topologically identical, then it is natural to use the sum of the branch lengths differences. To measure the topological distance, several tools have been proposed. One idea is to measure the difference by graphically counting the tree rearrangement operations between two trees. Popular tools

following this idea include Nearest-Neighbor-Interchange (NNI) distance (Robinson, 1971), Subtree-Prune-Regraft (SPR) distance (Penny and Hendy, 1985), and a generalization of SPR as Tree-Bisection-Regrafting (TRB) distance (Semple and Steel, 2003). However, the computation of the above distances are NP-hard problems (Dasgupta et al., 1997; Hickey et al., 2008; Allen and Steel, 2001).

Besides such tree rearrangement distances, an alternative is to derive the tree distance from the vector magnitudes in Euclidean space. The general idea is to map phylogenetic trees into Euclidean space by a vectorization function, and then the tree distance can be defined based on the squared Euclidean distances. Such distance measures are called inner product distances. The most frequently used measure is Robinson-Foulds distance (Robinson and Foulds, 1981), which counts the bipartitions that are in one tree but not in the other. Besides the Robinson-Foulds distance, triple distance (Critchlow et al., 1996) and quartet distance (Estabrook et al., 1985) are favored to quantify the topological difference. Compared with the tree rearrangement distances, such tools are computationally efficient because several dynamic programming algorithms have been developed (Day, 1985; Steel and Penny, 1993; Critchlow et al., 1996). However, the inference on trees, based on the above metric, needs further exploration.

The second problem in systematics is how to summarize a collection of trees. Similarly, if the trees share the same topology, then the branch length average is a natural summary statistic. If not, the conventional approach is to construct a consensus tree (Adams 1972, Bryant 2003) from a set of trees containing conflicting topologies. The simplest way is to extract a strict consensus tree (McMorris, 1983), which shows the branching pattern that is shared by all trees. Another way is to use majority-rule to construct a consensus tree that shows the support values on each branch, referred to as the majority-rule consensus tree (Margush and McMorris, 1981). However, if the maximum proportion is less than half, this construction can still result in a tree with polytomy, which is undesirable since it loses the information on certain nodes. In addition, several criteria and methods have been developed

to retrieve a consensus tree from a set of trees (e.g., Ragan, 1992; Baum, 1992). Compared with the consensus method that summarizes all trees by a single consensus tree, the analysis on the collection of trees can keep more information and allow a statistical perspective on the given trees (Willis, 2017).

In summary, compared with traditional methods on multiple phylogenies, quantitative and inferential analysis on tree sets can give us a better understanding of incongruent trees.

2.2 BHV Tree Space

Last, we discussed some traditional methods on multiple trees to induce our motivation to set up an infrastructure for analyzing a set of trees. This part gives a brief literature overview on BHV tree space, which is the groundwork of the analysis on tree sets. We will introduce the BHV tree space and the previous statistical work based on BHV tree space, and other statistical studies derived from BHV tree space.

Billera et al. (2011) use a geometric model to describe the set of rooted trees built on a given set of m taxa with positive internal branch lengths in a Hadamard space. The rooted tree with m leaves has at most $m-2$ internal branches. Each distinct binary tree topology is associated with a "top-dimensional" Euclidean orthant. Each tree can be viewed as a point in the orthant associated with its topology. The orthant coordinates represent the internal branch lengths in the tree. Thus, if there is polytomy, which can be explained as the corresponding internal branch decreasing to 0, then it resides on the orthant boundary, which is an orthant with lower dimension. For example, a binary tree topology is associated with an orthant with the top dimension $m-2$. By collapsing an internal branch, it moves to the orthant boundary (with dimension $m-3$) and becomes a tree with a polytomy. As Figure 2.4 shows, the boundary trees from two different orthants may describe the same multifurcating topology. This coincidence inspires the construction of BHV tree space by grafting the common orthant boundaries together, as in Figure 1.1.Thus, the entire BHV

treespace is the Cartesian product of the $(2m-3)!!$ Euclidean orthant meanwhile the $(2m-3)!!$ topologies constitute a non-Euclidean space (Billera et al., 2011).

BHV tree space has inspired a number of statistical inference work on phylogenetic trees. The difficulty of the statistical description on trees is that the trees are graph objects. Therefore, their center and variability are hard to describe. However, under the infrastructure of BHV tree space with the geometry metric, a number of statistical measures are allowed to be developed, especially the mean and variance of phylogenetic trees (e.g., Brown and Owen, 2017; Willis 2017).

Most previous work on statistical analysis in BHV tree space is based on the geodesic distance, which is the length of the unique shortest path between two trees. Though Owen and Provan develop an efficient algorithm (Owen and Provan, 2011) to calculate the geodesic, it is shown that the geodesics are unbounded in dimension (Lin et al., 2017).

Under the infrastructure of BHV tree space, two significant subjects have drawn the attention of researchers. One, as aforementioned, is to establish statistical analysis in the BHV tree space with geodesics (e.g., Barden et al., 2014; Barden et al., 2018; Willis, 2017). The statistic measures and theory are built analytically. However, the calculation based on the geodesics for a tree space with a large number of taxa is expected to be faced with computational difficulty. The alternative approach is to derive a setting, for tree space and metrics, based on BHV space. For example, Monod's study in 2019 utilizes the tropical geometry of BHV space to develop a palm tree space with the tropical metric based on the tropical line segments. From the BHV geodesic to the tropical metric, it has significant improvement in tree metric's computation. However, the following statistical inference in this tropical geometry tree space is limited. For example, Monod defines the population mean as the tree who minimizes the sum of squared tropical distances. But the reasonability of this estimation is tricky to demonstrate. Substantially, it did not propose a way to quantify the tree such that both the centroid and direction of the tree difference can be simultaneously described.

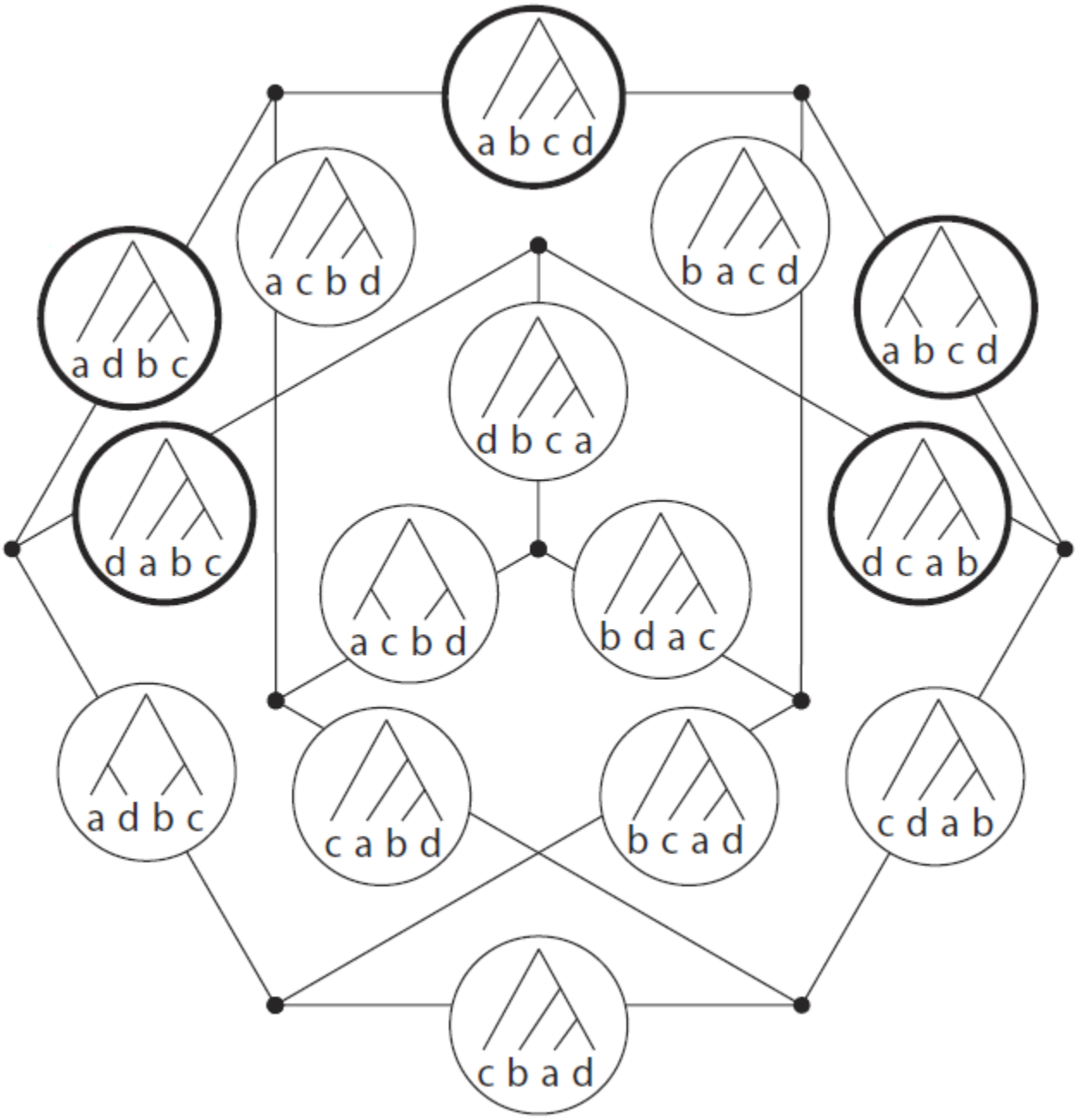

Figure 2.4: The Petersen graph depicting the 15 topologies of 4-taxa

Reprinted from *Statistics in the Billera-Holmes-Vogtmann Treespace*, by G.S. Weyenberg, 2015.

Our work will also exploit the geometry property of BHV tree space. However, instead of manipulating trees, we want to vectorize them into the Euclidean space by an injective mapping so that we can get a corresponding topological vector space isomorphic to the tree space. And then, we can construct an infrastructure for the tree analysis through the corresponding topological vector space.

2.3 Representation of Phylogenetic Trees

We have pointed out the necessity of analyzing tree sets, and have introduced the fundamental work of BHV tree space. Note that the computational complexity in analyzing a set of trees comes from the fact that a tree is a graph, the intuitive solution is to represent the tree by a mathematical object in Euclidean space. In this part, we will discuss the algebraic representation of phylogenetic trees. Besides the literature on the representation, we will discuss the other virtues brought by representing trees with matrices/ vectors.

2.3.1 Representation of Trees

In this part, some representation approaches previously used in the trees will be discussed. Two major approaches are briefly presented, matrix describing the descendance relationship and vector/matrix describing the pairwise relationship.

There is a long history of the study of graph-theoretic relationships. The most intuitive ones are the incidence matrix, adjacency matrix, and Laplace matrix (Kolaczyk and Csárdi, 2014). In each of the above matrices, the (i, j)th entry indicates the neighboring relationship between the pair of nodes (i, j). This relationship inspires researchers to use a matrix to describe the neighboring pattern in the trees. Such mathematical representations of phylogeny have been used to describe or compare the shape of trees (eg. Farris, 1973). However, the classical representation matrix from graph theory is not sufficient enough, because the phylogenetic tree emphasizes more on the descendance relationships between parents and children rather than the relationship between any two vertices in the graph.

Several approaches have been developed to quantitatively describe the topology of phylogenetic trees. One way is the matrix representation based on the additive binary coding of each node (Sokal and Sneath, 1963; Farris et al., 1970), which is substantially determined by the descendance relationship. The typical method is the matrix representation of the parsimony (MRP) method (Ragan, 1992). This method proposes a $n \times s$ ($s \leq n-1$) matrix with values 0 or 1 to characterize a tree with n tips (including the root) and s internal nodes. Each row in this matrix indicates the descendance between the tip and internal nodes. By combining two matrices from two trees using additive binary coding, a hybrid supertree can be reconstructed based on the parsimony analysis of the composite matrix.

Besides the matrix characterizing the descendant relationship between tips and internal nodes, another approach is based on the pairs of the tips. Typical method in this category is the cophenetic vector (Cardona et al., 2013), which proposes a $\binom{n}{2}$ vector for a tree built on n tips. Each entry for pair (i, j) is the cophenetic value, which is depth of the most recent common ancestor (MRCA) of (i, j). A great application of the cophenetic values is the triple distance advanced by Critchlow et al. (Critchlow et al., 1996). The triple distance counts the number of common triple subtrees in two trees. The calculation of the triple distance can be realized by the generational matrix, whose (i, j)th entry is actually the cophenetic value for (i, j).

Note that a tree topology can entirely be described by a matrix, or other mathematical objects similar to matrix. naturally such representation has been implemented to address the following two problems in phylogenetics. One is the construction of a metric between trees. The other topic is the construction of a supertree from a set of trees. Besides these applications, it opens up lots of possibilities in the analysis of phylogenetics. Next, we will discuss a significant merit of vectorization of the tree.

2.3.2 Moves in Tree space

Tree reconstruction is the most crucial topic in phylogenetics. As aforementioned, most commonly used reconstruction tools involve the optimization procedure over the tree space. Besides the need to manipulate various phylogenies, another challenge to our interest is that, the searching over the tree space across different tree topologies involves graphical changes in tree.

For any reconstruction method aiming to maximize a tree score, there are two levels of optimization. A tree score, based on the Markov substitution process, is a function of substitution rate and branch length parameters. The first step of optimization is to maximize the tree score under a given tree topology. The second step is to search over the maximal tree scores for an optimal topology.

Following such reconstruction approaches, the difficulty of finding the optimal tree is the complicated computation. Researchers have endeavored to simplify the calculation in both the steps. An example in simplifying the tree score calculation is the pruning algorithm (Felsenstein, 1973b) on likelihood score over a single tree, which saves much time by identifying common factors and calculating them only once. For the second step, however, the optimization over all the possible tree topologies is an NP-complete problem (Foulds and Graham, 1982). The exhaustive search, which is guaranteed to find the best tree out of all the possible topologies, is computationally infeasible for a tree with a large number of taxa. Thus, much effort has been taken to develop efficient algorithms for heuristic searching for the locally optimal tree, such as the maximum likelihood tree search (Guindon and Gascuel, 2003; Vinh and von Haeseler, 2004), and maximum posterior probability tree search (Drummond and Rambaut, 2000; Huelsenbeck and Ronquist, 2001).

Note that all the above searching procedures have to move in the tree space. The tree score is a function of topology, and the change of topology has to be realized by manipulating its branches graphically. This impediment in computation motivates us to find a way to enable the tree change in an algebraic way. If we can represent the phylogenetic tree by matrix/

vector following some specific rules, then instead of searching over tree space by pruning or grafting branches, we can manipulate their corresponding matrices/ vectors, and thus can reform the computation on trees to the computation in Euclidean space. The drawback of the aforementioned representation methods is that they only deal with topology. Thus, both the computational and inferential work is limited since trees with the same topology are viewed as identical. Therefore, an alternative representation method, that captures both the topology and branch lengths information, is desired.

Chapter 3

A topological vector space for phylogenetic trees

The obstacles in tree computation and analysis come from two aspects. One is the high dimensionality of the tree topologies. The other aspect is that both quantitative and statistical analyses are elusive for graphical objects. In this section, we will map the phylogenetic trees to vectors to manipulate phylogenetic trees in Euclidean space, and thus construct a topological vector space for phylogenetic trees.

3.1 Combinatorics of trees

In this section, we will introduce the combinatorics of trees from its graphic essence. Besides, the quartet tree is explicitly demonstrated as a prerequisite for the remaining sections.

3.1.1 Basic Concepts in Tree

Aiming at setting up an infrastructure on the trees, we will start by introducing the combinatorics of trees in this part.

The graph is an abstract idea describing the relationship between organisms by drawing the diagram. To be precise in the following analysis, we will first introduce the terminologies and notations in trees.

DEFINITION 2. A **graph** refers to $G = (V, E)$, where V is the set of **vertices/ nodes**, and E is the collection of **edges/ branches**.

Each edge $e \in E$ is a two-element set $e = (v_1, v_2)$ of vertices $v_1, v_2 \in V$. A weighted graph is a pair (G, b), where b is a **weight** function $b : E \to \mathbb{R}_{>0}$ such that each $e \in E$ is assigned

with a non-negative real number $b(e)$. If the edges in the graph have no specified weight, or say that edges have constant weight of 1, then the graph is called unweighted.

The **degree** of a vertex is the number of edges to which it is incident.

The **topology** of a graph G with a given vertex set V refers to its edge pattern representation of whether an edge between any pair of vertices exists.

We say v_1 and v_2 are the ends of e, as well as v_1 and v_2 are incident to e. e joins v_1 and v_2. v_1 and v_2 are adjacent nodes. If a vertex is the end of only one edge, it is called a tip/ leaf/ terminal node/ taxon. Otherwise, it is called an internal node.

In practice, for both the computational and analytical purposes, we do not want to present a graph each time when we want to exploit the graph tools. The topological information contained by a graph should be summarized by a mathematical object that can easily be stored and executed. As aforementioned, a common approach characterizing a topology of a graph is the adjacency matrix.

DEFINITION 3. For a graph $G = (V, E)$, its **adjacency matrix** is a $|V| \times |V|$ matrix A that is

$$A(i,j) = \begin{cases} 1, & \text{if } i,j \in E \\ 0, & \text{otherwise} \end{cases}$$

The adjacency matrix contains the information on the edge pattern of a graph, as well as the information of vertices set. However, the adjacency matrix is a sparse matrix. Thus, the usage of the adjacency matrix is often accompanied by a sparse tool in graph theory (Kolaczyk and Csárdi, 2014). This motivates us to find another representation of the topology of phylogenetic trees.

Biologically, each leaf on a phylogenetic tree represents an organism, such as a population, species, genera, families, orders, phyla, etc., that are present so that we can collect sequencing data. An internal vertex/ node in a phylogeny represents a common ancestor for several taxa. We typically do not have sequence data measured for the internal node. The edges in phylogenetic trees indicate lines of descent. The two ends of an edge are the parent/ ancestor

and child/ descendant. The edge weights, or called branch lengths, describe how closely the adjacent organisms are to each other. The edge length in the phylogeny is usually measured by the expected number of substitutions. The topology of a phylogeny represents all the evolutionary relationship, without the information on divergence time or genetic distance. The topology is characterized by the graph shape and leaf labels, and can be represented by adjacency matrix.

If we use graphs to depict the evolutionary relationship between a collection of organisms, note that not all types of graphs can reasonably describe evolution. Next, we will introduce the combinatorics of phylogenetic trees.

DEFINITION 4. A **path** from vertex v_0 to v_n is a sequence of distinct vertices $v_0, v_1, \ldots, v_n$ such that each v_i is adjacent to v_{i+1}. If there is a path between any two distinct vertices, then a graph is said to be **connected**. A **cycle** is a sequence of vertices $v_0, v_1, \ldots, v_n$ which are adjacent to each other while distinct from each other except for $v_0 = v_n$ with $n \geq 3$.

In a graph modeling the evolution, any vertex being an ancestor of itself should be ruled out. In general, we do not want any "loop" in the phylogeny.

DEFINITION 5. A **tree** $T = (V, E)$ is a connected acyclic graph.

There may be multiple paths in a graph, and the shortest path will be selected to measure the distance between vertices. However, multiple distinct paths form a cycle. Thus, we have the following statement.

LEMMA 6. *For any two vertices v_1 and v_2 in a tree T, there is* ***a unique path*** *between them.*

This property of tree validates the following definition of distance in a tree.

DEFINITION 7. The length of the unique path between two vertices is called (weighted) **graphical distance** between two vertices.

$$d(v_1, v_2) = \sum_{e \text{ on the path from } v_1 \text{ to } v_2} b(e)$$

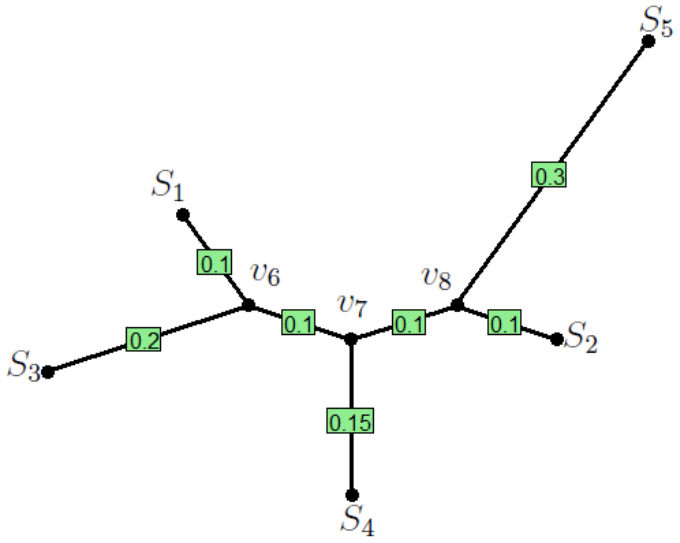

Figure 3.1: A 5-taxa unrooted binary tree with branch lengths

Note that each tip in the graph is incident to a terminal edge with positive length. Thus, the graphical distance between any pair of distinct vertices should be positive.

Example 8. Figure 3.1 presents a tree. It is connected while with no cycles. It has 8 nodes, including 5 tips $(S_1, S_2, S_3, S_4, S_5)$ and 3 internal nodes (v_6, v_7, v_8). The degrees of S_i $(i = 1, \ldots, 5)$ are 1 and degrees of $v_6/\ v_7/\ v_8$ are 3. There is a unique path between any pair of vertices. For example, $d(S_1, S_2) = b(e(S_1, v_6)) + b(e(v_6, v_7)) + b(e(v_7, v_8)) + b(e(v_8, S_2)) = 0.4$. $d(S_3, S_4) = 0.45$. The graphical distance d measures the distance between the vertices.

LEMMA 9. *The function $d : V \times V \to \mathbb{R}_{>0}$ has the properties of non-negativity, symmetry and triangle inequality. Thus, the graphical distance d is a* ***metric on $V(T)$****.*

Besides the basic concepts of vertices, edges, degrees, and paths, there are concepts such as root and polytomy that can group the phylogenetic trees into different presentation types.

In the biological sense, the ancestor of all taxa is the root. A rooted phylogenetic tree means that time is represented by a single direction. If a tree has no specified root, then

it is an unrooted tree. Under the assumption of the molecular clock that the evolutionary rate is constant over time, the root can be identified (Yang and Rannala, 2012). According to the pulley principle (Felsenstein, 1981), the position of the root does not impact the reconstruction. Also, an unrooted tree can be easily rooted by out-grouping (Iwabe et al., 1989). Thus, for the benefit of computation, our research will focus on unrooted trees. As the rooted trees and unrooted trees can switch to each other by pulley principle, all the inferences on unrooted trees can be applied parallelly to rooted trees.

For an unrooted tree, if each internal node has degree of 3, then this tree is said to be binary/ resolved. If it has vertices with degree more than 3, then it is said to have polytomy, or say, it is unresolved. A binary phylogenetic tree indicates that in the evolution process, all speciation events produce two taxa from one. Usually, phylogenetic trees are assumed to be binary because the possibility that several new species arise simultaneously is so small that it can be ignored. However, to cover up all possible phylogenetic trees, we want to take the tree with polytomies into consideration. Thus, our research will cover not only binary trees but also multifurcating trees.

Example 10. Figure 3.2 shows a rooted binary tree and an unrooted binary tree with 5 tips. For simplification, only the tips will be labeled. Such trees are referred to as semi-labeled trees. The rooted tree on the left-hand side has 3 internal edges while the unrooted tree on the right-hand side has 2 internal edges.

In the graphical presentation of trees, sometimes two trees will have different "shapes" while their combinatorial structures, or say, their adjacency matrices are the same. Then, they are identified as the same tree. In contrast, sometimes, two trees seem to have the same branching pattern, while their terminal nodes are labeled differently. Then, they are considered as different trees. For example, in Figure 3.3, T_1 and T_2 are the same tree, while T_1 and T_3 are different trees.

Consider an unrooted tree T with m ordered tips $S = (S_1, S_2, \ldots, S_m)$, terminal edges $(e_1, e_2, \ldots, e_m) \in \mathbb{R}^m_{>0}$ and internal edges $(e_{m+1}, e_{m+2}, \ldots, e_{m+r}) \in \mathbb{R}^r_{>0}$ where $r \leq m-3$. If

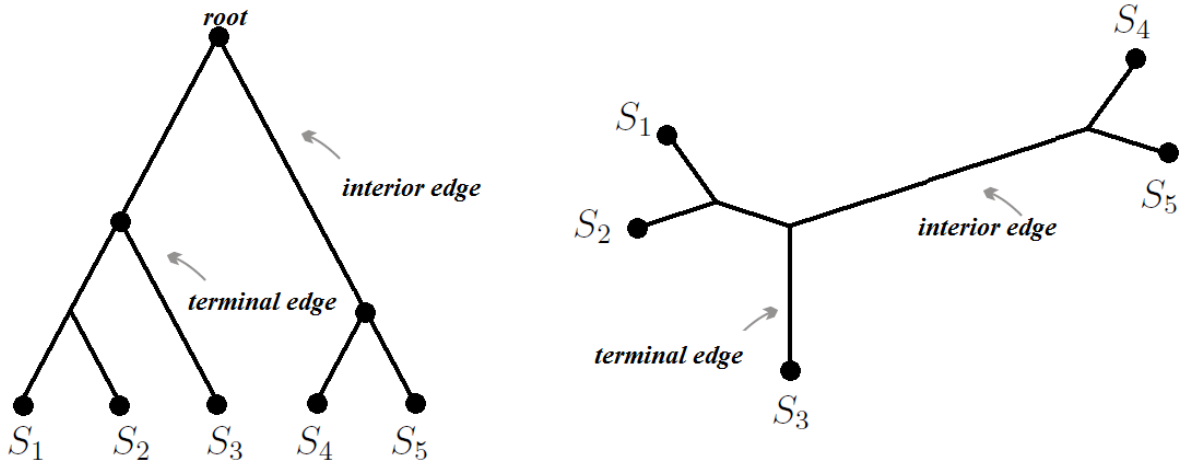

Figure 3.2: A rooted and an unrooted binary tree with 5 tips labeled

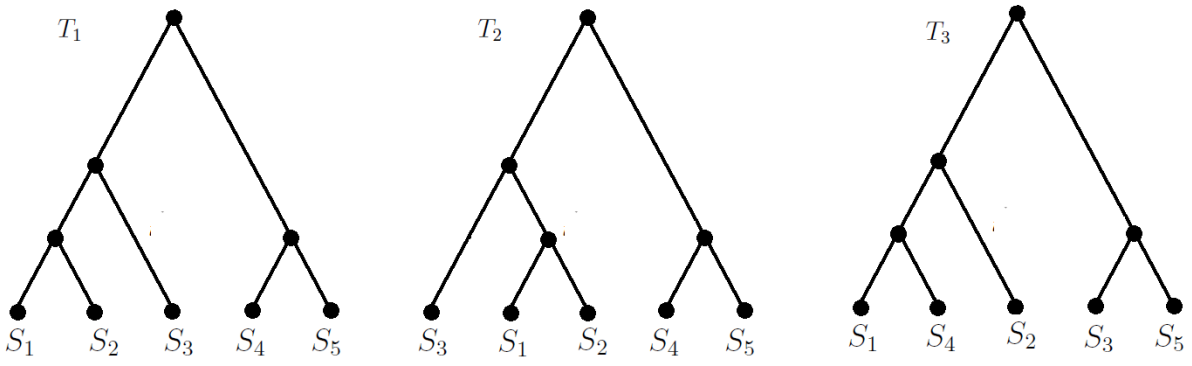

Figure 3.3: Three picture of rooted tree with 5 labeled tips

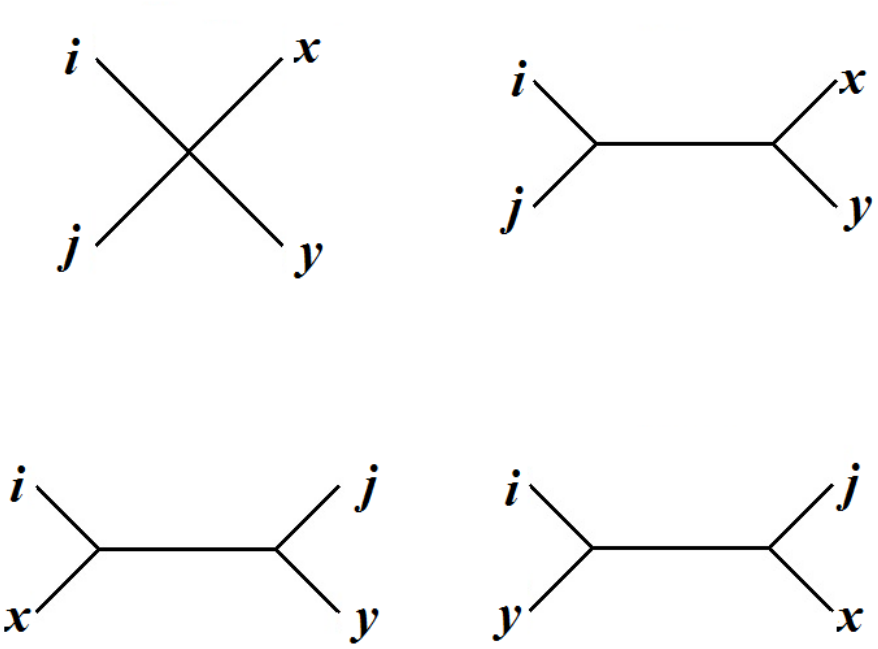

Figure 3.4: Quartet topologies

T is binary, then $r = m - 3$. There are $(2m-5)!!$ (Schroder, 1870) distinct binary topologies to be its potential topology.

3.1.2 Quartet Trees

The smallest informative rooted tree is a tree with three tips. Relabelling a 2-tips tree makes no difference in the adjacency matrix, but relabelling a 3-tips tree identifies distinct topology. Similarly, the smallest informative unrooted tree is a tree with four tips (Estabrook et al., 1985). For any four taxa $\{i, j, x, y\}$, there are 4 possible topologies, as shown in Figure 3.4. One is the star tree in which all four tips are adjacent to a common internal node, resulting in no internal edge. The others are binary trees. The four topologies correspond to distinct adjacency matrices.

We can check that for a quartet tree with any possible topologies, the graphical distance d satisfies the four-point condition (Buneman, 1974) , which is

DEFINITION 11 (Four-point Condition)**.** For any four points (i, j, x, y) in a tree, if the metric d satisfies the inequality:

$$d(i,j) + d(x,y) \leq \max \begin{cases} d(i,x) + d(j,y) \\ d(i,y) + d(j,x) \end{cases}$$

then it is said the metric d on the set $\{i, j, x, y\}$ satisfies the **four-point condition**.

The four-point condition, which involves the operations of addition (+) and max ($\max(a, b)$), encourages the introduction of max-plus semiring (Pin, 1998), which is isomorphic to the min-plus semiring (Speyer and Sturmfels, 2004). To describe the quartet topologies intuitively, we utilize the equivalent min-plus expression.

PROPOSITION 12. *A metric d satisfies the four-point condition if and only if the one of the following inequalities is valid:*

$$(1) \quad d(i,j) + d(x,y) \leq d(i,x) + d(j,y) = d(i,y) + d(j,x)$$
$$(2) \quad d(i,x) + d(j,y) \leq d(i,j) + d(x,y) = d(i,y) + d(j,x)$$
$$(3) \quad d(i,y) + d(j,x) \leq d(i,x) + d(j,y) = d(i,j) + d(x,y)$$

We call these relations as **min-plus inequalities**. Another equivalent expression of the four-point condition is the quadratic Plücker relations (Speyer and Sturmfels, 2004) in tropical geometry literature.

After the introduction of the basic tree concepts and quartet trees, in the next part we will introduce how we represent tree to vector and show the property of this representation by the quartet decomposition of a tree.

3.2 Vectorization of Trees

In this section, we will introduce the vectorization mapping, show its isomorphism, and induce a metric in tree space.

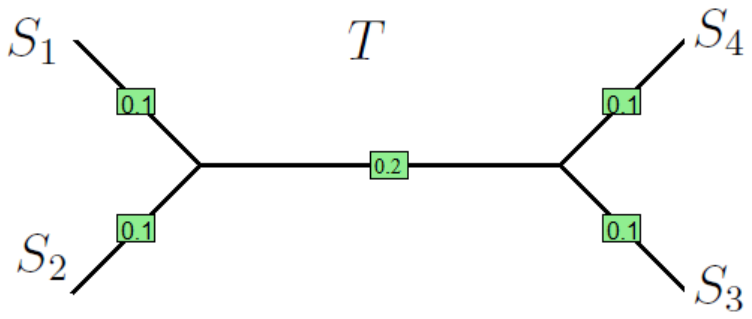

Figure 3.5: A 4-taxa unrooted weighted tree

3.2.1 Pairwise Path Mapping φ

T with m ordered tips $S = (S_1, S_2, \ldots, S_m)$ has $p = \binom{m}{2}$ pairs of taxa. For each pair of taxa (S_i, S_j), there is a graph-theoretic distance between them. The graphical distance between the pair of taxa $(S_i, S_j), i, j \in \{1, 2, \ldots, m\}$ is denoted as $d(S_i, S_j) = d(i, j) = d_{i,j}$. In the same manner exploited by cophenetic vector (Cardona et al., 2013), a vector that captures the distance between each pair of taxa can be defined.

DEFINITION 13 (Pairwise Path Vector)**.** For a tree T with a lexicographically ordered taxa set $S = (S_1, S_2, \ldots, S_m)$, without loss of generality, the **pairwise path vector** of T is

$$\varphi(T) = ((d(i,j)))_{1 \le i < j \le m} = (d_{1,2}, d_{1,3}, \ldots, d_{m-1,m}) \in \mathbb{R}^p_{>0}$$

The mapping φ is called **pairwise path mapping**.

Example 14. Given a tree T as shown by the following Figure 3.5, its pairwise path vector is $\varphi(T) = (0.2, 0.4, 0.4, 0.4, 0.4, 0.2)'$. The vector $\varphi(T)$ specifies a point in $\mathbb{R}^6_{>0}$.

Note that the pairwise path mapping φ projects each unrooted tree T with m tips to a point in $\mathbb{R}^p_{>0}$, and thus represents an element in tree space to the Euclidean space. Next we will explore the property of φ, to see if there is any correspondence between the tree space and its image in vector space.

3.2.2 Injection

Denote $\mathcal{T}_m$ as the set of all the unrooted trees built on the same set of m taxa $S = (S_1, S_2, \ldots, S_m)$. For two unrooted trees T_1 and T_2 in $\mathcal{T}_m$, the comparison of $\varphi(T_1)$ and $\varphi(T_2)$ is not intuitive. As aforementioned, in an unrooted tree, the smallest informative subtree is quartet subtree. Therefore, we will address this problem by decomposing the tree to quartet subtrees, and then the comparison can be made by comparing their sets of quartet subtrees.

DEFINITION 15 (Quartet subtree)**.** For an unrooted tree $T \in \mathcal{T}_m$, a **quartet subtree** is a set of four taxa $\{i, j, x, y\} \subset S$ that inherits from T. A more detailed description is that, when all the branches incident to the vertices not in the quartet $\{i, j, x, y\}$ are removed from T, it results to a subtree of T inherited by the quartet $\{i, j, x, y\}$.

There are $\binom{m}{4}$ quartet subtrees in T. Each quartet subtree will satisfy one of the min-plus inequalities concerning the metric d. The $\binom{m}{4}$ quartets are not independent of all, and a set of $\binom{m}{4}$ quartet subtrees uniquely determines its supertree.

LEMMA 16. *(Steel and Penny, 1993) Any two distinct topologies in $\mathcal{T}_m$ can not have all quartet topologies in common.*

For any two trees with distinct topologies, they cannot have the same set of quartet subtree topologies. Therefore, they must differ in at least one quartet, and thus their pairwise path associated with that quartet cannot be the same. Therefore, the following statement can be deduced.

THEOREM 17 (Injection)**.** $\varphi : \mathcal{T}_m \to \mathbb{R}^p_{>0}$ *is an injective function.*

Proof. For two distinct trees $T_1, T_2 \in \mathcal{T}_m$

(a). If their topologies are distinct, then there is at least one set of four taxa having distinct quartet topology. For this set of four taxa, their corresponding 6 entries in the

projected vectors should follow different min-plus inequalities. Thus, T_1 and T_2 cannot be the same.

(b). If they have the same topology, but distinct edge weight vector. Then for any pair of (i, j), the path from i to j in two trees covers the same set of edges. $d_{i,j}^{(1)}$ in $\varphi(T_1)$ and $d_{i,j}^{(2)}$ in $\varphi(T_2)$ are the same linear combination from $\boldsymbol{b}^{(1)}$ and $\boldsymbol{b}^{(2)}$ respectively. $\boldsymbol{b}^{(1)} \neq \boldsymbol{b}^{(2)}$ guarantees $\varphi(T_1) \neq \varphi(T_2)$. □

Therefore, the injective mapping $\varphi : \mathcal{T}_m \to \mathbb{R}_{>0}^p$ sends each $T \in \mathcal{T}_m$ to its pairwise path vector.

3.2.3 Pairwise Path Metric

Denote $image(\varphi) = \{v \in \mathbb{R}_{>0}^p | v = \varphi(T), \text{ where } T \in \mathcal{T}\}$, which is a subset in $\mathbb{R}_{>0}^p$. The one-to-one relationship between $\mathcal{T}$ and $image(\varphi)$ allows us to induce a metric on $\mathcal{T}_m$ based on the Euclidean norm of vector.

DEFINITION 18. Given two trees $T_k, T_l \in \mathcal{T}_m$, the **pairwise path distance** between two trees is defined as

$$D_{L2}(T_k, T_l) = \|\varphi(T_k) - \varphi(T_l)\|_2 = \sqrt{\sum_{1 \leq i < j \leq m} (d^{(T_k)}(i, j) - d^{(T_l)}(i, j))^2}$$

Obviously D_{L2} is a metric in $\mathcal{T}_m$. The metric tree space $\{\mathcal{T}_m, D_{L2}\}$ will be named as $L2$ **tree space**. For the sake of description, we denote the Euclidean distance between two vectors as $D(\cdot, \cdot)$.

The pairwise path distance is an inner product metric, as well as the Robinson-Foulds distance that is commonly used to measure the topological distance between trees.

Example 19. For the following three trees on the same set of 4 taxa shown in Figure 3.6. The pairwise path distance is $D_{L2}(T_1, T_2) = 0.17$, $D_{L2}(T_1, T_3) = 0.40$, $D_{L2}(T_2, T_3) = 0.44$.

The pairwise path distance between two topologically identical trees can be calculated straightforwardly. For two trees with different topologies, the preliminary work for calculating

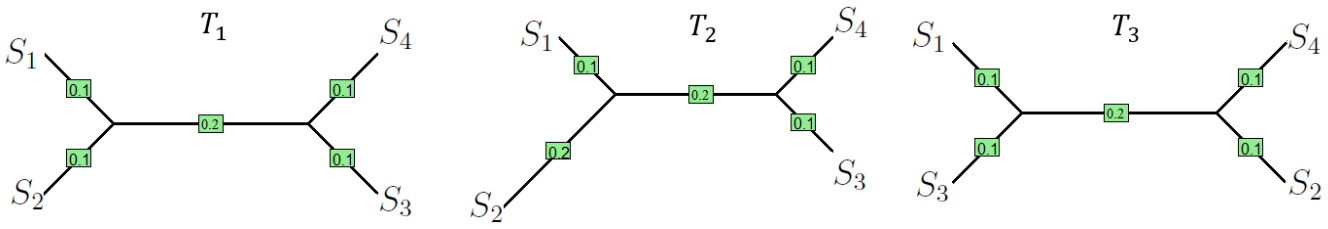

Figure 3.6: Three 4-tax trees with edge lengths

pairwise path distance is the construction of a the set of all paths. It takes $O(m^2)$ time for computation (Bryant and Waddell, 1997). Nonetheless, the statistical analysis of trees based on the pairwise path metric remains untapped.

3.2.4 A topological vector space $\mathcal{V}_m$

We want to explore the property of φ and the structure of $image(\varphi)$, which will be denoted as $\mathcal{V}_m \coloneqq \varphi(\mathcal{T}_m) = image(\varphi)$. The mapping $\varphi : \mathcal{T}_m \to \mathcal{V}_m$ is bijection. Actually, φ is distance preserving. Therefore, the L_2 tree space $(\mathcal{T}_m, D_{L2})$ and the topological vector space $(\mathcal{V}_m, D)$ are isomorphic.

Given that a BHV tree space characterizes the structure of $\mathcal{T}_m$, the structure of $\mathcal{T}_m$ inspires the exploration of $\mathcal{V}_m$.

As the orthant in BHV space shown by Figure 1.1, we define the trees that reside in the same orthant (including orthant boundary) as compatible.

DEFINITION 20. If a topology τ_1 can be got from removing internal edges from another topology τ_2, then τ_1 is called the degeneration of τ_2. If two trees are exactly topologically identical, or one's topology is the degeneration of the other's, then these two trees are **compatible**.

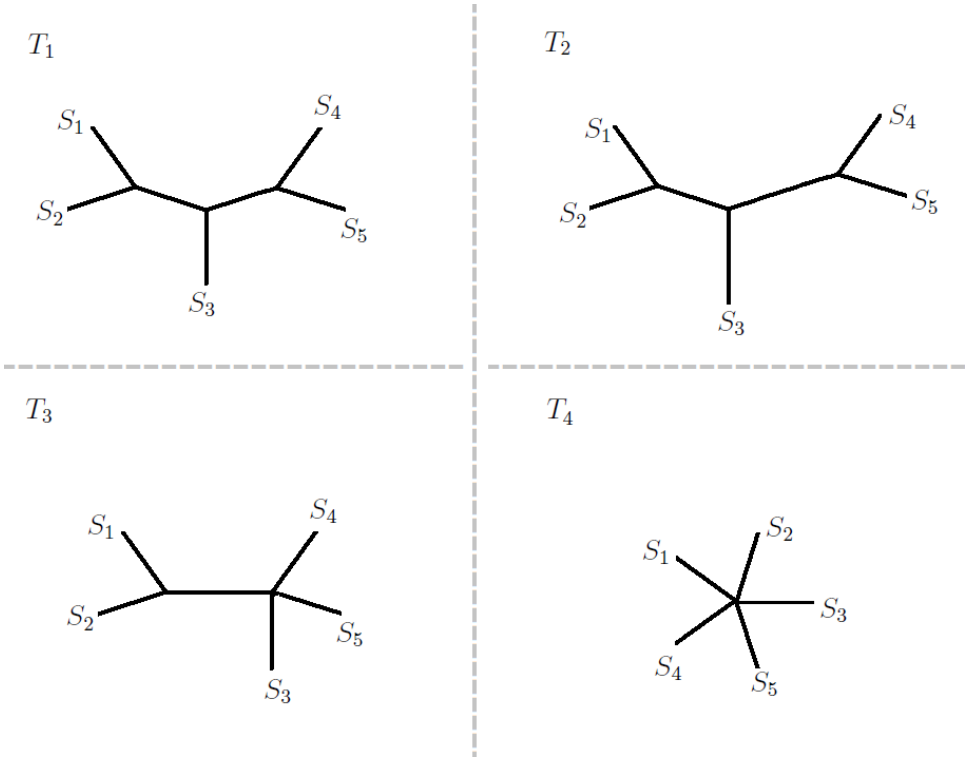

Figure 3.7: Four compatible trees with 5 tips

A tree topology can be viewed as a tree with unit branch lengths. Therefore, in our discussion, we will apply the concepts of degeneration and compatibility to both topologies and trees.

Example 21. The Figure 3.7 shows 4 trees built on the taxa $(S_1, S_2, S_3, S_4, S_5)$. T_1 and T_2 have the same topology and differ in branch lengths, so they are compatible. The topology of T_3 can be got by removing an internal edge from the topology of T_1/T_2, therefore T_3 are compatible with T_1 and T_2. Similarly, T_4 are compatible with T_1, T_2 and T_3. Furthermore, in Figure 3.7, for the tree T_3 with a polytomy with a 4 degree internal node, it is compatible with each of the three binary topology $\tau^{(1)}$, $\tau^{(2)}$ and $\tau^{(3)}$. However, $\tau^{(1)}$, $\tau^{(2)}$ and $\tau^{(3)}$ are not compatible with each other.

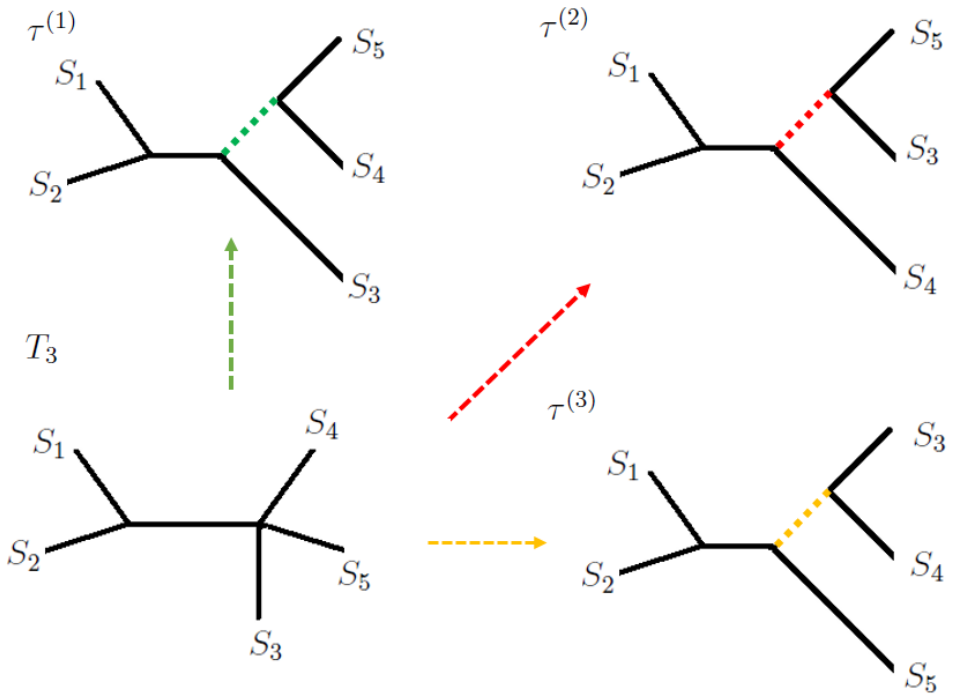

Figure 3.8: A multifurcating tree compatible with multiple binary trees

Any multifurcating tree is compatible with at least three binary topologies. But any two distinct binary topologies are not compatible. Note that there are $(2m-5)!!$ distinct binary topologies in $\mathcal{T}_m$. Any $T \in \mathcal{T}_m$ is compatible with at least one binary topology. Therefore, we can define the compatible tree class in $\mathcal{T}_m$ as induced by the binary topologies .

DEFINITION 22. The $(2m-5)!!$ distinct binary topologies in $\mathcal{T}_m$ are denoted as $\tau^{(k)}$ $(k = 1, 2, \ldots, (2m-5)!!)$. A **compatible tree class** induced by $\tau^{(k)}$ is defined as $\mathcal{T}_m^{(k)} = \{T \in \mathcal{T}_m | T \text{ is compatible with } \tau^{(k)}\}$, $(k = 1, 2, \ldots, (2m-5)!!)$.

Through the mapping φ, each compatible tree class can be projected to a compatible vector set.

DEFINITION 23. The image of $\mathcal{T}_m^{(k)}$ is $\mathcal{V}_m^{(k)} = \{v \in \mathbb{R}^p_{>0} | v = \varphi(T), T \in \mathcal{T}_m^{(k)}\}$, where $k = 1, 2, \ldots, (2m-5)!!$. $\mathcal{V}_m^{(k)}$ can be denoted as $\varphi(\mathcal{T}_m^{(k)})$, and is called a **compatible vector set**.

Note that in a compatible tree class $\mathcal{T}_m^{(k)}$ in BHV tree space, the binary trees which have strict positive internal edge lengths reside within the top-dimensional stratum (Willis, 2017), while the multifurcating trees reside at the boundary of the top-dimensional orthants. To facilitate our analysis, we introduce the following definitions.

DEFINITION 24. A **strictly compatible tree class** induced by $\tau^{(k)}$ is $\mathring{\mathcal{T}}_m^{(k)} = \{T \in \mathcal{T}_m^{(k)}$ and T is binary$\}$. Also, denote its image as $\mathring{\mathcal{V}}_m^{(k)} = \varphi(\mathring{\mathcal{T}}_m^{(k)})$

Since $\mathring{\mathcal{T}}_m^{(k)} \subset \mathcal{T}_m^{(k)}$, and φ is injective, thus $\mathring{\mathcal{V}}_m^{(k)} \subset \mathcal{V}_m^{(k)}$. The $(2m-5)!!$ compatible tree classes constitute the whole tree space. It is intuitive to deduce that the $\mathcal{V}_m$ is the collection of the $(2m-5)!!$ compatible vector sets. We are interested in $\mathcal{V}_m$ because, if we can characterize $\mathcal{V}_m$, then it is an alternative way to describe the tree space $\mathcal{T}_m$. In the next part, we will explore the structure of $\mathcal{V}_m$.

3.3 Exploration On Quartet Trees

We will start from the trivial situation of $\mathcal{T}_4$ and $\mathcal{V}_4$. Our exploration of the mapping φ and the topological vector space $\mathcal{V}_4$ includes two main aspects. First, the vectorization mapping φ is a piecewise linear transformation that can be operated by a matrix associated with topology. Second, by analyzing the composite of the linear operator matrix, the structure of the topological vector space $\mathcal{V}_4$ will be presented.

From the phylogenetic perspective, there are two essential elements featuring a tree T: the topology and edge lengths. We denote a phylogenetic tree as $T = (\tau, \boldsymbol{b})$. For any $T \in \mathcal{T}_4$, it has 4 possible topologies as shown in Figure 3.4. In $\mathcal{T}_4$, the star tree is compatible with either of the three binary topologies in 3.9, by viewing it as a special case with the internal

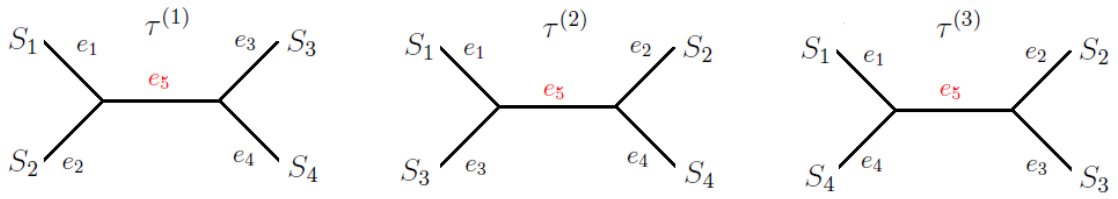

Figure 3.9: Three 4-taxa unrooted binary tree

edge length decreasing to 0. Denote the compatible tree class in $\mathcal{T}_4$ induced by $\tau^{(k)}$ as $\mathcal{T}_4^{(k)}$, where $k = 1, 2, 3$. Then,

PROPOSITION 25. $\bigcap_{k=1}^{3} \mathcal{T}_4^{(k)} = \{quartet\ star\ tree\} \subset \mathcal{T}_4 = \bigcup_{k=1}^{3} \mathcal{T}_4^{(k)}$

For any $T \in \mathcal{T}_4$, it has 4 positive terminal edges and 1 non-negative internal edge. Without loss of generality, we assign e_i as connected to S_i where $i = 1, 2, 3, 4$ and the last edge e_5 as the internal edge. The edge weights will be given as $\boldsymbol{b} = (b(e_1), b(e_2), b(e_3), b(e_4), b(e_5))' \in \mathbb{R}^5_{\geq 0}$. To differentiate the terminal edges and internal edge, we denote the edge length vector $\boldsymbol{b}$ as $\boldsymbol{b} = (\boldsymbol{b}'_{TER}, b_{INT})'$ where $\boldsymbol{b}_{TER} \in \mathbb{R}^4_{>0}$ represents the edge lengths for the 4 terminal edges and $b_{INT} \in \mathbb{R}_{\geq 0}$ is the edge length for the one internal edge. In other words, $\boldsymbol{b} = (\boldsymbol{b}'_{TER}, b_{INT})'$ can be denoted as $\boldsymbol{b} \in \mathbb{R}^4_{>0} \oplus \mathbb{R}_{\geq 0}$.

3.3.1 Characterizing The Mapping by Matrices

Assume that for $k = 1, 2, 3$, $T_k = (\tau^{(k)}, \boldsymbol{b}) \in \mathcal{T}_4^{(k)}$ as shown by Figure 3.9. Then,

$$\varphi(T_1) = \begin{pmatrix} b(e_1)+b(e_2) \\ b(e_1)+b(e_3)+b(e_5) \\ b(e_1)+b(e_4)+b(e_5) \\ b(e_2)+b(e_3)+b(e_5) \\ b(e_2)+b(e_4)+b(e_5) \\ b(e_3)+b(e_4) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \boldsymbol{b} \coloneqq M^{(1)}\boldsymbol{b}$$

$$\varphi(T_2) = \begin{pmatrix} b(e_1)+b(e_2)+b(e_5) \\ b(e_1)+b(e_3) \\ b(e_1)+b(e_4)+b(e_5) \\ b(e_2)+b(e_3)+b(e_5) \\ b(e_2)+b(e_4) \\ b(e_3)+b(e_4)+b(e_5) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \boldsymbol{b} \coloneqq M^{(2)}\boldsymbol{b}$$

$$\varphi(T_3) = \begin{pmatrix} b(e_1)+b(e_2)+b(e_5) \\ b(e_1)+b(e_3)+b(e_5) \\ b(e_1)+b(e_4) \\ b(e_2)+b(e_3) \\ b(e_2)+b(e_4)+b(e_5) \\ b(e_3)+b(e_4)+b(e_5) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \boldsymbol{b} \coloneqq M^{(3)}\boldsymbol{b}$$

The matrix $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$ are the 6×5 matrices. The mapping $\varphi : \mathcal{T}_4 \to \mathbb{R}^6_{>0}$ can be characterized by three linear operators $M^{(1)}, M^{(2)}, M^{(3)}$, according to the tree topology. Thus, we can present the mapping φ by the matrix.

PROPOSITION 26. *For a tree $T = (\tau, \boldsymbol{b})$, the pairwise path mapping $\varphi : \mathcal{T}_4 \to \mathbb{R}^6_{>0}$ can be presented as*

$$\varphi(T) = \begin{cases} M^{(1)}\boldsymbol{b}, & T \in \mathcal{T}_4^{(1)} \\ M^{(2)}\boldsymbol{b}, & T \in \mathcal{T}_4^{(2)} \\ M^{(3)}\boldsymbol{b}, & T \in \mathcal{T}_4^{(3)} \end{cases}$$

φ is a piecewise linear transformation on $T = (\tau, \boldsymbol{b})$.

Each quartet topology is described by a matrix, or say, a linear operator. From $\varphi(T = (\tau; \boldsymbol{b})) = M\boldsymbol{b}$, we can say that the topology of a quartet tree is associated with the linear map.

3.3.2 Visualization of $\mathcal{V}_4$

In each compatible tree class, $\varphi(T)$ is a linear transformation on the edge length vector. The strictly compatible vector set $\mathring{\mathcal{V}}_4^{(k)}$ corresponding to $\mathring{\mathcal{T}}_4^{(k)}$ is the collection of vectors corresponding to the binary trees with topology $\tau^{(k)}$.

PROPOSITION 27. $\mathring{\mathcal{V}}_4^{(k)} = \{M^{(k)}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^5_{>0}\}$ *is the positive orthant in the column space of* $M^{(k)}$.

Relating to BHV tree space, the vector set corresponding to the trees in the kth "top-dimensional stratum" is the positive orthant in the column space of $M^{(k)}$.

Note that in matrix $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$, the first 4 columns are the same. Denote the first four columns as $M^{(0)}$. The entries on the 6 rows of $M^{(0)}$ are the indicator of the $\binom{4}{2}$ combinations. The last column corresponds to the binary quartet topology by the min-plus inequalities. Note that in column5, the entry of (i, j) is 0 means there is no internal edge (e_5) between the pair of taxa (i, j). Thus, the last column is, in fact, an indicator of the bipartition.

DEFINITION 28. (Split Indicator) In a tree with edge e_l, the vector $\boldsymbol{I}_l$ with $\binom{m}{2}$ entries has values of 0 or 1 that

$$I_l(i,j) = \begin{cases} 1, & \text{the pair } i \text{ and } j \text{ are separated by } e_l \\ 0, & \text{the pair } i \text{ and } j \text{ are on the same part of the bipartition by } e_l \end{cases}$$

The indictor vector $\boldsymbol{I}_l$ is called the **split indicator** of e_l.

The column 5 in $M^{(k)}$ is denoted as $\boldsymbol{I}^{(k)}$ respectively for $k = 1, 2, 3$. In $\mathcal{T}_4^{(1)}$, S_1 and S_2 are on the same part, and S_3 and S_4 are on the same part. Thus, the entry of $I^{(1)}$ at the

position $(1,2)$ and $(3,4)$ are 0. Likewise,

$$M^{(k)} = \begin{pmatrix} M^{(0)} & \boldsymbol{I}^{(k)} \end{pmatrix}, \text{ where } \begin{cases} \boldsymbol{I}^{(1)} &= (0,1,1,1,1,0) \\ \boldsymbol{I}^{(2)} &= (1,0,1,1,0,1) \\ \boldsymbol{I}^{(3)} &= (1,1,0,0,1,1) \end{cases}$$

The split indicator $\boldsymbol{I}^{(k)}$ contains enough information to identify the topology. In this sense, $\boldsymbol{I}^{(k)}$ is related to the Robinson-Foulds distance (Robinson and Foulds, 1981), which counts the distinct splits among two trees.

DEFINITION 29. (Steel and Penny, 1993) The Robinson-Foulds distance between two trees T_1 and T_2 is

$\#\{\text{internal edges in } T_1\}+\#\{\text{internal edges in } T_2\}-2\#\{\text{internal splits shared by } T_1 \text{ and } T_2\}$

, where # indicates the number of elements in the set.

For two trees, if their split indicators corresponding to the internal edge e_l are the same, then the splits created by e_l are same. With respect to this edge, the number of internal splits shared by the two trees increases by 1. Otherwise, the shared split created by this edge will be 0. Thus, the Robinson-Foulds distance can be calculated by the split indicator.

PROPOSITION 30. *For two binary trees $T_1 \in \mathcal{T}_4^{(k_1)}$ and $T_2 \in \mathcal{T}_4^{(k_2)}$, where $k_1, k_2 \in \{1,2,3\}$, there is*

$$D_{RF}(T_1,T_2) = 2 - 2 \times \#\{\boldsymbol{I}^{(k_1)} = \boldsymbol{I}^{(k_2)}\} = \begin{cases} 2, & \text{if } \boldsymbol{I}^{(k_1)} = \boldsymbol{I}^{(k_2)} \\ 0, & \text{otherwise} \end{cases}$$

Remark 31. An equivalent expression for $\boldsymbol{I}^{(k_1)} = \boldsymbol{I}^{(k_2)}$ is $(\boldsymbol{I}^{(k_1)})'(\boldsymbol{j}_6 - \boldsymbol{I}^{(k_2)}) = 0$ where $\boldsymbol{j}_6$ is an all 1's vector with dimension 6. It can be checked that , for T_1 and T_2 as binary trees in $\mathcal{T}_4$, $D_{RF}(T_1,T_2) = 2 - (\boldsymbol{I}^{(k_1)})'(\boldsymbol{j}_6 - \boldsymbol{I}^{(k_2)}) = 4 - (\boldsymbol{I}^{(k_1)})'\boldsymbol{I}^{(k_2)}$.

The exception for the above proposition is the star tree. For the star tree, its internal edge length is 0. Therefore any linear operator $M^{(k)}$ on it gives the same result. Its corresponding vector set is $\mathcal{V}_4^{(0)} = \{\varphi(T) | T \text{ is star tree in } \mathcal{T}_4\}$.

PROPOSITION 32. *The image of quartet start trees* $\mathcal{V}_4^{(0)} = \bigcap_{k=1}^{3} \mathcal{V}_4^{(k)} = \{M^{(k)}\boldsymbol{b},$ *where* $b_{INT} = 0\} = \{M^{(0)}\boldsymbol{b}_{TER}, \boldsymbol{b}_{TER} \in \mathbb{R}^4_{>0}\}$ *is the positive orthant in the column space of* $M^{(0)}$.

A compatible tree class $\mathcal{T}_4^{(k)}$ is a top-dimensional stratum and its boundaries in tree space. Likewise,

PROPOSITION 33. $\mathcal{V}_4^{(k)} = \mathring{\mathcal{V}}_4^{(k)} \bigcup \mathcal{V}_4^{(0)}$

Though $v \in \mathcal{V}_4^{(k)}$ is a vector length of 6, $\mathcal{V}_4^{(k)}$ is vector set with dimension 5 since the column rank of $M^{(k)}$ is 5. In $\mathcal{T}_4$, the kth ($k = 1, 2, 3$) compatible tree class have the common boundary as the set of star trees. Correspondingly, $\mathcal{V}_4^{(k)}$ ($k = 1, 2, 3$) coincides with each other on $\mathcal{V}_4^{(0)}$. The structure of $\mathcal{V}_4^{(1)}$, $\mathcal{V}_4^{(2)}$ and $\mathcal{V}_4^{(3)}$ can be described as grafted with each other as 3.10 shows, similar to BHV tree space.

Figure 3.10 gives us a graphic visualization on the topological vector space $\mathcal{V}_4$. Each tree in tree class $\mathcal{T}_4^{(k)}$ with edge length vector $\boldsymbol{b} \in \mathbb{R}^4_{>0} \oplus \mathbb{R}_{\geq 0}$ is projected to a point in a semi-positive orthant of a space ranged by the 6×5 matrix $M^{(k)}$, which is a subset in $\mathbb{R}^6_{>0}$. Its coordinate is $M^{(k)}\boldsymbol{b} = M^{(0)}\boldsymbol{b}_{TER} + I^{(k)}b_{INT}$. As the internal edge b_{INT} decreases to 0, the corresponding vector moves to the "boundary" $\mathcal{V}_4^{(0)}$ with dimension 4.

Note that in Figure 3.10, the space ranged by $M^{(0)}$, $I^{(1)}$, $I^{(2)}$ and $I^{(3)}$ are not orthogonal, as none of the inner products is 0.

3.4 Generalization to Large Trees

We have shown several excellent properties about the tree space $(\mathcal{T}_4, D)$ and its corresponding vector set $\mathcal{V}_4$. Most conclusions in $\mathcal{T}_4$ are very straightforward. In this section, we will see if the parallel conclusions hold in the general $\mathcal{T}_m$ when $m > 4$.

LEMMA 34. *1.* $\mathcal{T}_m = \bigcup_{k=1}^{(2m-5)!!} \mathcal{T}_m^{(k)}$.

2. $\mathcal{T}_m^{(0)} := \bigcap_{k=1}^{(2m-5)!!} \mathcal{T}_m^{(k)}$ *is the set of (m−degree) star trees in* $\mathcal{T}_m$.

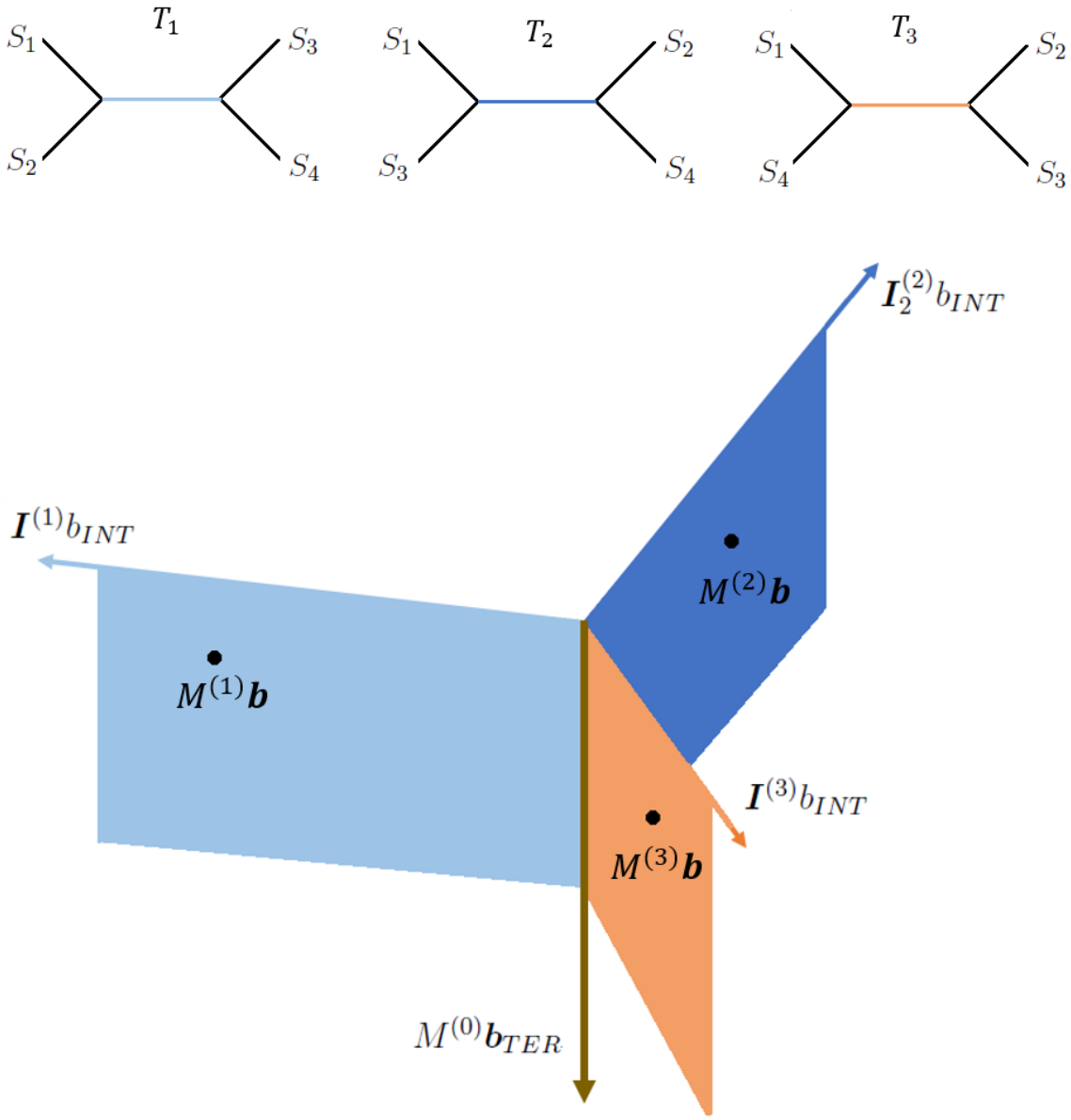

Figure 3.10: The structure of $\mathcal{V}_4$ in $\mathbb{R}^6$

Proof. (1). According to the definition of $\mathcal{T}_m^{(i)}$, there is $\bigcup_{k=1}^{(2m-5)!!} \mathcal{T}_m^{(i)} \subset \mathcal{T}_m$. For any $T \in \mathcal{T}_m$, if T is a binary tree with topology $\tau^{(k_0)}$, then T is in the tree class $\mathcal{T}_m^{(k_0)}$. If T is a tree with polytomy, then it can finally become a binary tree T^0 by adding branches into it, such that T is in the same tree class with the binary tree T^0.

(2)A star tree can become any other topology by adding branches, therefore the set of star tree is subset in $\mathcal{T}_m^{(0)}$. For a $T \in \bigcap_{k=1}^{(2m-5)!!} \mathcal{T}_m^{(k)}$, assume T has an internal edge e. e splits T into two parts, with each part have at least two tips. Arbitrarily select tips (i, j) from one part and (x, y) from the other part. The quartet subtree (i, j, x, y) has the quartet topology of $\tau^{(1)}$. Manipulate T by exchanging the label of j and x such that T becomes a different tree $\tilde{T}$. Then the quartet subtree (i, j, x, y) in $\tilde{T}$ has the topology of $\tau^{(2)}$. The quartet subtrees inherit by the same set of four tips from T and $\tilde{T}$ conflict, therefore T and $\tilde{T}$ are incompatible. This is contradiction to the assumption of $T \in \bigcap_{k=1}^{(2m-5)!!} \mathcal{T}_m^{(k)}$. □

3.4.1 Characterizing the mapping by matrices

For $T \in \mathcal{T}_m$, without loss of generality, order its edges lexicographically, such that the first m edges are terminal edges incident to $(S_1, S_2, \ldots, S_m)$ while the last $(m-3)$ are internal edges. Let $\boldsymbol{b}$ be the edge length vector. In $T \in \mathcal{T}_m$ with $(2m-5)!!$ possible topologies, the path from S_i to S_j definitely covers e_i and e_j while no other terminal edges. For $m+1 \leq l \leq 2m-3$, if e_l separates S_i and S_j, then the path from S_i to S_j covers e_l. Therefore, the mapping φ is determined by the topology, as well as the matrix operator representing the topology.

LEMMA 35. *For $T = (\tau, \boldsymbol{b}) \in \mathcal{T}_m^{(k)}$ $(1 \leq k \leq (2m-5)!!)$, $\varphi(T) = M^{(k)}\boldsymbol{b}$, where $M^{(k)} : p \times (2m-3)$ is the linear operator determined by the topology $\tau^{(k)}$.*

In matrix $M^{(k)}$, the first m columns, which will be multiplied with the first m edge lengths to contribute to the terminal composite of the path, constitute to a $p \times m$ matrix $M^{(0)}$ whose entries on the p rows are the indicators of the $\binom{m}{2}$ combinations. The last $(m-3)$ columns are the split indicators by the internal edges $\boldsymbol{I}_l^{(k)}$ where $(m+1 \leq l \leq 2m-3)$. The last $(m-3)$ columns constitute a $p \times (m-3)$ matrix, denoted as $M_{INT}^{(k)}$.

Within each compatible class, $\varphi(T)$ is a linear transformation the edge length vector. Thus the statement that $\varphi(T)$ is a piecewise linear transformation still holds.

In the tree class $\mathcal{T}_m^{(k)}$, φ is presented by a $p \times (2m-3)$ matrix $M^{(k)}$. We will characterize the mapping φ and the image $\varphi(\mathcal{T}_m^{(k)})$ by analyzing the matrix $M^{(k)}$.

LEMMA 36. *Each $M^{(k)}$ is full column rank.*

Proof. The proof is shown in the appendix. □

In order to compare two trees, we will look at the matrices corresponding to them.

LEMMA 37. *For two binary trees $T_1 \in \mathcal{T}_m^{(k_1)}$ and $T_2 \in \mathcal{T}_m^{(k_2)}$, where $k_1, k_2 \in \{1, 2, \ldots, (2m-5)!!\}$, there is*

$$D_{RF}(T_1, T_2) = 2(m-3) - 2 \times \#\{\boldsymbol{I}_l^{(k_1)} = \boldsymbol{I}_h^{(k_2)}, m+1 \leq l, h \leq 2m-3\}$$

Proof. $\boldsymbol{I}_l^{(k_1)} = \boldsymbol{I}_h^{(k_2)}$ indicates that the split created by edge e_l in T_1 is same with the split created by edge e_h in T_2, contributing to the same bipartition shared by T_1 and T_2. Therefore, $\#\{\boldsymbol{I}_l^{(i_1)} = \boldsymbol{I}_h^{(i_2)}, m+1 \leq l, h \leq 2m-3\}$ is the number of all the splits shared by T_1 and T_2. □

Example 38. For four trees T_1, T_2, $T_3$and T_4 shown in Figure 3.11.

The splits created by internal edges in T_1 are $\{S_1, S_2|S_3, S_4, S_5, S_6\}$, $\{S_1, S_2, S_3|S_4, S_5, S_6\}$, $\{S_1, S_2, S_3, S_4|S_5, S_6\}$.

Splits created by internal edges in T_2 are $\{S_1, S_2|S_3, S_4, S_5, S_6\}$, $\{S_1, S_2, S_5|S_3, S_4, S_6\}$, $\{S_1, S_3, S_4, S_5, S_6|S_3, S_4\}$.

Splits created by internal edges in T_3 are $\{S_1, S_3|S_2, S_4, S_5, S_6\}$, $\{S_1, S_3, S_4|S_2, S_5, S_6\}$, $\{S_1, S_3, S_4, S_5|S_2, S_6\}$.

Splits created by internal edges in T_4 are $\{S_1, S_2|S_3, S_4, S_5, S_6\}$, $\{S_3, S_4|S_1, S_2, S_5, S_6\}$, $\{S_5, S_6|S_1, S_2, S_3, S_4\}$.

In the graph, the internal edge colored as orange creates the same bipartition in T_1 and T_2. There is one common split ($\{S_1, S_2|S_3, S_4, S_5, S_6\}$) shared by T_1 and T_2. Moreover, the

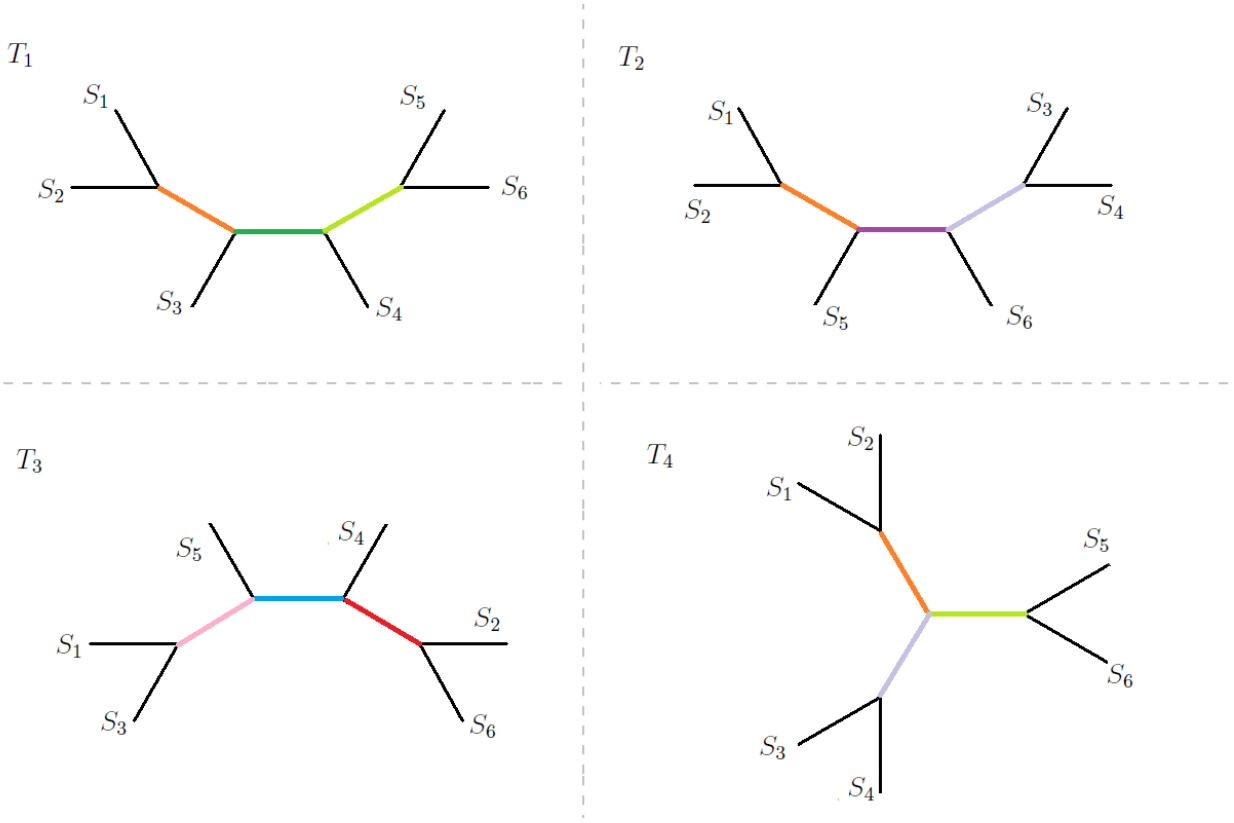

Figure 3.11: An example of four 6-taxa binary unrooted trees with distinct topologies

split indicators of this edge in the two trees are the same. The orange internal edge and light green internal edge create the same bipartition in T_1 and T_4. There are two common splits $(\{S_1, S_2|S_3, S_4, S_5, S_6\})$ and $(\{S_5, S_6|S_1, S_2, S_3, S_4\})$shared by T_1 and T_4. The split indicators in the matrix corresponding to the T_1 and T_4 have two common split indicators. T_3 has no common split with any of the other three trees. Likewise, there is no same split indicator in the associated matrix of T_3 and other matrices.

$D_{RF}(T_1, T_4) = D_{RF}(T_2, T_4) = 2$, $D_{RF}(T_1, T_2) = 4$, $D_{RF}(T_1, T_3) = D_{RF}(T_2, T_3) = D_{RF}(T_4, T_3) = 6$.

In the corresponding matrices $M^{(k_1)}$, $M^{(k_2)}$, $M^{(k_3)}$ and $M^{(k_4)}$,

$\#\{\boldsymbol{I}_l^{(k_1)} = \boldsymbol{I}_h^{(k_4)}, 7 \leq l, h \leq 9\} = \#\{\boldsymbol{I}_l^{(k_4)} = \boldsymbol{I}_h^{(k_2)}, 7 \leq l, h \leq 9\} = 2$,

$\#\{\boldsymbol{I}_l^{(k_1)} = \boldsymbol{I}_h^{(k_2)}, 7 \leq l, h \leq 9\} = 1$,

$\#\{\boldsymbol{I}_l^{(k_1)} = \boldsymbol{I}_h^{(k_3)}, 7 \leq l, h \leq 9\} = \#\{\boldsymbol{I}_l^{(k_2)} = \boldsymbol{I}_h^{(k_3)}, 7 \leq l, h \leq 9\} = \#\{\boldsymbol{I}_l^{(k_4)} = \boldsymbol{I}_h^{(k_3)}, 7 \leq l, h \leq 9\} = 0$

This example provides a perspective on the relationship between bipartition and the structure of $M_{INT}^{(k)}$. The $(m-3)$ columns in $M_{INT}^{(k)}$ are independently from each other, and independent with $M^{(0)}$. However, the $(2m-5)!! \times (m-3)$ columns in the collection of $\{M_{INT}^{(k)}, k = 1, 2, \ldots, (2m-5)!!\}$ will have a huge number of repeats. So, we are interested in the configuration that how $M^{(k)}\boldsymbol{b}$ constitute $\mathcal{V}_m$.

$M_{INT}^{(k)}$ has $(m-3)$ columns as split indicators determined by the internal edges. Because one split indicator corresponds to one split. Thus, the total number of different split indicators in $\{M_{INT}^{(k)}, k = 1, 2, \ldots, (2m-5)!!\}$ equals the number of different splits, denoted as $\#\{splits\}$.

$$\#\{splits\} = \begin{cases} \binom{m}{2} + \binom{m}{3} + \cdots + \binom{m}{[(m-1)/2]}, & \text{when } m \text{ is odd} \\ \binom{m}{2} + \binom{m}{3} + \cdots + \frac{1}{2}\binom{m}{m/2}, & \text{when } m \text{ is even} \end{cases} = 2^{m-1} - m - 1$$

.

3.4.2 Structure of $\mathcal{V}_m$

The Petersen graph shown by Figure 3.12 depicts the structure of $\mathcal{T}_5$. In this graph, there are 10 vertices (presented by the bullets) and 15 edges (edges connecting the bullets). Each vertex represents a tree with a polytomy with a 4-degree internal node. Each edge represents a binary topology. Three edges sharing a common end vertex means that three distinct binary topologies can degenerate to the same multifurcating topology with a polytomy with a 4-degree internal node.

In this part, we will explore how the topological vector sets $\mathcal{V}_m^{(k)}$ constitute the image $\mathcal{V}_m$ based on the repetition among the columns in $\{M^{(k)}, k = 1, 2, \ldots, (2m-5)!!\}$.

As we have pointed out, in matrix $M^{(k)}$ $(k = 1, 2, \ldots, (2m-5)!!)$, the first m columns are $M^{(0)}$. The last $(m-3)$ columns are the split indicators $\boldsymbol{I}_{m+1}^{(k)}, \boldsymbol{I}_{m+2}^{(k)}, \ldots, \boldsymbol{I}_{2m-3}^{(k)}$. Denote the edge length vector as $\boldsymbol{b} = (\boldsymbol{b}'_{TER}, \boldsymbol{b}'_{INT}) \in \mathbb{R}_{>0}^m \oplus \mathbb{R}_{\geq 0}^{m-3}$, where $\boldsymbol{b}_{TER} \in \mathbb{R}_{>0}^m$ represents the edge lengths of the m terminal edges, and $\boldsymbol{b}_{INT} = (b(e_{m+1}), b(e_{m+2}), \ldots, b(e_{2m-3}))' \in \mathbb{R}_{\geq 0}^{m-3}$ represents the lengths of the $(m-3)$ internal edges. For $T \in \mathcal{T}_m^{(k)}$, its pairwise path vector is

$$\varphi(T) = M^{(k)}\boldsymbol{b} = M^{(0)}\boldsymbol{b}_{TER} + \sum_{l=m+1}^{2m-3} b(e_l)\boldsymbol{I}_l^{(k)}$$

The image of the top-dimensional stratum $\mathring{\mathcal{T}}_m^{(k)}$ is the vector set characterized by matrix $M^{(k)}$ in space $\mathbb{R}_{>0}^p$.

PROPOSITION 39. $\mathring{\mathcal{V}}_m^{(k)} = \{M^{(k)}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}_{>0}^{2m-3}\}$ *is the positive orthant in the column space of* $M^{(k)}$.

As one internal edges e_h in $T \in \mathring{\mathcal{T}}^{(k)}$ decreasingly approaches to 0, the corresponding vector $\varphi(T)$ moves to the boundary of the $\mathring{\mathcal{V}}_m^{(k)}$, which is a lower dimensional stratum characterized by a matrix with column rank $(2m-2)$.

Example 40. Three distinct binary trees with 5-taxa T_1, T_2 and T_3 in Figure 3.13 are part of the Petersen graph shown by Figure 3.12. From the Petersen graph, we can see that the

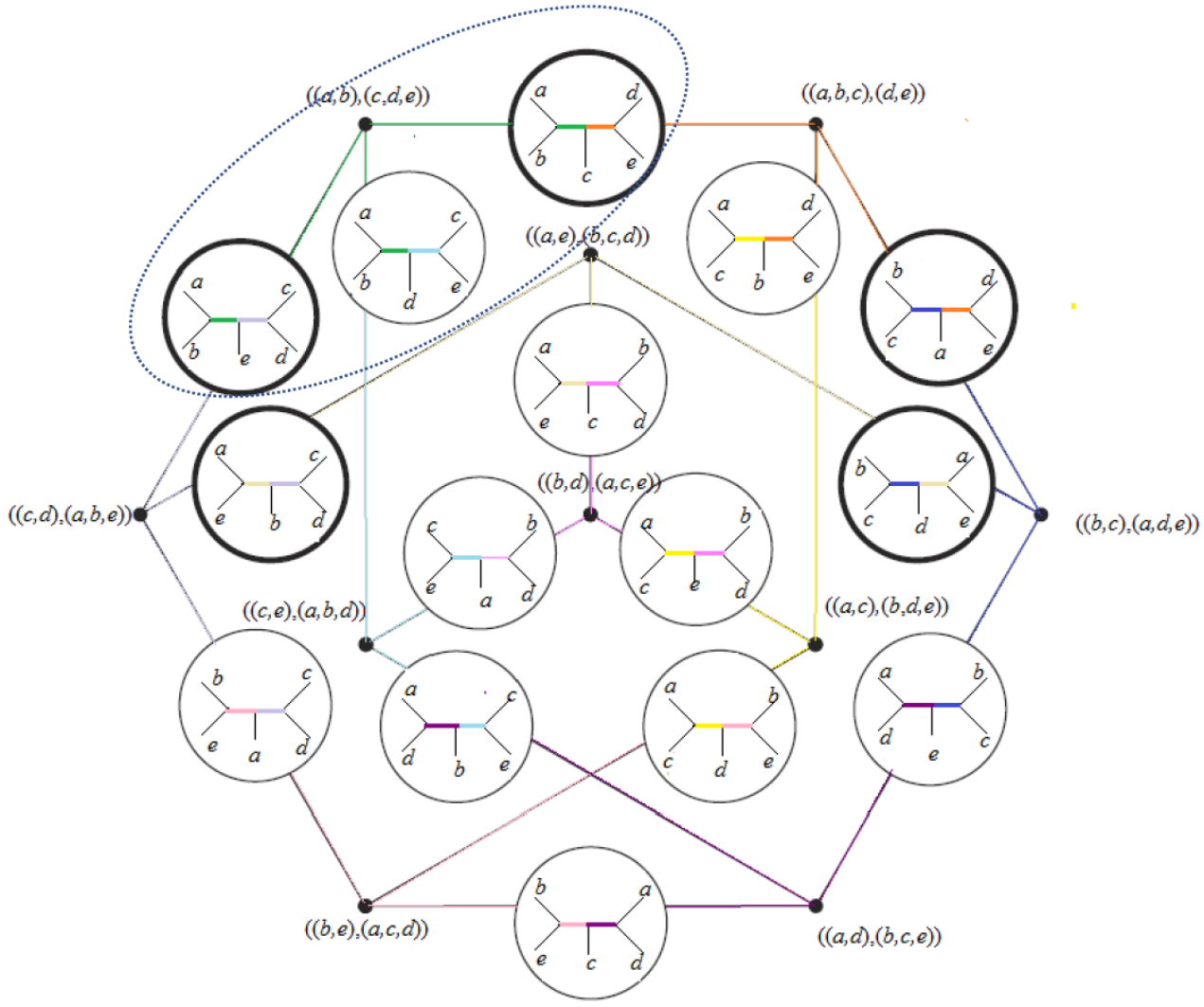

Figure 3.12: The Peterson graph depicting the 15 unrooted topologies of 5-taxa

three trees are connected to a common vertex which represents the tree $((a,b),(c,d,e))$ with a polytomy and maintains the internal edge who create the split as $\{a,b|c,d,e\}$.

From the view of vectors, the corresponding vectors are in the facet restricted by $M^{(k)}$ as

$$\varphi(T_k) = M^{(k)}\boldsymbol{b} = M^{(0)}\boldsymbol{b}_{TER} + b(e_6)\boldsymbol{I}_6^{(k)} + b(e_7)\boldsymbol{I}_7^{(k)}$$

Since $\boldsymbol{I}_6^{(k)}$, $\boldsymbol{I}_7^{(k)}$ and $M^{(0)}$ are independent, thus the existence of each internal edge increases the dimension of the range of a $p-$length vector $\varphi(T_k)$ by 1. Reversely, removal of one internal edge moves the point to a lower dimensional space, which is the vector set corresponding to a tree with polytomy. Beside the first 5 columns as $M^{(0)}$, $M_{INT}^{(1)}, M_{INT}^{(2)}$ and $M_{INT}^{(3)}$ have a common column of $\boldsymbol{I}_6^{(k)}$, denoted as $\boldsymbol{I}_6^{(0)}$, and presented in the graph by the internal edge colored in green. By removing the "orange" edge in T_1, the "light blue" edge in T_2, and "light purple" edge in T_3, the three trees degenerate to their consensus tree as $((a,b),(c,d,e))$, which has the "green" edge whose split indicator is $\boldsymbol{I}_6^{(0)}$.

In general, the topological vector set isomorphic to the tree space $\mathcal{T}_m$ is

$$\begin{aligned}
\mathcal{V}_m &= \{\varphi(T) \in \mathbb{R}_{>0}^p | T \in \textstyle\bigcup_{k=1}^{(2m-5)!!} \mathcal{T}^{(k)}\} \\
&= \{M^{(k)}\boldsymbol{b} | k = 1,2,\ldots,(2m-5)!!, \boldsymbol{b} \in \mathbb{R}_{>0}^m \oplus \mathbb{R}_{\geq 0}^{m-3}\} \\
&= \{M^{(0)}\boldsymbol{b}_{TER} + \sum_{l=m+1}^{2m-3} b(e_l)\boldsymbol{I}_l^{(k)} | k = 1,2,\ldots,(2m-5)!!, \boldsymbol{b}_{TER} \in \mathbb{R}_{>0}^m, b(e_l) \geq 0\}
\end{aligned}$$

Since there are $2^{m-1} - m - 1$ possible distinct $\boldsymbol{I}_l^{(k)}$, thus besides $M^{(0)}$, the top-dimensional stratums $\{\mathring{\mathcal{V}}_m^{(k)}, k = 1,2,\ldots,(2m-5)!!\}$ have $(2m-5)!!$ facets with $2^{m-1} - m - 1$ boundaries in total.

For any two matrices $M^{(k)}$ and $M^{(h)}$ $(k \neq h)$, the number of their common split indicators ranges from 0 to $(m-2)$. If the two matrices have no common split indicator, then $\mathcal{T}_m^{(k)} \bigcap \mathcal{T}_m^{(h)}$ is the set of star trees. If $M^{(k)}$ and $M^{(h)}$ have a collection of common split indicators , then their corresponding topology $\tau^{(k)}$ and $\tau^{(h)}$ have the strict consensus tree with the internal edges associated with the common split indicators.

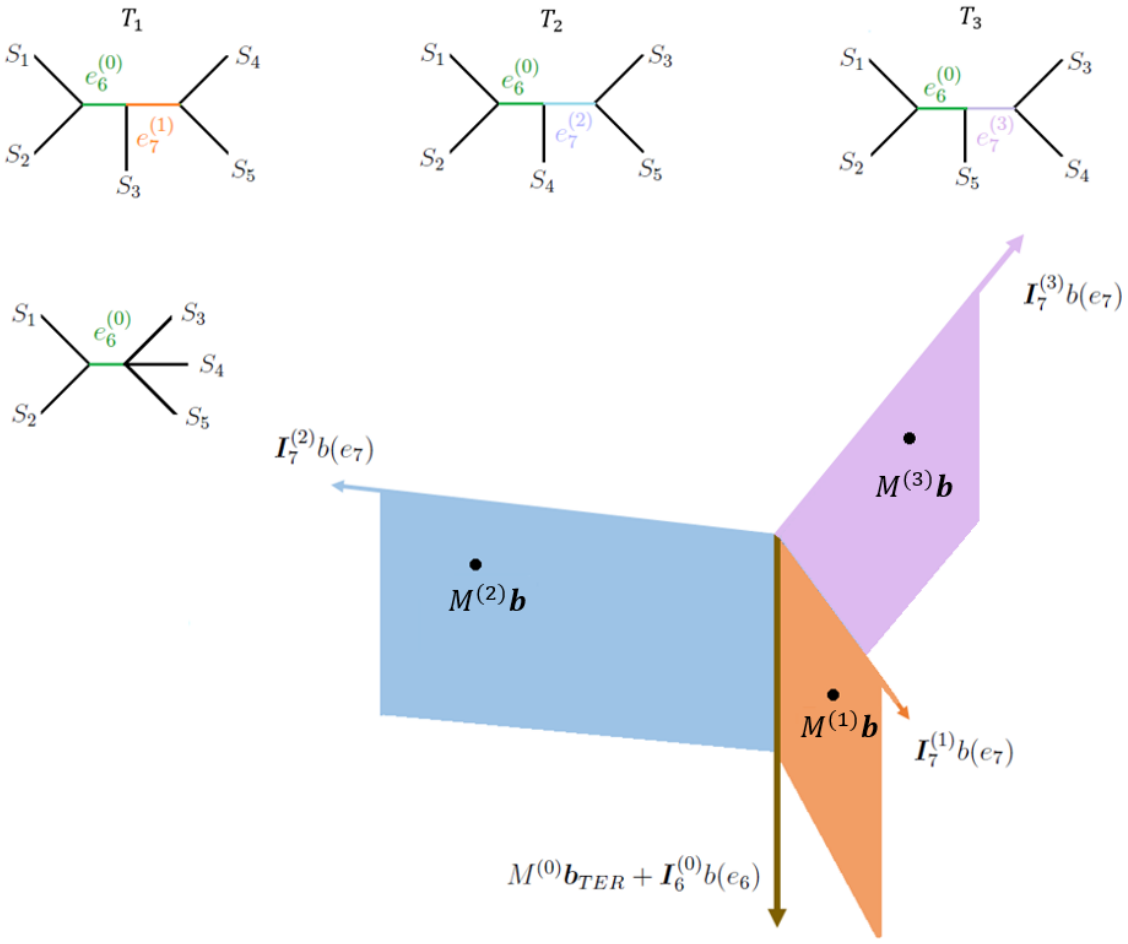

Figure 3.13: A part of $\mathcal{V}_5$ in $\mathbb{R}^{10}$

PROPOSITION 41. $\mathcal{V}_m^{(k)} \bigcap \mathcal{V}_m^{(h)} = \{M^{(0)}\boldsymbol{b}_{TER} + \sum_{\boldsymbol{I}_z \text{ exists both in } M^{(k)} \text{ and } M^{(h)}} \boldsymbol{I}_z b(e_z), \boldsymbol{b} \in \mathbb{R}^m_{>0}, b(e_z) > 0\} = \{\varphi(T) | T \text{ is compatible with the consensus tree of } \tau^{(k)} \text{ and } \tau^{(h)}\}$.

PROPOSITION 42. $\mathcal{V}_m^{(0)} \coloneqq \bigcap_{k=1}^{(2m-5)!!} \mathcal{V}_m^{(k)} = \{M^{(0)}\boldsymbol{b}, \boldsymbol{b} \in \mathbb{R}^m_{>0}\}$ *is the positive orthant in the column space of* $M^{(0)}$*, and is the vector set corresponding to* $m-$*degree star trees.*

Because $\mathcal{V}_m$ is characterized by $M^{(0)}$ and the $2^{m-1} - m - 1$ split indicators, $\mathcal{V}_m$ cannot range over $\mathbb{R}^p_{>0}$. It is a strict subset of $\mathbb{R}^p_{>0}$. This arises the question that what vectors in $\mathbb{R}^p_{>0}$ have preimages in $\mathcal{T}_m$.

LEMMA 43. *For two trees* $T_1, T_2 \in \mathcal{T}_m$*,* $\varphi(T_1) + \varphi(T_2) \in \mathcal{V}_m$ *if and only if* T_1 *and* T_2 *are compatible.*

Proof. The proof is provided in appendix. □

This statement suggests that, though each phylogenetic tree has its corresponding vector in $\mathbb{R}^p_{>0}$, not the vector yielded from the calculation of trees has its associated phylogenetic tree.

Example 44. There are three trees T_1, T_2 and T_3 in $\mathcal{T}_6$ as shown by Figure 3.14 (a). T_2 has a polytomy and T_2 is compatible with both T_1 and T_3, while T_1 and T_3 are incompatible. Their corresponding vectors are

$$\begin{aligned}
\varphi(T_1) &= (0.20, 0.45, 0.50, 0.60, 0.70, 0.45, 0.50, 0.60, 0.70, 0.45, 0.55, 0.65, 0.40, 0.50, 0.30)' \\
\varphi(T_2) &= (0.20, 0.30, 0.30, 0.40, 0.40, 0.30, 0.30, 0.40, 0.40, 0.20, 0.30, 0.30, 0.30, 0.30, 0.20)' \\
\varphi(T_3) &= (0.25, 0.40, 0.40, 0.60, 0.60, 0.45, 0.45, 0.65, 0.65, 0.20, 0.60, 0.60, 0.60, 0.60, 0.40)'
\end{aligned}$$

The additions are

$$\begin{aligned}
\varphi(T_1) + \varphi(T_2) &= (0.40, 0.75, 0.80, 1.00, 1.10, 0.75, 0.80, 1.00, 1.10, 0.65, 0.85, 0.95, 0.70, 0.80, 0.50)' \\
\varphi(T_2) + \varphi(T_3) &= (0.45, 0.70, 0.70, 1.00, 1.00, 0.75, 0.75, 1.05, 1.05, 0.40, 0.90, 0.90, 0.90, 0.90, 0.60)' \\
\varphi(T_1) + \varphi(T_3) &= (0.45, 0.85, 0.90, 1.20, 1.30, 0.90, 0.95, 1.25, 1.35, 0.65, 1.15, 1.25, 1.00, 1.10, 0.70)'
\end{aligned}$$

To project vector back to tree, the first step is to identify its topology. Assume T_1 is in the compatible tree class $\mathcal{T}^{(k)}$, then $\varphi(T_1) + \varphi(T_2)$ is a vector in the compatible vector set

$\mathcal{V}^{(k)}$ which can be characterized by $M^{(k)}$. The tree associated with $\varphi(T_1) + \varphi(T_2)$ has the topology as $\tau^{(k)}$ and has the branch length vector as $ginv(M^{(k)}) \cdot (\varphi(T_1) + \varphi(T_2))$, where $ginv(M^{(k)})$ is the left generalized inverse of $M^{(k)}$. This tree is shown as the tree $\tilde{T}_1$ in Figure 3.14 (b).

Likewise, $\tilde{T}_3$ is the tree associated with the vector $\varphi(T_2) + \varphi(T_3)$.

For $\varphi(T_1) + \varphi(T_3)$, note that T_1 and T_3 differ in the quartet subtree (S_1, S_3, S_4, S_6). The pairwise paths in $\varphi(T_1)+\varphi(T_3)$ corresponding to this quartet is $(0.85, 0.9, 1.3, 0.65, 1.25, 1.1)$. None of the min-plus inequalities is valid for this quartet. Therefore, $\varphi(T_1) + \varphi(T_3) \notin \mathcal{V}_m$.

3.4.3 Geometry of $\mathcal{V}_m$

As a supplementary to the structure of $\mathcal{V}_m$, we will explore the geometry of $\mathcal{V}_m$. Though $\mathcal{V}_4$ can be visualized as in Figure 3.10, the graphical visualization of the topological vector space corresponding to large trees is infeasible. We want a more systematic way to describe the structure of $\mathcal{V}_4$. The fact that vector $\varphi(T)$ is related to min-plus inequalities, and the mapping φ is a piecewise linear transformation, stimulates us to explore the structure of $\mathcal{V}_4$ in the literature of tropical geometry.

Study of Speyer and Sturmfels (2004) shows that in tropical algebraic geometry, the structure of BHV tree space $\mathcal{T}_m$ is a tropical Grassmannian $\mathcal{G}_{2,m}$, which is a polyhedral complex in $\mathbb{R}^p$. This polyhedral complex in $\mathbb{R}^p$ has $2^{m-1} - m - 1$ vertices and $(2m-5)!!$ facets. Each facet (also referred to as top dimensional stratum in Willis (2017)) has the same dimension of $(2m-3)$.

From the Figure 3.10 which shows the trivial situation of $\mathcal{V}_4$, the topological vector space $\mathcal{V}_4$ is a polyhedral fan with three five-dimensional cones $\mathbb{R}^4_{>0} \oplus \mathbb{R}_{\geq 0}$ glued along $\mathbb{R}^4_{>0}$. As general case, the geometry of $\mathcal{V}_m$ can be detected by the matrices. The $(2m-5)!!$ nonnegative orthants in column space of $M^{(k)}$ constitute $\mathcal{V}_m$, and there are $2^{m-1} - m - 1$ possible split indicators in $\{M^{(k)}_{INT}, k = 1, 2, \ldots, (2m-5)!!\}$. Therefore, when $m \geq 5$, the $\mathcal{V}_m$ is a polyhedral complex in $\mathbb{R}^p_{>0}$ with $2^{m-1} - m - 1$ vertices and $(2m-5)!!$ facets, and each facet $\mathcal{V}^{(k)}_m$ has

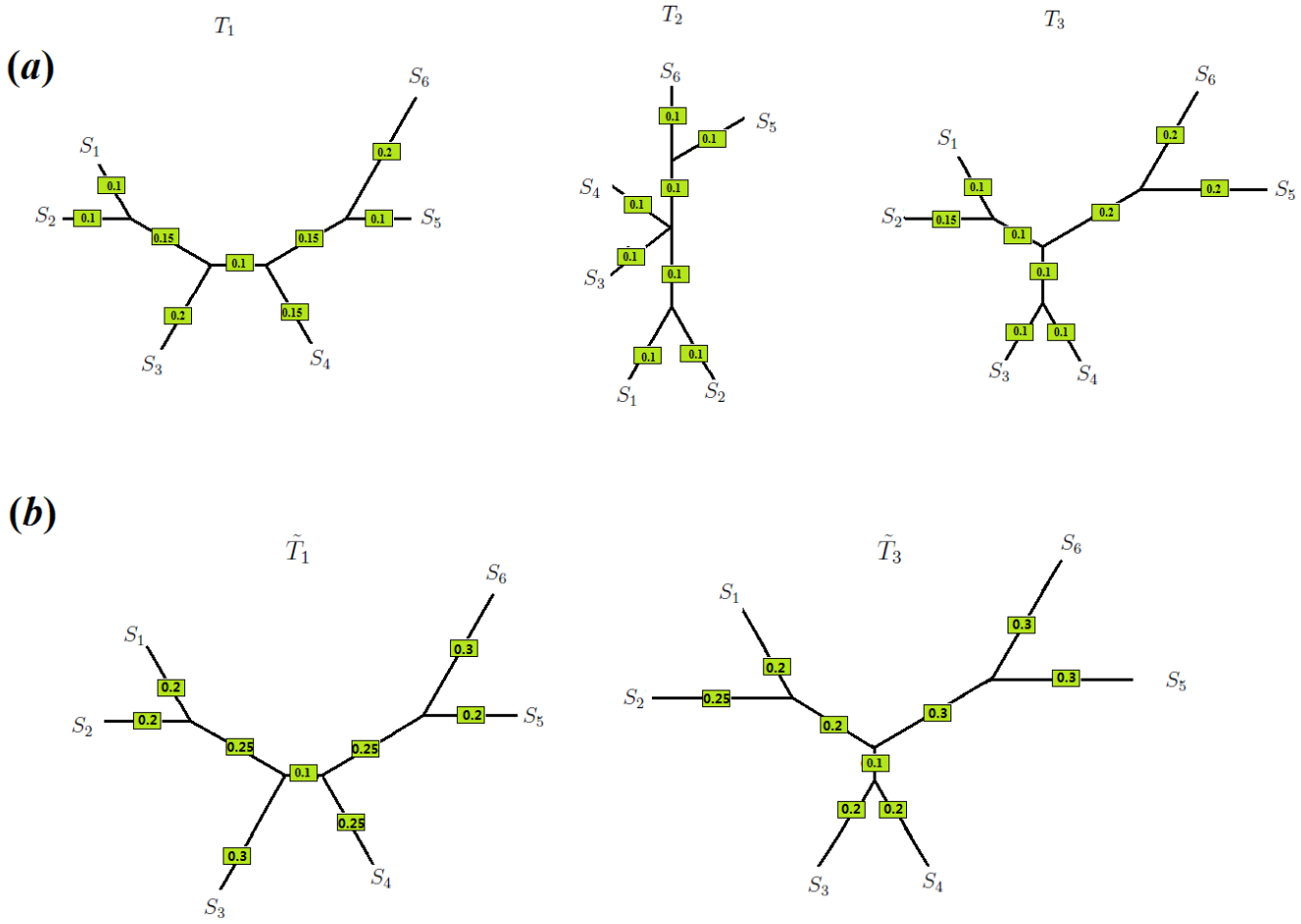

Figure 3.14: An example of manipulating tree with 6 taxa by vectors

the dimension of $(2m-3)$. Two facets $\mathcal{V}_m^{(k)}$ and $\mathcal{V}_m^{(h)}$ are glued along the set of vectors corresponding to the strict consensus tree of $\tau^{(k)}$ and $\tau^{(h)}$.

3.4.4 $(\mathcal{T}_m, D_{L2})$ and $(\mathcal{V}_m, D)$ are Polish Space

To conduct probabilistic and statistical analysis in the $L2$ tree space $(\mathcal{T}_m, D_{L2})$ as well as the topological vector space $(\mathcal{V}_m, D)$, in this part, we point out that they are Polish space, which is separable completely metrizable. Since $(\mathcal{T}_m, D_{L2})$ and $(\mathcal{V}_m, D)$ are isometry, for simplicity, we will show that $(\mathcal{V}_m, D)$ is complete and separable.

LEMMA 45. *A topological vector space $(\mathcal{V}_m, D)$ is complete.*

Proof. The proof is provided in the appendix. □

$\mathcal{V}_m$ is a subset in Euclidean vector space, therefore it is separable. The establishment that $(\mathcal{T}_m, D_{L2})$ and $(\mathcal{V}_m, D)$ allows the probability measures under this setting. In next chapter, we will show the application of this vectorization setting from a statistical perspective.

Chapter 4

Mean and Variance of Phylogenetic Trees

Because the tree is a graph-like object, it is hard to define computation on trees. This poses an impediment to the statistical inference on a collection of trees. Under the settings constructed in the last chapter, we can manipulate the corresponding vectors instead of phylogenetic trees. In this chapter, several classical statistical works will be explored, including the definition of mean and variance, and the proposal of an estimation method for the mean tree. For simplicity, we will omit the number of taxa in the notation, but note that the tree space $\mathcal{T}$ is built on a given set of taxa.

4.1 Centroid and Variability Measure

In Chapter 3, we build a pairwise path distance D_{L2} as a metric in the tree space $\mathcal{T}$. As aforementioned, the possibility that there may be multiple gene trees from an underlying species tree makes the phylogenetic tree can be viewed as a random variable with a distribution. In order to study the random phylogenetic tree from a statistical perspective, we aim to set a statistical infrastructure for the phylogenetic tree in the metric tree space $(\mathcal{T}, D_{L2})$.

Assume there's a tree distribution $T \sim F$ in the tree space $(\mathcal{T}, D_{L2})$. The mean of distribution F is $\int_{T\in\mathcal{T}} T dF(T)$. If the tree population resides in a compatible tree class $\mathcal{T}^{(k)}$, then the mean tree can be naturally defined as having the topology $\tau^{(k)}$ and with the mean branch lengths. If we consider the trees with conflicting topologies, the integration $\int_{T\in\mathcal{T}} T dF(T)$ is not intuitive. Fortunately, based on the relationship between trees and vectors defined in Chapter 3, we can define a mean vector for the tree distribution.

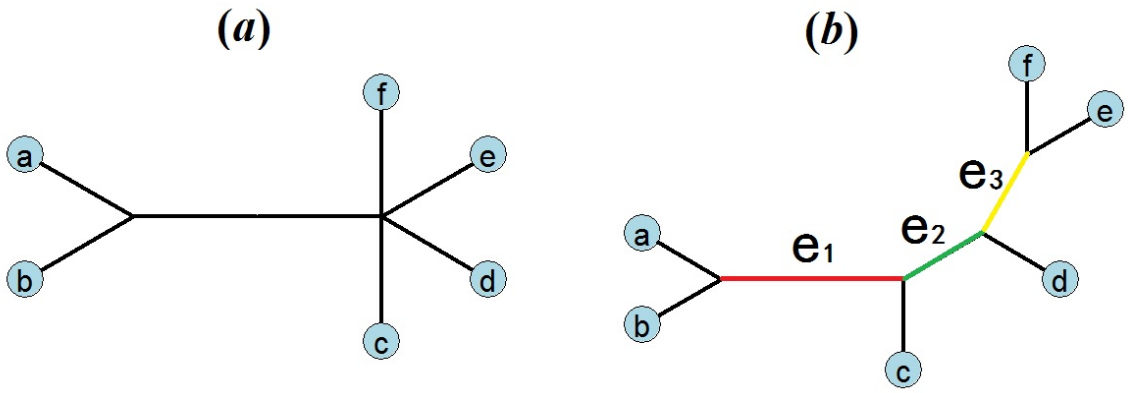

Figure 4.1: An example of a multifurcating tree as a special case of binary tree

Because there are finite possible topologies in $\mathcal{T}$, so the integration in $\mathcal{V}$ can be divided to $\mathcal{V}^{(k)}$. Note that a multifurcating tree is compatible with at least three distinct binary topologies. For simplification as well as avoiding double-counting, we treat a multifurcating tree as a special case of the binary tree, in a way by adding the internal edges subsequently to the lexicographically ordered tips within the polytomy. For example, the trees in Figure 4.1 (a) is compatible with 15 distinct binary topologies. In computation, it is treated as a special case of the topology in Figure 4.1(b), which is obtained by adding internal edges connected to tip c and d. The multifurcating tree in Figure 4.1(a) will be calculated as it is in the compatible tree class yielded from the tree in Figure 4.1(b).

4.1.1 Proposal of Mean and Variance

The mean of a random tree is hard to define. Chapter 3 constructs a setting that relates each phylogenetic tree to a vector. This representation encourages us to define the mean of the corresponding random vector instead. Since the tree space $(\mathcal{T}, D_{L2})$ and topological vector space $(\mathcal{V}, D)$ are Polish space, therefore we can define the probability triples in them.

DEFINITION 46. (mean vector of F) Given a probability space $(\mathcal{T}, \mathcal{B}(\mathcal{T}), P_F)$ for a random phylogenetic tree, it has an associated probability triple in the isometric topological vector space as $(\mathcal{V}, \mathcal{B}(\mathcal{V}), P_F)$. Define its **mean vector** as

$$V_F = E_F(\varphi(T)) = \int_{T\in\mathcal{T}} \varphi(T)dF(T) = \int_{\varphi(T)\in\mathcal{V}} \varphi(T)dF(T)$$

Since $\mathcal{V} = \bigcup_{k=1}^{(2m-5)!!} \mathcal{V}^{(k)}$, the integration of $\varphi(T)$ is actually the sum of integrations on $\mathcal{V}^{(k)}$. Therefore,

$$\begin{aligned} V_F &= \sum_{k=1}^{(2m-5)!!} \textstyle\int_{T\in\mathcal{T}^{(k)}} \varphi(T)P(T\in\mathcal{T}^{(k)})dF(T|\mathcal{T}^{(k)}) \\ &= \sum_{k=1}^{(2m-5)!!} P(T\in\mathcal{T}^{(k)}) \textstyle\int_{T\in\mathcal{T}^{(k)}} \varphi(T)dF(T|\mathcal{T}^{(k)}) \\ &= \sum_{k=1}^{(2m-5)!!} E_F(\varphi(T)|\mathcal{T}^{(k)})P(T\in\mathcal{T}^{(k)}) \\ &\doteq \sum_{k=1}^{(2m-5)!!} V_{F|\mathcal{T}^{(k)}} P(T\in\mathcal{T}^{(k)}) \end{aligned}$$

V_F is the weighted average of the conditional mean vectors from the compatible tree classes. As an integration, $V_F \in \mathbb{R}^p_{>0}$ exists if and only if $\int_{\mathcal{T}} \varphi(T)dF(T) < \infty$. However, LEMMA 43 points out that V_F is not necessarily in $\mathcal{V}$. In other words, there may not exist a tree in $\mathcal{T}$ such that $\varphi(T) = V_F$. In general, a natural way is to define the tree whose corresponding vector is "closest" to V_F as the optimal tree associated with V_F.

DEFINITION 47. (mean tree of F) For a tree distribution F with mean vector V_F, its **mean tree** is defined as

$$T_F = \arg\min_{T\in\mathcal{T}} \|\varphi(T) - V_F\|_2$$

Given this definition, next we will discuss the existence and uniqueness of T_F. Since $\mathcal{V}$ is complete, therefore, if V_F exists, then T_F exists.

For the uniqueness of T_F, note that $\|\varphi(T) - V_F\|_2$ achieves 0 if and only if $\varphi(T) = V_F$. Therefore, if $V_F \in \mathcal{V}$, then it is guaranteed that T_F is unique.

LEMMA 48. *Given that the mean tree T_F for a random tree $T \sim F$ exists, then*

1. *If T distributes within a tree class, then its mean tree is unique and is in the same tree class.*

2. *If the mean vector $V_F \in \mathcal{V}$, then the mean tree is unique as $\varphi^{(-1)}(V_F)$.*

3. *In each tree class, there is at most one tree minimizing the distance to V_F.*

Proof. Here we give a brief proof to the last statement. If in T_1 and T_2 are two distinct trees in $\mathcal{T}^{(i)}$ who minimize the distance to V_F. $\|\varphi(T_1) - V_F\|_2^2 = \|\varphi(T_2) - V_F\|_2^2 = \delta$. Let $V_0 = (V_1 + V_2)/2$, then $\varphi^{-1}(V_0) \in \mathcal{T}^{(i)}$ and $\|V_0 - V_F\|_2^2 < \delta$. This is a contradiction to the assumption that both T_1 and T_2 are minimizers. □

Generally the uniqueness of T_F can not be guaranteed.

Example 49. Assume a random tree has the same probability of being either of the two trees in Figure 4.2. Then the mean vector $V_F = (3, 3, 4, 4, 3, 3)'$. To find the minimizer to V_F, search in three compatible vector classes $\mathcal{V}^{(i)}$ that constitute $\mathcal{V}_4$. Two trees are selected because they give the same minimal distance to V_F. One tree is compatible with the first tree, and has the corresponding vector as $V_1 = (3, 3.5, 3.5, 3.5, 3.5, 3)'$. The other tree is compatible with the second tree, and has the corresponding vector as $V_2 = (3.5, 3, 3.5, 3.5, 3, 3.5)'$.

In this example, there are two minimizers because the two distinct tree classes have the same weight on the tree distribution, so the minimization gives symmetric results in the two tree classes. If the tree is inclined to a particular tree class, or say, if there is a compatible vector set $\mathcal{V}^{(i)}$ such that the distance of the mean vector V_F to $\mathcal{V}^{(i)}$ is less than the distance to any other compatible vector set, then the minimizer is unique.

PROPOSITION 50. *Given that the mean tree T_F for a random tree $T \sim F$ exits, if there is k_0, such that $\min_{V \in \mathcal{V}^{(k_0)}} \|V - V_F\|_2^2 < \min_{V \in \mathcal{V}^{(j)}} \|V - V_F\|_2^2$ for any $j \neq k_0$, then T_F is unique and is in $\mathcal{T}^{(k_0)}$.*

The definition of variability in a set of trees is even more troublesome than the centroid. The variability measure should capture both the direction and latitude of the tree difference.

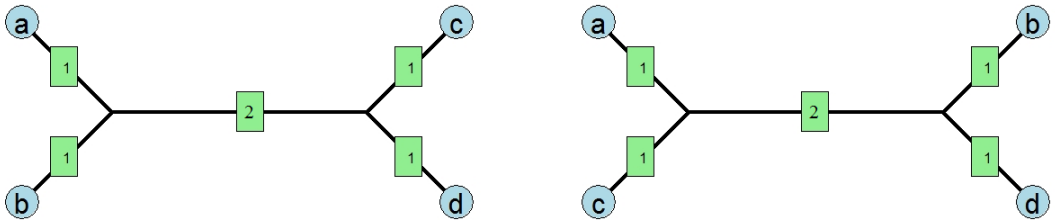

Figure 4.2: An example of 4-taxa tree distribution

Traditional tree metrics such as Robinson-Foulds distance can present the latitude; however, the direction of tree difference is hard to describe. Given the relationship between trees and vectors constructed by the mapping φ, our intuition is to characterize the variability in a set of trees by the variability in the corresponding set of vectors. Therefore, similar to the definition of mean vector for a tree distribution, based on the tree's vectorization, we can define a variance of a phylogenetic tree.

DEFINITION 51. For a tree $T \sim F$ with mean vector V_F, its **covariance matrix** is defined as

$$\Sigma_F = E_F\left((\varphi(T) - V_F)(\varphi(T) - V_F)'\right) = \int_{\mathcal{T}} (\varphi(T) - V_F)(\varphi(T) - V_F)' dF(T)$$

Likewise,

$$\begin{aligned}
\Sigma_F &= \sum_{k=1}^{(2m-5)!!} P(T \in \mathcal{T}^{(k)}) \int_{T \in \mathcal{T}^{(k)}} (\varphi(T) - V_F)(\varphi(T) - V_F)' dF(T|\mathcal{T}^{(k)}) \\
&= \sum_{k=1}^{(2m-5)!!} P(T \in \mathcal{T}^{(k)}) MSE_{V_F}(\varphi(T)|\mathcal{T}^{(k)}) \\
&= \sum_{k=1}^{(2m-5)!!} P(T \in \mathcal{T}^{(k)}) \left[VAR_T(\varphi(T)|\mathcal{T}^{(k)}) + (E_T(\varphi(T|\mathcal{T}^{(k)})) - V_F)(E_T(\varphi(T|\mathcal{T}^{(k)})) - V_F)'\right] \\
&\doteq \sum_{k=1}^{(2m-5)!!} P(\mathcal{T}^{(k)}) \left[VAR_T(\varphi(T)|\mathcal{T}^{(k)}) + (V_{F|\mathcal{T}^{(k)}} - V_F)(V_{F|\mathcal{T}^{(k)}} - V_F)'\right]
\end{aligned}$$

Σ_F is the weighted average of the conditional variance and the squared bias of conditional mean vector to V_F from each compatible tree class. It measures the variability in $\varphi(T)$ where $T \sim F$.

4.1.2 Sample Mean Tree

For a collection of trees $\{T_1, T_2, \ldots, T_n\}$, its average tree is not intuitive. However, by the vectorization mapping φ, we can propose a sample mean tree based on the average of their corresponding vectors. Given a sample of trees, similar to the definition of mean tree, we can first get their average vector, and then look for the tree that optimally "matches" the average vector.

DEFINITION 52. (sample mean tree) Given a random sample of trees $\{T_1, T_2, \ldots, T_n\}$ drawn from distribution F , the **sample mean vector** is defined as the mean of the sample trees' corresponding vectors.

$$\overline{V}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(T_i)$$

The **sample mean tree** is defined as the tree closest to the sample mean vector.

$$\overline{T}_n = \arg\min_{T \in \mathcal{T}} \left\| \varphi(T) - \overline{V}_n \right\|_2$$

The sample mean tree is defined as the tree whose corresponding vector is closest to the sample mean vector in the L_2 norm. Since the tree metric D_{L2} is defined from the L_2 norm in the vector space, hence the sample mean tree is the tree that minimizing the pairwise path distance to the sample trees. Therefore, the sample mean tree is the Fréchet mean (Nye, 2011) of the sample trees. This argument can be proved.

LEMMA 53. *$\overline{T}_n$ is the sample Fréchet mean of $\{T_1, T_2, \ldots, T_n\}$ in the metric space $(\mathcal{T}, D_{L2})$.*

Proof. Denote $V_T = \varphi(T) = (v_{kl}^{(T)})_{kl}$, $V_i = \varphi(T_i) = (v_{kl}^{(i)})_{kl}$, and $\overline{V}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(T_i) = (\overline{v}_{kl})_{kl}$.

$$\sum_{i=1}^{n} \|V_T - V_i\|_2^2 = \sum_{i=1}^{n} \left\| V_i - \overline{V}_n + \overline{V}_n - V_T \right\|_2^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \sum_{k<l} (v_{kl}^{(i)} - \overline{v}_{kl} + \overline{v}_{kl} - v_{kl}^{(T)})^2 \\
&= \sum_{i=1}^{n} \sum_{k<l} \left((v_{kl}^{(i)} - \overline{v}_{kl})^2 + (\overline{v}_{kl} - v_{kl}^{(T)})^2 + 2(\overline{v}_{kl} - v_{kl}^{(T)})(v_{kl}^{(i)} - \overline{v}_{kl}) \right) \\
&= \sum_{i=1}^{n} \left\| V_i - \overline{V}_n \right\|_2^2 + \sum_{i=1}^{n} \left\| V_t - \overline{V}_n \right\|_2^2 + 2 \sum_{k<l} (\overline{v}_{kl} - v_{kl}^{(t)}) \sum_{i=1}^{n} (v_{kl}^{(i)} - \overline{v}_{kl}) \\
&= \sum_{i=1}^{n} \left\| V_i - \overline{V}_n \right\|_2^2 + \sum_{i=1}^{n} \left\| V_t - \overline{V}_n \right\|_2^2
\end{aligned}
$$

Given $\{T_1, \ldots, T_n\}$, $\sum_{i=1}^{n} \left\| V_i - \overline{V}_n \right\|_2^2$ is a fixed value. Thus $\min \sum_{i=1}^{n} \|V_T - V_i\|_2^2 \Leftrightarrow \min \sum_{i=1}^{n} \left\| V_t - \overline{V}_n \right\|_2^2$. $\overline{T}_n = \arg\min_{T \in \mathcal{T}} \sum_{i=1}^{n} D_{L2}^2(T, T_i)$ □

In this part, we define the mean tree and propose a sample mean tree based on the vectorization of trees. Next, we will show that the sample mean tree is an excellent estimator for the mean tree.

4.2 Asymptotic Behavior of Sample Mean Tree

The mean tree T_F cannot be guaranteed to be unique in general cases. The LEMMA 43 also indicates that for a given sample $\{T_1, T_2, \ldots, T_n\}$, the tree minimizing the pairwise path distance may not be unique. From LEMMA 48, if the tree distributes within a tree class, the minimizer is guaranteed to be unique.

LEMMA 54. *If the random tree T distributes within a tree class $\mathcal{T}^{(i)}$ and its mean vector exists, then its unique sample mean tree is unbiased estimator (UE) for its unique mean tree.*

Proof. The population of T reside in $\mathcal{T}^{(i)}$, thus its mean vector V_F and sample mean vector $\overline{V}_n$ are in $\mathcal{V}^{(i)}$. Hence, both the mean tree and sample mean tree are unique. $T_F = \varphi^{-1}(V_F)$ and $\overline{T}_n = \varphi^{-1}(\overline{V}_n)$. For $\overline{T}_n$ as a random tree, its mean vector is $E(\varphi(\overline{T}_n)) = E(\overline{V}_n) = E(\frac{1}{n} \sum_{i=1}^{n} \varphi(T_i)) = \frac{1}{n} \sum_{i=1}^{n} E(\varphi(T_i)) = E(\varphi(T)) = V_F$. Therefore, the mean of $\overline{T}_n$ is $\varphi^{-1}(V_F) = T_F$. □

In this literature, the unbiasedness of tree estimator means its corresponding vector is unbiasedness. From the proof, $\varphi(\overline{T}_n) = \overline{V}_n$ is the sufficient condition for unbiasedness .

PROPOSITION 55. *$\overline{T}_n$ is UE for T_F if and only if the sample mean vector $\overline{V}_n \in \mathcal{V}$.*

If there are conflicting topologies in tree population, there is no guarantee that $\overline{V}_n \in \mathcal{V}$. Hence $\overline{T}_n$ is not necessarily to be unbiased. Nevertheless, $\overline{T}_n$ has the following satisfying property of consistency.

THEOREM 56. *(Consistency of Sample Mean Tree) For a random sample generated from $T \sim F$, if the mean tree T_F uniquely exists, then the sample mean tree $\overline{T}_n \in \mathcal{T}^{(k)}$ is a consistent estimator for T_F in $(\mathcal{T}, D_{L2})$.*

Proof. For a random sample $\{T_1, T_2, \ldots, T_n\}$ generated from F, by the strong Law of Large Numbers (L.L.N.), $\lim_{n\to\infty} \overline{V}_n = V_F$ almost surely. Also, we have

$$\begin{aligned} \forall T \in \mathcal{T}, \|\varphi(T_F) - V_F\|_2 &\leq \|\varphi(T) - V_F\|_2 \\ \forall T \in \mathcal{T}, \forall n, \left\|\varphi(\overline{T}_n) - \overline{V}_n\right\|_2 &\leq \left\|\varphi(T) - \overline{V}_n\right\|_2 \end{aligned}$$

Take limits on both sides of the second inequality and set T as $\overline{T}_n$ in the first inequality, then

$$\begin{aligned} \forall n, \|\varphi(T_F) - V_F\|_2 &\leq \left\|\varphi(\overline{T}_n) - V_F\right\|_2 \\ \forall T \in \mathcal{T}, \left\|\lim_{n\to\infty} \varphi(\overline{T}_n) - V_F\right\|_2 &\leq \|\varphi(T) - V_F\|_2 \text{ almost surely} \end{aligned}$$

Take limits on both sides of the first inequality and set T as T_F in the second inequality, then

$$\begin{aligned} \|\varphi(T_F) - V_F\|_2 &\leq \left\|\lim_{n\to\infty} \varphi(\overline{T}_n) - V_F\right\|_2 \text{ almost surely} \\ \left\|\lim_{n\to\infty} \varphi(\overline{T}_n) - V_F\right\|_2 &\leq \|\varphi(T_F) - V_F\|_2 \text{ almost surely} \end{aligned}$$

Thus, we have

$$\|\varphi(T_F) - V_F\|_2 = \left\|\lim_{n\to\infty} \varphi(\overline{T}_n) - V_F\right\|_2 \text{ almost surely}$$

Since T_F uniquely exists, we can assume $T_F \in \mathcal{T}^{(k)}$. We assert that there exists $N \in \mathbb{N}$ such that $\forall n > N$, $\overline{T}_n$ is in $\mathcal{T}^{(k)}$.

Otherwise, since there are finite possible tree classes, there is at least one tree class, other than $\mathcal{T}^{(k)}$, that will contain infinite $\overline{T}_n$ for $n \geq 1$. Assume that tree class is $\mathcal{T}^{(h)}$ $(h \neq k)$. There is a subsequent $\overline{T}_{n_h}$ such that $\overline{T}_{n_h} \in \mathcal{T}^{(h)}$ and

$$\left\| \lim_{n_h \to \infty} \varphi(\overline{T}_{n_h}) - V_F \right\|_2 = \|\varphi(T_F) - V_F\|_2 \text{ almost surely}$$

Therefore, there exists a $T_F^{(h)} \in \mathcal{T}^{(h)}$ such that $\lim_{n_h \to \infty} \varphi(\overline{T}_{n_h}) = \varphi(T_F^{(h)})$, i.e.,

$$\left\| \varphi(T_F^{(h)}) - V_F \right\|_2 = \|\varphi(T_F) - V_F\|_2$$

$T_F^{(h)} \in \mathcal{T}^{(h)}$ is the mean tree. This contradicts with the assumption that the mean tree is unique.

There exists $N \in \mathbb{N}$ such that $\forall n > N$, $\overline{T}_n$ is in $\mathcal{T}^{(k)}$. Thus, $\lim_{n \to \infty} \varphi(\overline{T}_n) = T_F$ almost surely. That is $D_{L2}(\overline{T}_n, T_F) \to 0$. Hence, the sample mean tree converges to the mean tree. □

Moreover, the uniqueness of $\overline{T}_n$ cannot be guaranteed unless the sample mean vector $\overline{V}_n$ is "inclined" to a particular tree class. If the mean tree T_F exists but is not unique, the sample mean tree $\overline{T}_n$ may not converge in $(\mathcal{T}, D_{L2})$. However, from the above inference, if we focus on $T_F \in \mathcal{T}^{(k)}$ and restrict the sample mean tree sequence to the subsequence in the tree class $\mathcal{T}^{(k)}$, then the subsequence of $\overline{T}_n$ in the tree class $\mathcal{T}^{(k)}$ converges to the mean tree T_F in $\mathcal{T}^{(k)}$.

Since it is assumed that $\{T_1, T_2, \ldots, T_n\} \overset{iid}{\sim} F$, then their corresponding vectors $\{\varphi(T_1), \varphi(T_2), \ldots, \varphi(T_n)\}$ are also i.i.d., with the mean vector $V_F = E(\varphi(T))$ and covariance matrix $\Sigma_F = VAR(\varphi(T))$. According to the multidimensional Central Limit Theorem (C.L.T.), there is

$$\sqrt{n}(\overline{V}_n - V_F) \xrightarrow{\boldsymbol{d}} \mathcal{N}_p(\mathbf{0}, \Sigma_F)$$

Based on the isomorphism between trees and vectors, we attempt to extend this statement to tree space.

THEOREM 57 (Asymptotic Normality of Sample Mean Tree)**.** *For the random sample* $\{T_1, T_2, \ldots, T_n\}$ *generated from* $T \sim F$*, if the mean tree* T_F *uniquely exists, then*

$$\sqrt{n}(\varphi(\overline{T}_n) - \varphi(T_F)) \xrightarrow{D} \mathcal{N}_p(\mathbf{0}, \tilde{\Sigma}_F)$$

Proof. There exists N such that $\forall n > N$, $\overline{T}_n$ reside in the same tree class with T_F. Without loss of generality, assume they are in the tree class $\mathcal{T}^{(k)}$ which is characterized by $M^{(k)}$: $p \times (2m-3)$. Then $\mathcal{V}^{(k)}$ is the semi-positive orthant in the column space of $M^{(k)}$.

For $n > N$, $\overline{T}_n = \arg\min_T \left\|\varphi(T) - \overline{V}_n\right\|_2 \in \mathcal{T}^{(k)}$, thus $\overline{T}_n = \arg\min_{\varphi(T)=M^{(k)}\boldsymbol{b}} \left\|M^{(k)}\boldsymbol{b} - \overline{V}_n\right\|_2$. Note $M^{(k)}$ is full column rank, assume $ginv(M^{(k)}) : (2m-3) \times p$ is the left generalized inverse of $M^{(k)}$. Therefore, $\varphi(\overline{T}_n) = M^{(k)} ginv(M^{(k)}) \overline{V}_n \doteq G \cdot \overline{V}_n$, where $G = M^{(k)} \cdot ginv(M^{(k)})$.

Note that $\varphi(T_F)$ is the projection of V_F on $\mathcal{V}^{(k)}$, thus $\varphi(T_F) = G \cdot V_F$. Explicitly, $\varphi(T_F) = M^{(k)} \cdot \boldsymbol{b}(T_F) = M^{(k)} \cdot ginv(M^{(k)}) V_F = G \cdot V_F$.

According to $\sqrt{n}(\overline{V}_n - V_F) \xrightarrow{\boldsymbol{d}} \mathcal{N}_p(\mathbf{0}, \Sigma_F)$, we have

$$\sqrt{n}(G \cdot \overline{V}_n - G \cdot V_F) \xrightarrow{\boldsymbol{d}} \mathcal{N}_p(\mathbf{0}, G\Sigma_F G')$$

That is

$$\sqrt{n}(\varphi(\overline{T}_n) - \varphi(T_F)) \xrightarrow{\boldsymbol{d}} \mathcal{N}_p(\mathbf{0}, \tilde{\Sigma}_F)$$

where $\tilde{\Sigma_F} = M^{(k)} \cdot ginv(M^{(k)}) \Sigma_F (ginv(M^{(k)}))' (M^{(k)})'$. □

This theorem validates the asymptotic normality of the vectors corresponding to the sample mean tree. Based on this infrastructure, a series of statistical inferences can be developed. For example, given a sample of trees $\{T_1, T_2, \ldots, T_n\} \overset{iid}{\sim} F$, we can build a confidence procedure as follows. First, calculate the sample mean tree $\overline{T}_n$. Then, the confidence region in the Euclidean space can be built as $\left\{x \in \mathbb{R}^p_{>0} | n \left(\varphi(\overline{T}_n) - x\right)' \tilde{\Sigma_F}^{-1} \left(\varphi(\overline{T}_n) - x\right) < \frac{p}{n-p} F_{p,n-p}(1-\alpha)\right\}$. Next,for an arbitrary tree T_0, we can determine if it is in the confidence set by identifying if $\varphi(T_0)$ is in the confidence region.

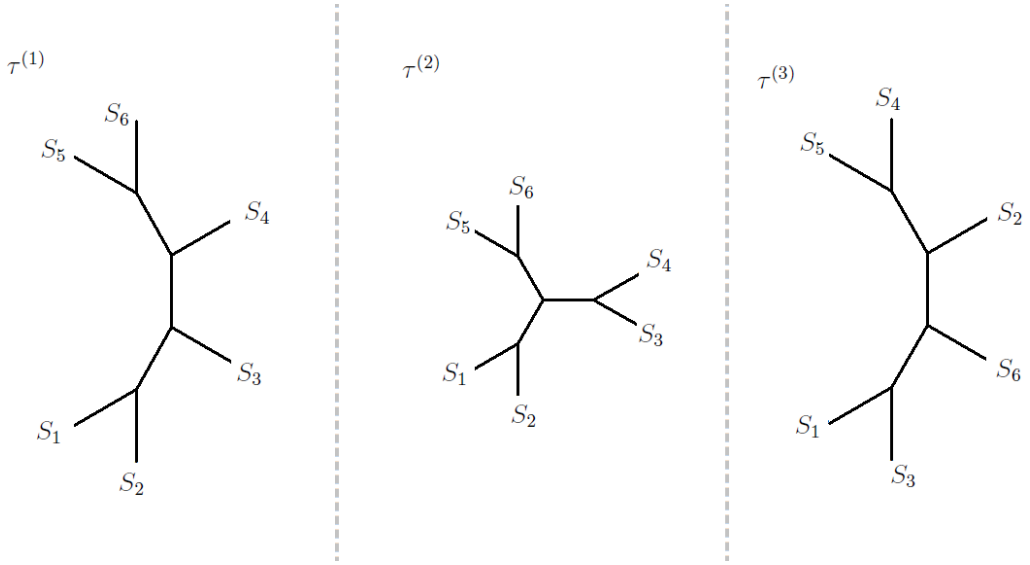

Figure 4.3: Potential topologies for the random sample tree

4.3 A Simulation Study

In this section, we will conduct a simulation to show the asymptotic performance of $\varphi(\overline{T}_n) - \varphi(T_F)$.

Step 1. Assume the tree distribution F.

In this simulation study, to show the asymptotic behavior, we use an arbitrary distribution. For a more complicated construction, we can assume a species tree and generate a sample of phylogenetic trees according to the distribution from the coalescent process (Rosenberg and Nordborg, 2002; Liu et al., 2015).

Our assumed distribution is that the tree has three possible topologies as shown by Figure 4.3. $P(T \in \mathcal{T}^{(1)}) = 1/2$, $P(T \in \mathcal{T}^{(2)}) = 1/4$, $P(T \in \mathcal{T}^{(3)}) = 1/4$. The 6 terminal edges are fixed to be with length 1. The three internal edges are assumed to be exponentially distributed as $exp(0.5)$ for each tree class.

It can be calculated that the mean vector is

$$V_F = (3.0, 4.0, 6.5, 7.5, 6.5, 5.0, 5.5, 6.5, 6.5, 4.5, 6.5, 5.5, 4.0, 5.0, 3.0)'$$

$V_F \notin \mathcal{V}_6$, while the mean tree T_F whose corresponding vector is closest to V_F is $T_F \in \mathcal{T}^{(1)}$ and has the edge lengths as

$$\boldsymbol{b}_{TER} = (1.625, 1.375, 1.750, 1.750, 1.625, 1.375)', \boldsymbol{b}_{INT} = (1.250, 1.333, 1.250)'$$

. The corresponding vector of the mean tree is

$$\varphi(T_F) = (3.000, 4.625, 5.958, 7.083, 6.833, 4.375, 5.708, 6.833, 6.583, 4.833, 5.958, 5.708, 4.625, 4.375, 3.000)'$$

Step 2. Generate $K = 100$ samples.

Each sample contains n ($n = 10, 50, 300$) trees randomly sampling from F. For each sample $\{T_1, T_2, \ldots, T_n\}$, calculate the corresponding vectors of each tree, and then get the sample mean vector $\overline{V}_n$. Next, reconstruct the sample mean tree $\overline{T}_n$ from the sample mean vector. The last step is to vectorize the sample mean tree $\overline{T}_n$ to get its corresponding vector $\varphi(\overline{T}_n)$.

Step 3. Calculate the differences between $\varphi(T_F)$ and $\varphi(\overline{T}_n)$ for 100 samples.

We have 100 replicates of $\varphi(\overline{T}_n) - \varphi(T_F)$ ($n = 10, 50, 300$), which are vectors of $\binom{6}{2} = 15$ entries.

To demonstrate the multidimensional normality of $\varphi(\overline{T}_n) - \varphi(T_F)$, we present the histograms of its 15 marginal distributions. In addition, we randomly generate a linear combination of $\varphi(T_F) - \varphi(\overline{T}_n)$ and show its distribution. Furthermore, we demonstrate the distribution of $\left\|\varphi(\overline{T}_n) - \varphi(T_F)\right\|_2^2$, which should be approximately Chi-squared.

From Figure 4.4, Figure 4.5 and Figure 4.6, the marginal distribution of $\varphi(\overline{T}_n) - \varphi(T_F)$ seems more and more normal as n increases.

Figure 4.7 shows the distribution of a linear combination $\boldsymbol{a}'(\varphi(\overline{T}_n) - \varphi(T_F))$ for a random 15-length vector $\boldsymbol{a}$. As n increases, the $\boldsymbol{a}'(\varphi(\overline{T}_n) - \varphi(T_F))$ seems to be more normally distributed.

Figure 4.8 shows the distribution of $\left\|\varphi(\overline{T}_n) - \varphi(T_F)\right\|_2^2$. As n increases, the distance between $\varphi(\overline{T}_n)$ and $\varphi(T_F)$ decreasingly approaches to the expectation of the Chi-square distribution.

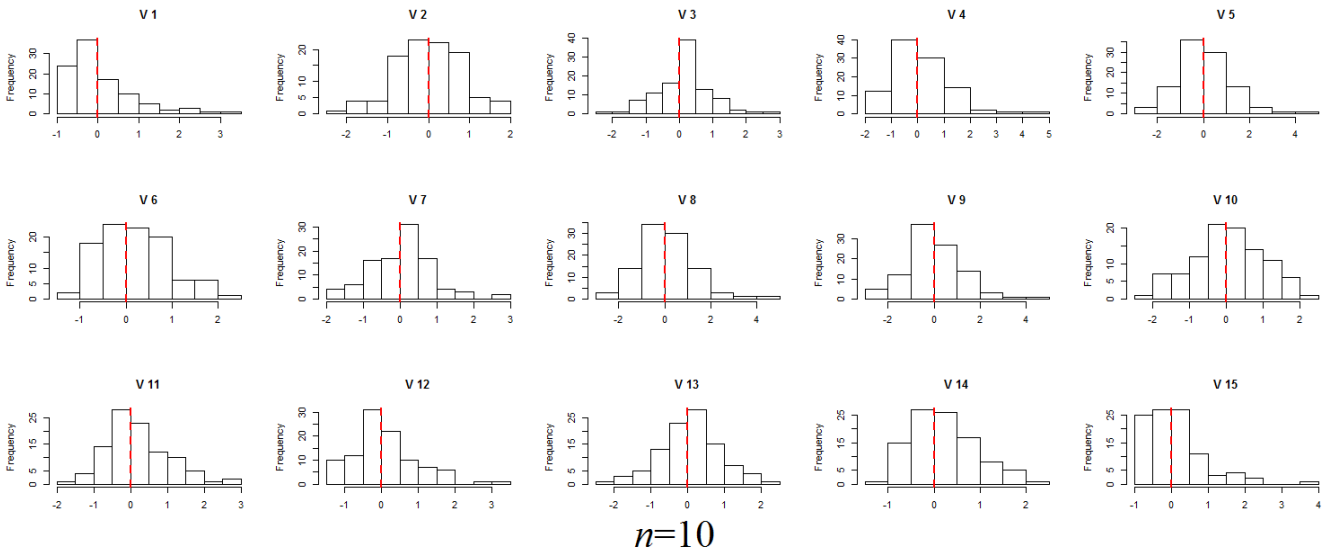

Figure 4.4: Histogram for the marginal distribution of $\varphi(\overline{T}_n) - \varphi(T_F)$ when $n = 10$

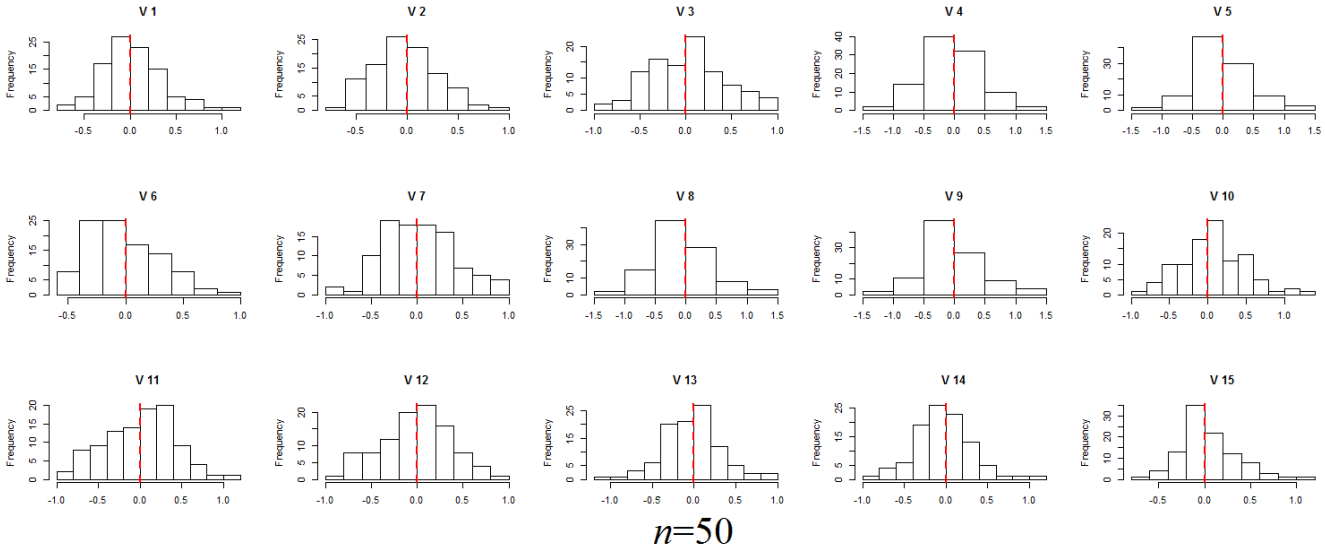

Figure 4.5: Histogram for the marginal distribution of $\varphi(\overline{T}_n) - \varphi(T_F)$ when $n = 50$

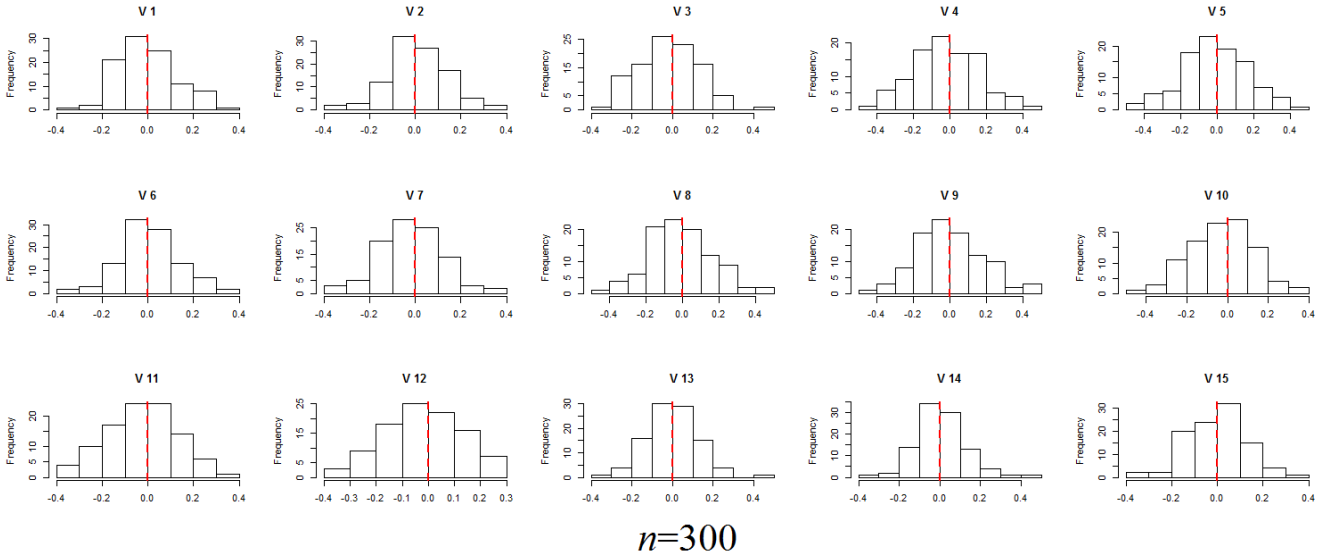

Figure 4.6: Histogram for the marginal distribution of $\varphi(\overline{T}_n) - \varphi(T_F)$ when $n = 300$

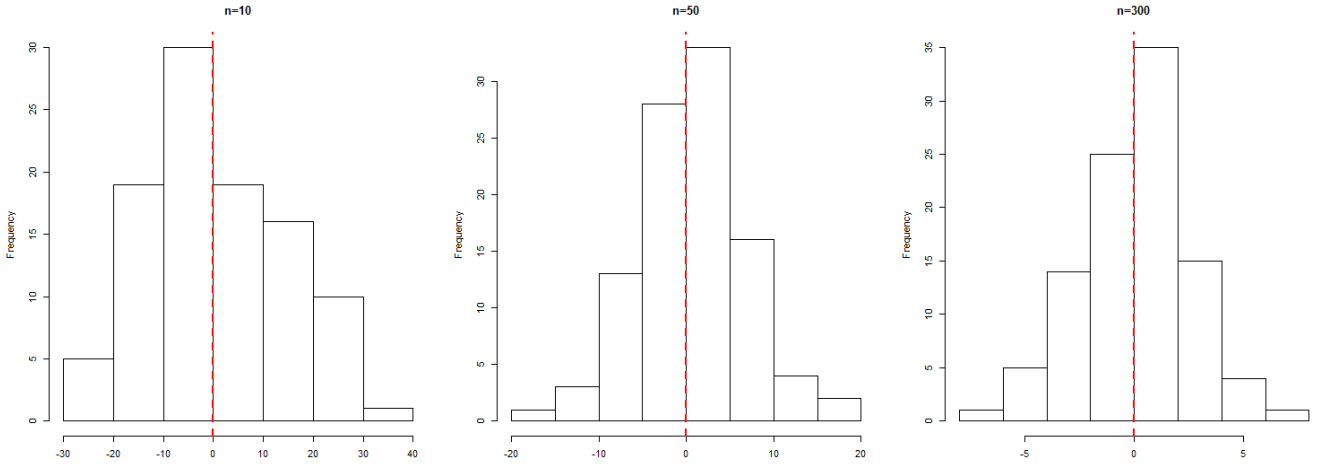

Figure 4.7: Histogram of a random linear combination of $\varphi(\overline{T}_n) - \varphi(T_F)$ when $n = 10, 50, 300$

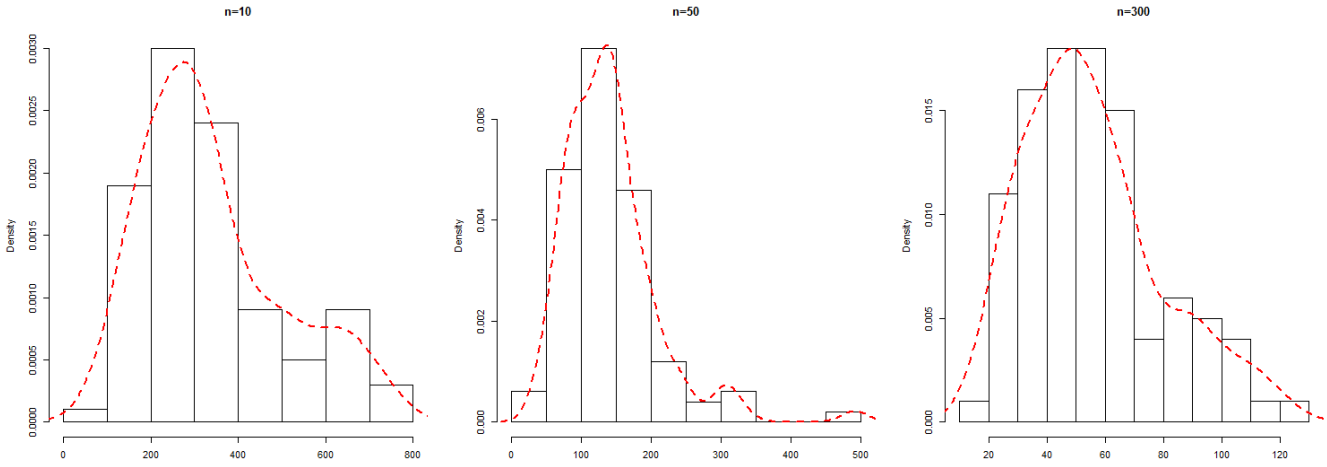

Figure 4.8: Histogram of the distance between $\varphi(\overline{T}_n)$ and $\varphi(T_F)$ when $n = 10, 50, 300$

4.4 A Real Case Study

In this section, we will use a real case to show how to build a confidence procedure in tree space, based on the vectorization mapping.

Step 1. The material collection. We use a portion of Multiz alignment block from the RefSeq Genes published on *genome.ucsc.edu*. The alignment block 5 of 102 in windows, 79935440-79935592 (153 bps) collected from (*Chicken, Cow, Dog, Rat*) is used.

Step 2. Use RAxML (Kozlov et al., 2019; Stamatakis, 2014) to generate 100 bootstrap phylogenetic trees $(T_1, T_2, \ldots, T_{100})$ from the alignment. The substitution model is assumed as GTR. The best tree built from RAxML is shown by Figure 4.9.

The 100 bootstrap trees distribute in the three tree classes in $\mathcal{T}_4$. Order the taxa lexicographically as *Chicken, Cow, Dog, Rat*. There are 7 trees with topology $\tau^{(1)}$, 5 trees with topology $\tau^{(2)}$ and 88 trees with topology $\tau^{(3)}$ as shown by Figure 4.10. The internal edge lengths are presented in Figure 4.11. The sizes of trees in $\mathcal{T}^{(1)}$ and in $\mathcal{T}^{(2)}$ are small, but it can be told that the lengths of internal edge for trees in $\mathcal{T}^{(3)}$ concentrate around 0.06.

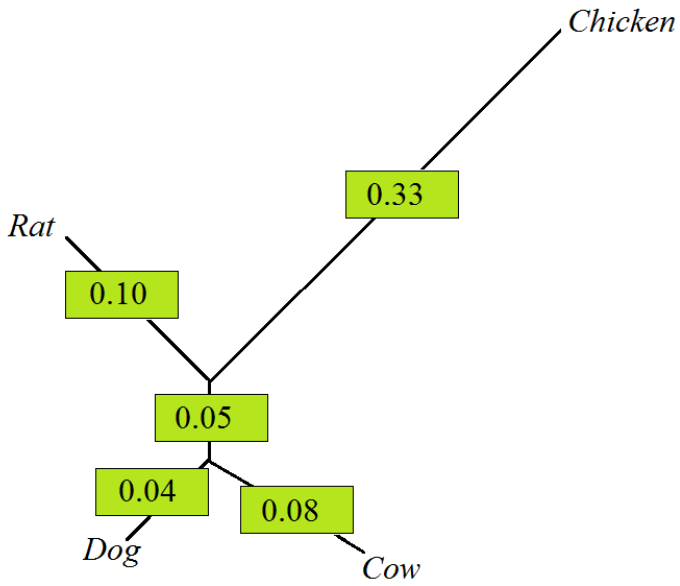

Figure 4.9: The maximum likelihood tree from RAxML built from the alignment of 4 species and 153 bps

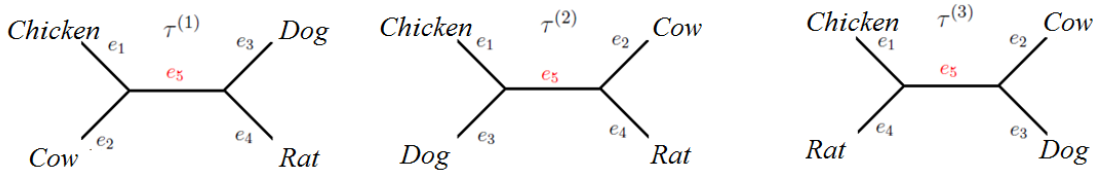

Figure 4.10: Three possible topologies for the 100 bootstrap trees

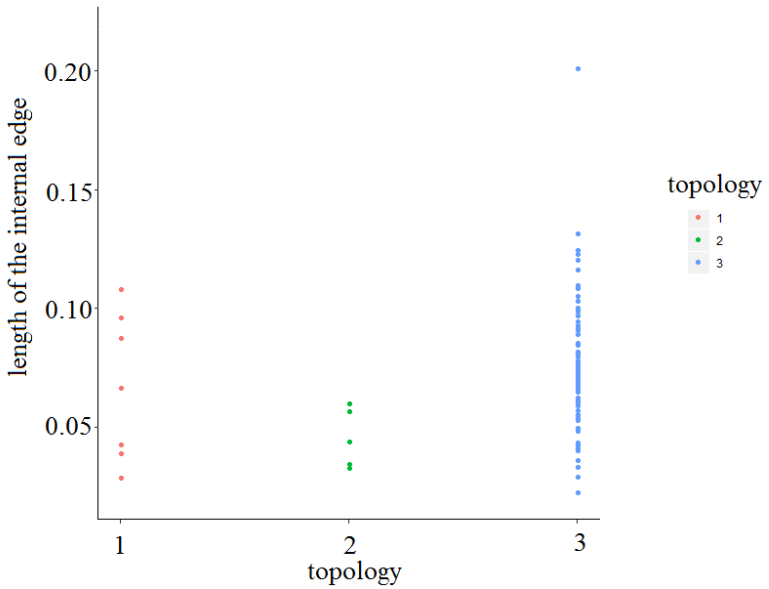

Figure 4.11: The internal edge lengths of the 100 bootstrap trees

Step 3. Calculate the statistics from the bootstrap trees. The sample mean vector is

$$\overline{V}_{100} = (0.467, 0.430, 0.422, 0.123, 0.242, 0.199)'$$

$\overline{V}_{100} \notin \mathcal{V}_4$. Minimizing the distance over the tree space gives the sample mean tree $\overline{T}_{100}$. $\overline{T}_{100}$ has the topology $\tau^{(3)}$ and has the edge lengths as

$$\boldsymbol{b} = (0.325, 0.085, 0.041, 0.097, 0.062)'$$

. The corresponding vector of $\overline{T}_{100}$ is

$$\varphi(\overline{T}_{100}) = (0.468, 0.428, 0.422, 0.123, 0.240, 0.200)'$$

The sample covariance matrix S is

$$S = \begin{pmatrix} 0.009 & 0.006 & 0.005 & 0.001 & 0.002 & 0.001 \\ & 0.007 & 0.005 & 0.001 & 0.001 & 0.002 \\ & & 0.008 & 0.001 & 0.001 & 0.001 \\ & & & 0.001 & 0.001 & 0.000 \\ & & & & 0.003 & 0.002 \\ & & & & & 0.002 \end{pmatrix}$$

Step 4. According to the central limit theorem, the $100(1-\alpha)\%$ confidence region in vector space built from this sample is

$$\left\{ \boldsymbol{x} \in \mathbb{R}^6_{>0} | (\varphi(\overline{T}_{100}) - \boldsymbol{x})' S^{-1} (\varphi(\overline{T}_{100}) - \boldsymbol{x}) < \frac{p(n-1)}{n(n-p)} F_{p,n-p}(1-\alpha) \right\}$$

The confidence set in tree space built from this sample is

$$\left\{ T \in \mathcal{T}_4 | (\varphi(\overline{T}_{100}) - \varphi(T))' S^{-1} (\varphi(\overline{T}_{100}) - \varphi(T)) < \frac{p(n-1)}{n(n-p)} F_{p,n-p}(1-\alpha) \right\}$$

It can be checked that, the sample mean vector $\overline{V}_{100}$ is in the 95% confidence region, and the ML tree estimated from RAxML is in the 95% confidence set.

Chapter 5

Conclusion and Future Work

Aiming at manipulating object in Euclidean space instead of in tree space, this dissertation utilizes a pairwise path mapping to vectorize the phylogenetic tree. By the establishment of the isomorphism between the tree space and the topological vector space, two properties are shown in Chapter 3. First, the mapping is a piecewise linear transformation that can be realized by matrices related to bipartition. Second, the topological vector sets in the Euclidean space is the positive orthant in the column space of associated matrices and structured similar to the tropical variety as a polyhedral complex in $\mathbb{R}^p$. After setting up the above mathematical infrastructure for trees in vector space, we develop its statistical application. The mean tree and tree variance are defined based on the corresponding vectors of trees, as centroid and variability measure in tree space. For the foremost application, we propose a sample mean tree as the estimator for the mean tree. The unbiasedness, consistency, and asymptotically normal distribution of the sample mean tree's vectorization are proved. A simulation is conducted to show the asymptotic normality, and a real study demonstrates a confidence procedure based on the asymptotic normality.

The significance of this work is that it constructs a work-frame on phylogenetic trees based on the preliminary vectorization, including the inducing of metric in L_2 tree space. The structure of the image vector set is presented. Furthermore, the computation and inference work on trees can be easily manipulated through the vectors by vectorization of trees.

Future work that can be perceived through this dissertation includes two major parts. The first is to develop the representation algorithms to facilitate computation on trees. The linear transformation by matrices is ideal for computation. The principles of the matrices encourage

the development of dynamic programming algorithms yielding matrices for large trees. The natural idea is that the programming should be based on the quartet decomposition of the tree. By systemizing the matrix operators, all manipulations on trees can be conducted by altering the vectors in real number space.

Another aspect is the statistical application of the vectorization of trees. If we push the sample mean tree approach further, we can utilize the property of corresponding vectors to develop various data-analytic work. For example, by construction of a confidence set from a sample of trees, the questions like "whether a particular branch should be rejected at the 0.05 significance level" can be answered quantitatively. The usage of confidence level instead of bootstrap support percentage (Yang, 2006) will give a statistically interpretable meaning. Furthermore, some adjustment methods can be exploited to make the confidence procedure more accurate.

Bibliography

[1] Abadi, S., Azouri, D., and Pupko, T. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communication*, 10, 934.

[2] Adams, E. N. (1972). Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, 21, 390-397.

[3] Allen, B. L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5 (1), 1-15.

[4] Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6: 319.

[5] Barden, D., Le, H., and Owen, M. (2014). Central limit theorems for Fréchet means in the space of phylogenetic trees. *Electronic Journal of Probability*, 18 (25).

[6] Barden, D., Le, H., and Owen, M. (2018). Limiting behaviour of Fréchet means in the space of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*, 70, 99-129.

[7] Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability for combining gene trees. *Taxon*, 41, 1- 10.

[8] Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4), 733-767.

[9] Brown, D. G. and Owen, M. (2017). Mean and Variance of Phylogenetic Trees. *arXiv:1708.00294*

[10] Bruno, W. J., Socci, N. D., and Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17, 189-197.

[11] Bryant, D. and Waddell, P. (1997). Rapid evaluation of least squares and minimum evolution criteria on phylogenetic trees. *Molecular Biology and Evolution*, 15(10), 1346-1359.

[12] Bryant, D. (2003). A classfication of consensus methods for phylogenetics. In: *BioConsensus*, 163-183.

[13] Buneman, P. (1974). A Note on the Metric Properties of Trees. *Journal of Combinatorial Theory, Series B*, 17 (1), 48-50.

[14] Cardona, G., Mir, A., Rossello, F., Rotger, A. and Sanchez, D. (2013). Cophenetic Metrics for Phylogenetic Trees, After Sokal and Rohlf. *BMC Bioinformatics*, 14 (1), 3.

[15] Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, 21, 550-570.

[16] Critchlow, D. E., Pearl, D. K. and Qian, C. (1996). The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45, 323-334.

[17] Dasgupta, B., He, X., Jiang, T., Li, M., Tromp, J. and Zhang, L. (1997). On computing the nearest neighbor interchange distance. In: Proc. DIMACS Workshop on Discrete Problems with Medical Applications, 125-143.

[18] Day, W. H. E. (1985). Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2,7-28.

[19] Do, C. B., Woods, D. A., and Batzoglous, S. (2006) CONTRAlign: discriminative training for proetin sequence align,ent. In: *Computational Molecular Biology*, Berlin/Heidelberg: Springer, 160-174.

[20] Drummond, A. J. and Rambaut, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1), 214.

[21] Du, Y., Wu, S. , Edwards, S. V., and Liu, L. (2019). The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *BMC Evolutionary Biology*, 19: 203.

[22] Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985). Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34, 193-200.

[23] Farries, J. S., Kluge, A. G., and Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Systematic Zoology*, 19, 172-191.

[24] Farris, J. S.(1973). On comparing the shapes of taxonomic trees. *Systematic Zoology*, 22, 50-54.

[25] Felsenstein, J. (1973a). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25, 471-492.

[26] Felsenstein, J. (1973b). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22, 240-249.

[27] Felsenstein, J. 1981. Evolutionary trees fromDNAsequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17, 368-376.

[28] Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155, 279-284.

[29] Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zooloogy*, 20, 406-416.

[30] Foulds, L. R. and Graham, R. L. (1982). The Steiner Problem in Phylogeny is NP-Complete. *Advances in Applied Mathematics*, 3 (1), 43-49.

[31] Galtier, N. and Daubin, V. (2008). Dealing with incongruence in phylogenomic analysis. *Philosophical Transactions of the Royal Society B*, 363, 4023-4029.

[32] Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular Biology and Evolution*, 17, 401-405.

[33] Goloboff, P. A. (1999). Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics*, 15, 415-428.

[34] Gromov, M. (1987). Hyperbolic groups. In: *Essays in group theory*, Springer, New York, 75-263.

[35] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52, 696-704.

[36] Hickey, G., Dehne, F., Rau-Chaplin, A. and Blouin, C. (2008). Spr distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4, 17-27.

[37] Hillis, D. M. and Huelsenbeck, J. P. (1994). Support for dental HIV transmission. *Nature*, 369, 24-25.

[38] Hoff, M. , Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics*, 17, 143.

[39] Huelsenbeck, J.P. and Ronquist, F. (2001). Mrbayes: Bayesian inference in phylogenetic trees. *Bioinformatics*, 17, 754-755.

[40] Iwabe, N., Kuma, K. I., Hasegawa, M., Osawa, S., and Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences of the USA, 86, 9355-9359.

[41] Jukes, T. H. and Cantor, C. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism*, New York Academic Press, 21-32.

[42] Kolaczyk, E. D. and Csárdi, D. (2014). Statistical Analysis of Network Data with R. (2nd version), Springer. 17.

[43] Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, btz305.

[44] Lemmon, A. R. and Moriarty, E. C. (2004). The importance of proper model assumption in bayesian phylogenetics. *Systematic Biology*, 53(2), 265-277.

[45] Lin, B., Sturmfels, B., Tang, X. and Yoshida, R. (2017). Convexity in Tree Spaces. *SIAM Journal on Discrete Mathematics*, 31 (3), 2015-2038.

[46] Liu, L., Wu, S., and Yu L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution*, 53 (5), 380-90.

[47] Liu, L., Xi, Z., and Davis, C. C. (2015). Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution*, 32(3), 791-805.

[48] Liu, L. and Yu L. (2020). On the evolution of word usage of classical Chinese poetry. *arXiv: 1509.04556[physics.soc-ph]*

[49] Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3), 523-536.

[50] Margush, T. and McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Bioloby*, 43 (2), 239-244.

[51] McMorris, F. R. and Neumann, D. A. (1983). A view of some consensus methods for trees. In: *Numerical Taxonomy*, Springer-Verlag, 122-125.

[52] Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. (2003). Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology*, 52, 674-683.

[53] Monod, A., Lin, B., Yoshida, R.,and Kang, Q. (2019). Tropical geometry of phylogenetic tree space: a statistical perspective. *arXIV:1805.12400v5 [math.MG]*.

[54] Morrison, D. A. and Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Molecular Biology and Evolution*, 14(4), 428-441.

[55] Mugridge, N. B., Morrison, D. A., Jakel, T., Heckeroth, A. R., Tenter, A. M., and Johnson, A. M. (2000). Effects of sequence alignment and structural domains of ribosomal DNA on phylogeny reconstruction for the protozoan family sarcocystidae. *Molecular Biology and Evolution*, 17(12), 1842-1853.

[56] Nixon, K. C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15, 407-414.

[57] Nye, T. M. (2011). Principal components analysis in the space of phylogenetic trees. *The Annals of Statistics*, 39(5), 2716-2739.

[58] Owen, M. and Provan, J. S. (2011). A Fast Algorithm for Computing Geodesic Distances in Tree Space. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 8 (1), 2-13.

[59] Penny, D. and Hendy, M. D. (1985). The use of tree comparison metrics. *Systematic Zoology*, 34(1), 75-82.

[60] Pin, J.E. (1998) Tropical semirings. In: *Idempotency*,Cambridge University Press, 50-69.

[61] Posada, D. and Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4), 580-601.

[62] Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1, 53-58.

[63] Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43, 304-311.

[64] Reid, N. M., Hird, S. M., Brown, J. M., Pelletier, T. A., McVay, J. D., Satler, J. D. and Carstens, B. C. (2014). Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data. *Systematic Biology*, 63(3), 322-333.

[65] Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3, 217-223.

[66] Richter-Gebert, J., Sturmfels, B., and Theobald, T. (2003). First Steps in Tropical Geometry. In: *Idempotent mathematics and mathematical physics, Contemporary Mathematics*, vol. 377, American Mathematical Society, 289-317.

[67] Robinson, D.F. (1971). Comparison of labeled trees with valency three. *Journal of Combinatorial Theory*, 11, 105-119.

[68] Robinson, D. F. and Foulds, L. R. (1981). Comparison of Phylogenetic Trees. *Mathematical Biosciences*, 53 (1), 131-147.

[69] Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5), 380-390.

[70] Rzhetsky, A. and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*, 9, 945-967.

[71] Salemi, M., Vandamme, A. M., and Lemey, P. (2009). The phylogenetic handbook: A pratical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK: Cambridge University Press.

[72] Schroder, E. (1870). Vier kombinatorische Probleme. *Zeitschrift fur Angewandte Mathematik and Physik*, 15, 361-376.

[73] Semple, C. and Steel, M. (2003). Phylogenetics. In: *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford. ISBN 0-19-850942-1.

[74] Shen, X. X., Hittinger, C. T, and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology and Evolution, 1(5), 126.

[75] Sneath, P. H. A. and Sokal, R. R. (1973). Numerical Taxonomy. W.H. Freeman and Company, San Francisco, CA.

[76] Speyer, D. and Sturmfels B. (2004). The Tropical Grassmannian. *Advances in Geometry*, 4(3), 389-411.

[77] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and postanalysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313.

[78] Standley, K. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772-780.

[79] Steel, M. and Penny, D. (1993). Distributions of Tree Comparison Metrics: Some New Results. *Systematic Biology*, 42 (2), 126-141.

[80] Stoye, J. (1998). Multiple sequence alignment with the Divide-and-Conquer method. *Gene*, 211 (2), GC45-GC56.

[81] Stoye, J., Moulton, V., and Dress, A. W. (1997). DCA: An efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Computer Applications in the Biosciences*, 13, 625-626.

[82] Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4), 564-577.

[83] Tavar, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. In: *Lectures on mathematics in the life sciences*, 17, 57-86.

[84] Vinh, Y. and von Haeseler, A. (2004). IQPNNI: Moving fast through tree space and stopping in time. *Molecular Biology and Evolution*, 21, 1565-1571.

[85] Yang, Z. (2006). Computational molecular evolution. Oxford University Press.

[86] Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13 (5), 303-314.

[87] Weyenberg, G. S. (2015). Statistics in the Billera-Holmes-Vogtmann Treespace. *Theses and Dissertations-Statistics.12.*

[88] Willis, A.(2017). Confidence sets for phylogenetic trees. *Journal of the American Statistical Association*,14, 235-244.

[89] Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862), 473476.

[90] Wu, M., Chatterji, S., Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PLoS One*, 7(1), e30288.

[91] Wu, S., Edwards, S. V., Liu, L. (2018). Genome-scale DNA sequence data and the evolutionary history of placental mammals. *Data in Brief*, 18, 1971-1975.

Appendix

Proofs

Proof to LEMMA 36

LEMMA 36. *Each $M^{(k)}$ is full column rank.*

Proof.

The values in every column of $M^{(k)}$ are either 0 or 1. Each pairwise path will cover a set of edges, while any two pairwise paths will not cover the same set of edges. Therefore, the columns in $M^{(k)}$ will be different from each other.

Assume there is $\sum_{l=1}^{2m-3} a_l \boldsymbol{I}_l = \mathbf{0}$, where $\{a_l, l = 1, \ldots, 2m-3\}$ are real number coefficients.

Note that there are at least two "end" edges in the tree. Here, end edge refers to the internal edge that is adjacent to three terminal edges. Assume one such edge is e_{l_0} as shown in Figure 5.1. It appears that this edge is incident to three tips and one cluster of tips. The

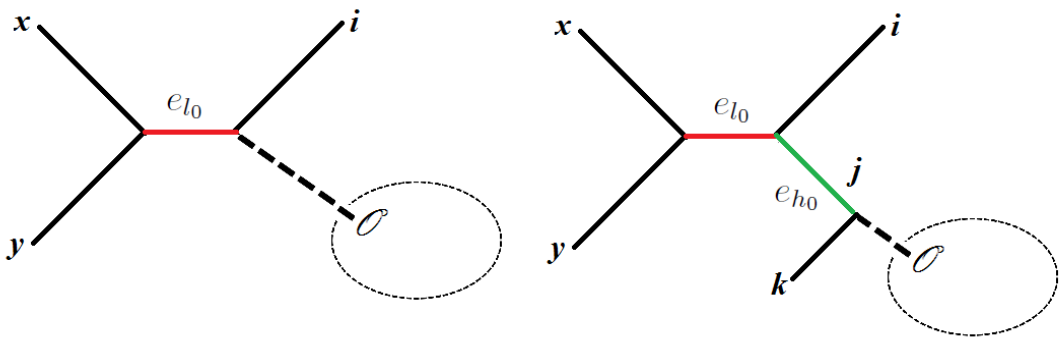

Figure 5.1: A general situation of the "end" edge and the edge adjacent to "end" edge

path between the pair (x, i) covers e_{l_0} and no other internal edges. That is,

$$\begin{aligned}\textstyle\sum_{l=1}^{2m-3} a_l I_l(x, y) &= a_x + a_y = 0\\ \textstyle\sum_{l=1}^{2m-3} a_l I_l(x, i) &= a_x + a_{l_0} + a_i = 0\\ \textstyle\sum_{l=1}^{2m-3} a_l I_l(y, i) &= a_y + a_{l_0} + a_i = 0\end{aligned}$$

This gives $a_x = a_y = 0$ and $a_{l_0} + a_i = 0$.

Then, assume the internal edge e_{h_0} is adjacent to e_{l_0}, as shown by Figure 5.1. The path between the pair (x, j) covers e_{l_0} and e_{h_0}. Therefore,

$$\begin{aligned}\textstyle\sum_{l=1}^{2m-3} a_l I_l(x, j) &= a_{l_0} + a_{h_0} = 0\\ \textstyle\sum_{l=1}^{2m-3} a_l I_l(i, j) &= a_i + a_{h_0} = 0\end{aligned}$$

Combined with $a_{l_0} + a_i = 0$, this gives $a_i = a_{l_0} = a_{h_0} = 0$.

Extend this step further, until to the farthest pair that covers the most internal edges in the graph. We can get that every a_l, $(l = m + 1, \ldots, 2m - 3)$ is 0.

Therefore, the $2m - 3$ columns $\boldsymbol{I}_l$ in the topology matrix $M^{(k)}$ are independent. □

Proof to LEMMA 43

LEMMA 43. For two trees $T_1, T_2 \in \mathcal{T}_m$, $\varphi(T_1) + \varphi(T_2) \in \mathcal{V}_m$ if and only if T_1 and T_2 are compatible.

Proof.

(1). Assume T_1 and T_2 are in the compatible tree class $\mathcal{T}^{(k)}$, and $T_1 = (\tau^{(k)}, \boldsymbol{b}_1)$, $T_2 = (\tau^{(k)}, \boldsymbol{b}_2)$. Then, $\varphi(T_1) + \varphi(T_2) = M^{(k)}(\boldsymbol{b}_1 + \boldsymbol{b}_2)$. There exists $T = (\tau^{(k)}, \boldsymbol{b}_1 + \boldsymbol{b}_2)$ such that $\varphi(T) = \varphi(T_1) + \varphi(T_2)$.

(2). Assume T_1 and T_2 are in two distinct tree classes. Then, there is at least one quartet (x, y, i, j) such that the quartet subtrees (x, y, i, j) inherited from T_1 and T_2 have different topologies. Without loss of generality, assume the quartet subtree (x, y, i, j) in T_1 has the topology of $\tau^{(1)}$ while the quartet subtree (x, y, i, j) in T_2 has the topology of $\tau^{(2)}$, which is

shown by Figure 3.9. Then, their four-point conditions are respectively,

$$\begin{aligned} d_{xy}^{(1)} + d_{ij}^{(1)} &= x_1 - \delta_1 \\ d_{xi}^{(1)} + d_{yj}^{(1)} &= x_1 \\ d_{xj}^{(1)} + d_{yi}^{(1)} &= x_1 \end{aligned}$$

$$\begin{aligned} d_{xy}^{(2)} + d_{ij}^{(2)} &= x_2 \\ d_{xi}^{(2)} + d_{yj}^{(2)} &= x_2 - \delta_2 \\ d_{xj}^{(2)} + d_{yi}^{(2)} &= x_2 \end{aligned}$$

If there is a tree T such that $\varphi(T) = \varphi(T_1) + \varphi(T_2)$, then the pairwise path distance in T associated with the quartet (x, y, i, j) is

$$\begin{aligned} d_{xy} + d_{ij} &= x_1 + x_2 - \delta_1 \\ d_{xi} + d_{yj} &= x_1 + x_2 - \delta_2 \\ d_{xj} + d_{yi} &= x_1 + x_2 \end{aligned}$$

It does not satisfy any of the min-plus inequalities. Therefore, (x, y, i, j) is not a quartet in T. It contradicts the assumption. Therefore, if the summation of two vectors corresponding to two trees can be projected back into tree space, then these two trees are compatible. Its contrapositive proposition also holds. □

Proof to LEMMA 45

LEMMA 45. $\mathcal{V} = \{\varphi(T) | T \in \mathcal{T}\}$ is a complete subset in $\mathbb{R}^p_{>0}$.

Proof. Assume a series of vectors $\{V_1, V_2, \dots\}$ in $\mathcal{V}$ has its limitation $V = \lim_{k\to\infty} V_k \in \mathbb{R}^p_{>0}$, i.e., $\|V_k - V\|_2 \to 0$. For an arbitrary quartet (x, y, i, j), denote the corresponding subvectors as $\tilde{V}_k$ and $\tilde{V}$. Then, $\left\|\tilde{V}_k - \tilde{V}\right\|_2 \to 0$.

We claim that, there are two possibilities.

(1). There exists $N \in \mathbb{N}$, such that $\forall k > N$, $\tilde{V}_k$ satisfies the same min-plus inequality with $\tilde{V}$. Without loss of generality, assume the series of quartets has topology of $\tau^{(1)}$, then

$$\begin{aligned}
\lim_{k\to\infty}(d^k_{xy} + d^k_{ij}) &= \lim_{k\to\infty}(x^{(k)} - \delta^{(k)}) \\
\lim_{k\to\infty}(d^k_{xi} + d^k_{yj}) &= \lim_{k\to\infty} x^{(k)} \\
\lim_{k\to\infty}(d^k_{xj} + d^k_{yi}) &= \lim_{k\to\infty} x^{(k)}
\end{aligned}$$

Then, we can see $\tilde{V}$ satisfies the min-plus inequality (1).

(2). Otherwise, the series of quartets (x, y, i, j) will convergence to the star tree. Without loss of generality, assume there are two subsequences $\tilde{V}_{k1}$ and $\tilde{V}_{k2}$ converging to $\tilde{V}$, and respectively have the topology of $\tau^{(1)}$ and $\tau^{(2)}$. Then

$$\begin{aligned}
\lim_{k1\to\infty}(d^{(k1)}_{xy} + d^{(k1)}_{ij}) &= \lim_{k1\to\infty}(x^{(k1)}_1 - \delta^{(k)}_1) &&= \lim_{k2\to\infty}(d^{(k2)}_{xy} + d^{(k2)}_{ij}) = \lim_{k2\to\infty} x^{(k2)}_2 \\
\lim_{k1\to\infty}(d^{(k1)}_{xi} + d^{(k1)}_{yj}) &= \lim_{k1\to\infty} x^{(k1)}_1 &&= \lim_{k2\to\infty}(d^{(k2)}_{xi} + d^{(k2)}_{yj}) = \lim_{k2\to\infty}(x^{(k2)}_2 - \delta^{(k2)}_2) \\
\lim_{k1\to\infty}(d^{(k1)}_{xj} + d^{(k1)}_{yi}) &= \lim_{k1\to\infty} x^{(k1)}_1 &&= \lim_{k2\to\infty}(d^{(k2)}_{xj} + d^{(k2)}_{yi}) = \lim_{k2\to\infty} x^{(k2)}_2
\end{aligned}$$

That gives,

$$\lim_{k1\to\infty} x^{(k1)}_1 = \lim_{k2\to\infty} x^{(k2)}_2$$

and

$$\lim_{k1\to\infty} \delta^{(k1)}_1 = \lim_{k2\to\infty} \delta^{(k2)}_2 = 0$$

This suggests that these two subsequences converge to $\tilde{V}$ corresponding to star tree.

In either of the case, the limitation $\tilde{V}$ satisfies the four-point condition and corresponds to a quartet subtree that is the "limitation" of the series of quartet subtrees associated with $\{V_1, V_2, \dots\}$ in L_2 tree space. Therefore, the vector V corresponds to a tree. $\mathcal{V}$ is closed. Because $\mathbb{R}^p_{>0}$ is complete, therefore $\mathcal{V}$ is complete. □

R code for simulation

Simulation

```
library(ape)
```

```
library(phytools)

m0<-matrix(0,ncol=6,nrow=15)
for(i in 1:15) m0[i,t(combn(6,2))[i,]]<-1

par(mfrow=c(1,3))
t1<-read.tree(text="((a,b),c,(d,(e,f)));")
plot(t1,type="u",edge.width = 3)
m1.int<-matrix(c(0,0,0,
                 1,0,0,
                 1,1,0,
                 1,1,1,
                 1,1,1,
                 1,0,0,
                 1,1,0,
                 1,1,1,
                 1,1,1,
                 0,1,0,
                 0,1,1,
                 0,1,1,
                 0,0,1,
                 0,0,1,
                 0,0,0),byrow=T,ncol = 3)

t2<-read.tree(text="((a,b),(c,d),(e,f));")
plot(t2,type="u",edge.width = 3)
m2.int<-matrix(c(0,0,0,
                 1,1,0,
                 1,1,0,
                 1,0,1,
                 1,0,1,
                 1,1,0,
                 1,1,0,
                 1,0,1,
                 1,0,1,
                 0,0,0,
                 0,1,1,
                 0,1,1,
                 0,1,1,
                 0,0,0),byrow=T,ncol = 3)
t3<-read.tree(text="(((a,c),f),b,(d,e));")
plot(t3,type="u",edge.width = 3)
m3.int<-matrix(c(1,1,0,
                 0,0,0,
```

```
                  1,1,1,
                  1,1,1,
                  1,0,0,
                  1,1,0,
                  0,0,1,
                  0,0,1,
                  0,1,0,
                  1,1,1,
                  1,1,1,
                  1,0,0,
                  0,0,0,
                  0,1,1,
                  0,1,1),byrow=T,ncol=3)
m1<-cbind(m0,m1.int)
m2<-cbind(m0,m2.int)
m3<-cbind(m0,m3.int)
mlist<-list(m1,m2,m3)

t1.mean<-m1%*%c(rep(1,6),rep(2,3))
t2.mean<-m2%*%c(rep(1,6),rep(2,3))
t3.mean<-m3%*%c(rep(1,6),rep(2,3))

Vf<-t1.mean/2+t2.mean/4+t3.mean/4
opt2<-function(V)
{
  D<-matrix(0,ncol=6,nrow=6)
  D[t(combn(6,2))]<-V
  D[lower.tri(D)]<-t(D)[lower.tri(D)]
  dimnames(D)<-list(c("a","b","c","d","e","f"),c("a","b","c","d","e","f"))
  return(optim.phylo.ls(D))
}
Tf<-opt2(Vf)
ordered.Tf.V<-m1%*%Tf$b

opt2<-function(V)
{
  min1<-sum((V-m1%*%ginv(m1)%*%V)^2)
  min2<-sum((V-m2%*%ginv(m2)%*%V)^2)
  min3<-sum((V-m3%*%ginv(m3)%*%V)^2)
  ind<-which.min(c(min1,min2,min3))
  return(list(type=ind,b=ginv(mlist[[ind]])%*%V))
}

#--------------------------
#
```

```
#    100 replicates, n=10
#
#--------------------------
res1<-NULL
RF.d1<-NULL
for(i in 1:100)
{
  V<-NULL
  for(j in 1:10)
  {ind<-sample(c(1,1,2,3),1)
  V<-cbind(V,mlist[[ind]]%*%c(rep(1,6),rexp(3,0.5)))}
  V.tmp<-apply(V,1,mean)
  ordered.T.avg.V<-mlist[[opt2(V.tmp)$type]]%*%opt2(V.tmp)$b
  DIFF<-ordered.T.avg.V-ordered.Tf.V
  res1<-cbind(res1,DIFF)
  RF.d1<-c(RF.d1,opt2(V.tmp)$type==1)
}


par(mfrow=c(3,5))
for(i in 1:15)
  {hist(res1[i,],main=paste("V",i),xlab="")
  abline(v=0,col="red",lty=2,lwd=2)}
dev.off()

res.square1<-NULL
for(i in 1:100) res.square1<-c(res.square1,norm(res1[,i]))
#--------------------------
#
#    100 replicates, n=50
#
#--------------------------
res2<-NULL
RF.d2<-NULL
for(i in 1:100)
{
  V<-NULL
  for(j in 1:50)
  {ind<-sample(c(1,1,2,3),1)
  V<-cbind(V,mlist[[ind]]%*%c(rep(1,6),rexp(3,0.5)))}
  V.tmp<-apply(V,1,mean)
  ordered.T.avg.V<-mlist[[opt2(V.tmp)$type]]%*%opt2(V.tmp)$b
  DIFF<-ordered.T.avg.V-ordered.Tf.V
  res2<-cbind(res2,DIFF)
  RF.d2<-c(RF.d2,opt2(V.tmp)$type==1)
```

```
}


par(mfrow=c(3,5))
for(i in 1:15)
  {hist(res2[i,],main=paste("V",i),xlab="")
  abline(v=0,col="red",lty=2,lwd=2)}
dev.off()

res.square2<-NULL
for(i in 1:100) res.square2<-c(res.square2,norm(res2[,i]))

#---------------------------
#
#    100 replicates, n=300
#
#---------------------------
res3<-NULL
RF.d3<-NULL
for(i in 1:100)
{
  V<-NULL
  for(j in 1:300)
  {ind<-sample(c(1,1,2,3),1)
  V<-cbind(V,mlist[[ind]]%*%c(rep(1,6),rexp(3,0.5)))}
  V.tmp<-apply(V,1,mean)
  ordered.T.avg.V<-mlist[[opt2(V.tmp)$type]]%*%opt2(V.tmp)$b
  DIFF<-ordered.T.avg.V-ordered.Tf.V
  res3<-cbind(res3,DIFF)
  RF.d3<-c(RF.d3,opt2(V.tmp)$type==1)
}


par(mfrow=c(3,5))
for(i in 1:15)
{hist(res3[i,],main=paste("V",i),xlab="")
 abline(v=0,col="red",lty=2,lwd=2)}
dev.off()


res.square3<-NULL
for(i in 1:100) res.square3<-c(res.square3,norm(res3[,i]))


#---------------------------
```

```
#
#    random linear combination
#
#--------------------------

a<-runif(15,min=0,max=3)
par(mfrow=c(1,3))
hist(t(res1)%*%a,xlab="",main="n=10")
abline(v=0,col="red",lty=2,lwd=2)
hist(t(res2)%*%a,xlab="",main="n=50")
abline(v=0,col="red",lty=2,lwd=2)
hist(t(res3)%*%a,xlab="",main="n=300")
abline(v=0,col="red",lty=2,lwd=2)
round(a,3)
dev.off()

#--------------------------
#
#    Distance Distribution
#
#--------------------------
par(mfrow=c(1,3))
hist(100*res.square1,xlab="",main="n=10",prob=T)
lines(density(100*res.square1),col="red",lty=2,lwd=2)
hist(100*res.square2,xlab="",main="n=50",prob=T)
lines(density(100*res.square2),col="red",lty=2,lwd=2)
hist(100*res.square3,xlab="",main="n=300",prob=T)
lines(density(100*res.square3),col="red",lty=2,lwd=2)
dev.off()
```

Real Study

```
library(ape)
library(phangorn)
library(ggplot2)

a<-read.table("bstrees.txt",header=F)

best.tree<-read.tree(text="((Dog:0.040262,Cow:0.083899):0.050133,Chicken:0.334728,Rat:
0.102365):0.0;")
plot(best.tree,type="u",edge.width = 3)
edgelabels(round(best.tree$edge.length,2))
```

```
tau1<-read.tree(text="((Chicken,Cow),(Dog,Rat));")
tau2<-read.tree(text="((Chicken,Dog),(Cow,Rat));")
tau3<-read.tree(text="((Chicken,Rat),(Cow,Dog));")

a<-unfactor(a)
tree.list<-list()
ee.list<-list()
type<-c()
for(i in 1:100)
{
t<-read.tree(text=a[i,])
ee<-setNames(t$edge.length[sapply(1:4,function(x,y) which(y==x), y=t$edge[,2])],
t$tip.label)
ee[5]<-t$edge.length[which(!t$edge.length%in%ee)]
names(ee)[5]<-"Internal"
ee<-c(ee[order(names(ee)[1:4])],ee[5])
tree.list[[i]]<-t
ee.list[[i]]<-ee
if(RF.dist(tau1,t)==0) type<-c(type,1)
else if(RF.dist(tau2,t)==0) type<-c(type,2)
else if(RF.dist(tau3,t)==0) type<-c(type,3)
else type<-c(type,0)
}
table(type)

m0<-matrix(c(1,1,0,0,
             1,0,1,0,
             1,0,0,1,
             0,1,1,0,
             0,1,0,1,
             0,0,1,1),byrow=T,ncol=4)
I1<-c(0,1,1,1,1,0)
I2<-c(1,0,1,1,0,1)
I3<-c(1,1,0,0,1,1)
m1<-cbind(m0,I1)
m2<-cbind(m0,I2)
m3<-cbind(m0,I3)
mlist<-list(m1,m2,m3)

#plot for internal edges

internal.length<-NULL
for(i in 1:100) internal.length<-c(internal.length,ee.list[[i]][5])
summary<-data.frame(type,internal.length)
gg<-ggplot(summary[type!=0,],aes(x=internal.length,y=type,color=factor(type)))+
```

```
geom_point()
gg+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour="black"))

# sample mean vector

V.set<-NULL
for(i in 1:100)
{
V<-mlist[[type[i]]]%*%ee.list[[i]]
V.set<-rbind(V.set,t(V))
}
V.average<-apply(V.set,2,mean)

# sample mean tree
norm<-function(v) sqrt(sum(v^2))
norm(m1%*%ginv(m1)%*%V.average-V.average)
norm(m2%*%ginv(m2)%*%V.average-V.average)
norm(m3%*%ginv(m3)%*%V.average-V.average)
#type 3 is the optimal
T.average<-ginv(m3)%*%mV.average
V.T.average<-m3%*%ginv(m3)%*%V.average

# sample covariance
S<-cov(V.set)

# Is Vf in the confidence region
t(V.average-V.T.average)%*%solve(S)%*%(V.average-V.T.average)<6*99/(100*94)*
qf(0.95,6,94)

# Is the best tree from RAxML in the confidence region
best.ee<-setNames(best.tree$edge.length[sapply(1:4,function(x,y) which(y==x),
y=t$edge[,2])],best.tree$tip.label)
best.ee[5]<-best.tree$edge.length[which(!best.tree$edge.length%in%best.ee)]
names(best.ee)[5]<-"Internal"
best.ee<-c(best.ee[order(names(best.ee)[1:4])],best.ee[5])
V.best.tree<-m3%*%best.ee
t(V.best.tree-V.T.average)%*%solve(S)%*%(V.best.tree-V.T.average)<6*99/(100*94)*
qf(0.95,6,94)
```