Title
Combinatorial Phylogenetics of Reconstruction Algorithms

Author
Kleinman, Aaron Douglas

Publication Date
2012

Peer reviewed|Thesis/dissertation

Combinatorial Phylogenetics of Reconstruction Algorithms

by

Aaron Douglas Kleinman

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Chair
Professor Bernd Sturmfels
Professor Satish Rao

Spring 2012

Combinatorial Phylogenetics of Reconstruction Algorithms

Abstract

Combinatorial Phylogenetics of Reconstruction Algorithms

by

Aaron Douglas Kleinman

Doctor of Philosophy in Mathematics

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Lior Pachter, Chair

Phylogenetics is the study of the evolutionary history of different organisms. A reconstruction algorithm is a technique for producing a tree from molecular or morphological data that is believed to have evolved in a tree-like fashion. In this thesis, we present a number of new combinatorial results that have implications for the accuracy and significance of some of these methods.

We begin by exploring generalizations of phylogenetic trees known as PQ- and PC-trees. In Chapter 2, we show how these objects, which have appeared repeatedly throughout computer science literature, arise naturally by relaxing the combinatorial condition in the splits equivalence theorem for regular trees. We determine the appropriate analog of the four-point condition and precisely characterize the metrics that come from these trees. One of our main results is an algorithm to constructively produce the PQ- or PC-tree that best describes a certain class of metrics. Throughout, we describe a single framework that unites a number of different known combinatorial objects, many for the first time.

In Chapter 3, we study the robustness of a class of distance-based reconstruction algorithms known as minimum evolution. We focus on those methods that are linear in the elements of the dissimilarity. Our main result is that one such method, known as balanced minimum evolution, is the unique method with a certain accuracy guarantee. Although this is a significant result in its own right, it is especially important because balanced minimum evolution is the theoretical underpinning of neighbor-joining, the gold standard of distance-based reconstruction algorithms. Our theorem is the last in a long line of results, stretching back over 20 years, that "explain" neighbor-joining, and it completes our understanding of the algorithm. We next compute the robustness of the traveling salesman linear form as a reconstruction method for Kalmanson dissimilarities. Lastly, we define families of balanced minimum evolution- and traveling salesman-like forms parameterized by real functions on the set of X-splits and investigate their robustness.

In Chapter 4, we examine the problem of bounding the size of maximum agreement subtrees between pairs of trees. Several polynomial-time algorithms already exist for this, but the extremal problem of how large the agreement subtree must be remains open. In 1992, Kubicka, Kubicki and McMorris conjectured that there exists a constant c such that a pair of trees on c^n leaves has a maximum agreement subtree on n leaves. We make substantial progress toward this conjecture and show $c^{n \log n}$ leaves suffice. This represents a large improvement over the previous best bound, c^{c^n}. We also adapt a proof of the Erdös-Szekeres theorem to give an interval of length 2 containing the minimum size $|X|$ such that two caterpillar X-trees must have an agreement subtree of size n.

To my family.

Contents

List of Figures

Acknowledgments

I want to begin by thanking my adviser, Lior Pachter. Without his enthusiasm, encouragement and generosity this dissertation could not have been written. He has been a true role model, not merely in how to conduct research, but also in how to conduct oneself as a mathematician and as a scientist, and I am and will remain deeply grateful for his guidance and support.

My thanks go to Bernd Sturmfels for several helpful conversations, for a very careful reading of this thesis and for asking a question that led to one of the results contained herein. Thanks also to Satish Rao for sitting on my dissertation committee, to Craig Evans for chairing my qualifying exam, and to the Berkeley math department, the NSF and the Institute for Pure and Applied Mathematics for their financial support during my years of graduate study.

I am deeply grateful for the many people throughout my life who have spent their time and energy fostering my love of mathematics, including Suzanne Williams, Rita Hellyer, Ted Alper and Joshua Zucker.

Thank you to the other members of the Pachter lab, who made Evans a pleasant place to work. I'd also like to thank my friends who helped make my time at Berkeley so enjoyable, including George Schaeffer, David Zureick-Brown and the members of Pretty Basic, the math department ultimate team. I owe a special thank you to Daniel Cristofaro-Gardiner, my office mate these past five years; my graduate school experience has been much improved with him alongside.

Last but not least, I want to thank those who matter most. To Andrea, Dan and my parents: Thank you for your unwavering support and love these past five years. You have always believed in me, and that has made the difference.

Chapter 1

Introduction

1.1 Preliminaries

In his "Notebook B: Transmutation of species" [16], Charles Darwin drew a single figure to illustrate the shared ancestry of extant species (Figure 1.1). That figure is a depiction of a graph known as a *rooted X-tree* [63].

Definition 1.1.1. Let X be a finite set. A *tree* is a connected acyclic graph. A *phylogenetic X-tree* is a pair (T, ϕ), where T is an unrooted tree $T = (V, E)$, and ϕ is a map $\phi : X \to V$ such that $\phi(X)$ contains every vertex of degree at most 2. Two X-trees (T_1, ϕ_1), (T_2, ϕ_2) are isomorphic if there exists a graph isomorphism $\Phi : T_1 \to T_2$ such that $\phi_2 = \Phi \circ \phi_1$. A *rooted X-tree* is an X-tree with a special vertex, denoted r, of degree at least 2. Two rooted X-trees T_1, T_2 with roots r_1, r_2 are isomorphic if there exists a graph isomorphism $\Phi : T_1 \to T_2$ such that $\phi_2 = \Phi \circ \phi_1$ and $r_2 = \Phi(r_1)$.

Rooted X-trees are useful because they graphically describe how evolution occurs. Extant species are represented by the leaves of the tree, while internal vertices represent speciation events. The root represents a shared ancestor of the extant species, and vertices on a path from a vertex v to the root correspond to ancestors of v. Such trees are useful in other classification problems as well. For example, they have been used when X is a set of individuals in a population, or a set of languages [24]. In most cases the vertices of primary interest are the leaves.

Definition 1.1.2. An X-tree (T, ϕ) is a *phylogenetic X-tree* if ϕ is a bijection from X to the leaves of T. We call a phylogenetic X-tree *trivalent* if, in addition, each internal vertex is of degree 3. A *rooted binary X-tree* is a rooted X-tree whose leaves are in bijection with the elements of X, and with every internal vertex of degree 3 except for the root, which is of degree 2.

The fundamental question of phylogenetics is: Given data D on X that arose from a phylogenetic X-tree T, is it possible to reconstruct T? The kind of data at our disposal can

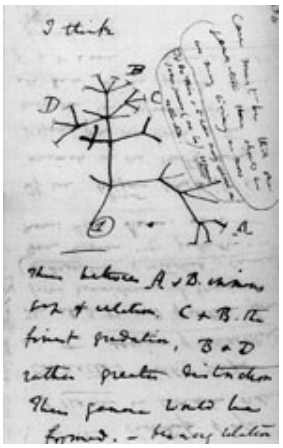


Figure 1.1: The first recorded phylogenetic tree, as seen in one of Darwin's notebooks.

take many forms. In biology, one often obtains D by sequencing the genomes of the extant species and computing a multiple alignment of homologous regions. Each column of this alignment gives us a single nucleotide at each leaf, and the whole alignment is considered as many independent samples from the same underlying distribution. We assume these nucleotides evolved according to some Markov process running on the hidden tree T with rates parameterized by the branch lengths of the tree; with a suitable prior, the problem then becomes that of computing the maximum likelihood tree. Bayesian approaches to tree reconstruction have been very effective in practice, but they can also be computationally expensive.

Another popular approach uses *character data*. A character consists of a discrete set S of states and a map $\chi : X \to S$ that assigns one of these states to each taxon. These characters can correspond to a nucleotide at a particular genomic site, a morphological feature or some other inheritable trait. The data D is a collection of characters, and the method of *maximum parsimony* seeks to reconstruct the tree which explains these characters with minimal mutations. This can be effective, but maximum parsimony is known to not be statistically consistent [31] – that is, even as the amount of data tends to infinity, a character-based approach may not return the correct tree topology.

Our main focus will be on *distance-based methods*. Here, the data takes the form of a matrix of pairwise distances between the taxa. This can be computed from a multiple alignment; one simple method would be to assign two taxa a distance equal to the number of sites where their alignment differs. There are ways of computing genomic distances directly without aligning sequences; also, the distances can come from other characteristics of the taxa and not be genetic at all. Trees are fundamentally combinatorial objects, and in this thesis we will study how their combinatorics both inform and are informed by distance-based methods in phylogenetics.

1.2 Combinatorics of X-trees

We begin with an introduction to some of the fundamental combinatorics that will serve as background to the upcoming chapters. General references include [23, 63].

Definition 1.2.1. A *split* of X is a partition of X into two nonempty pieces $A|B$. A split is *trivial* if $|A|$ or $|B|$ is 1. A set of splits containing all the trivial splits called a *split system*. Two X-splits $A_1|B_1$, $A_2|B_2$ are *compatible* if $A_1 \subseteq A_2$, $A_1 \subseteq B_2$, $B_1 \subseteq A_2$ or $B_1 \subseteq B_2$, and are *incompatible* otherwise.

Removing an edge e of a phylogenetic X-tree T disconnects the tree and partitions the vertices into two pieces V_1, V_2, and thus gives rise to an X-split $S_e = \phi^{-1}(V_1)|\phi^{-1}(V_2)$. Doing this for each edge of $T = (V, E)$ allows us to associate a split system $\mathcal{S}(T) = \{S_e\}_{e \in E}$ to T. If $S \in \mathcal{S}(T)$ we say T *contains* the split S. It is easy to check that each pair of splits in such a system is compatible. A classic result says that the converse also holds [10]:

Theorem 1.2.2 (Splits equivalence theorem)**.** *A split system is of the form $\mathcal{S}(T)$ if and only if the elements of the split system are pairwise compatible.*

Horizontal gene transfers, hybridization and reticulation mean real-world biological data does not always arise from a tree-like topology. And even when it does, noise might inject conflicting signals into the data; coercing the data to fit a tree might require throwing away valuable information. Theorem 1.2.2 suggests a convenient way of generalizing trees by relaxing the compatibility condition. In Chapter 2 we will show precisely how PQ- and PC-trees, combinatorial objects that have arisen in a number of different contexts in computer science, are derived in this way.

We can use splits to define a distance between trees. Given two binary X-trees T_1, T_2, let $d(T_1, T_2) = |\Delta(\mathcal{S}(T_1), \mathcal{S}(T_2))|$, where $\Delta(A, B)$ is the symmetric difference. Two trees are said to be separated by a *nearest-neighbor interchange (NNI)* if $d(T_1, T_2) = 2$, the smallest possible nonzero value. This means one tree can be obtained from the other by transposing two subtrees that are precisely three edges apart. We can then represent the set $\mathcal{T}_X$ of trivalent X-trees as a graph whose vertices correspond to trees, and whose edges correspond to pairs of trees separated by an NNI. This graph gives a convenient way of exploring tree space, and in Chapter 3, we will prove our main theorem by comparing a tree to each of its NNIs.

Definition 1.2.3. A *clade* of X is a subset $A \subset X$ such that $A|X \setminus A$ is a split of X

In Figure 1.2, and in other figures throughout, circles represent vertices while triangles represent subtrees or clades.

Definition 1.2.4. A *quartet* is a 4-tuple of distinct taxa $(ij : kl)$ partitioned into two sets of size two. We say a tree T *contains* $(ij : kl)$ if there exists a split $S = A|B$ of T such that $i, j \in A$, $k, l \in B$.

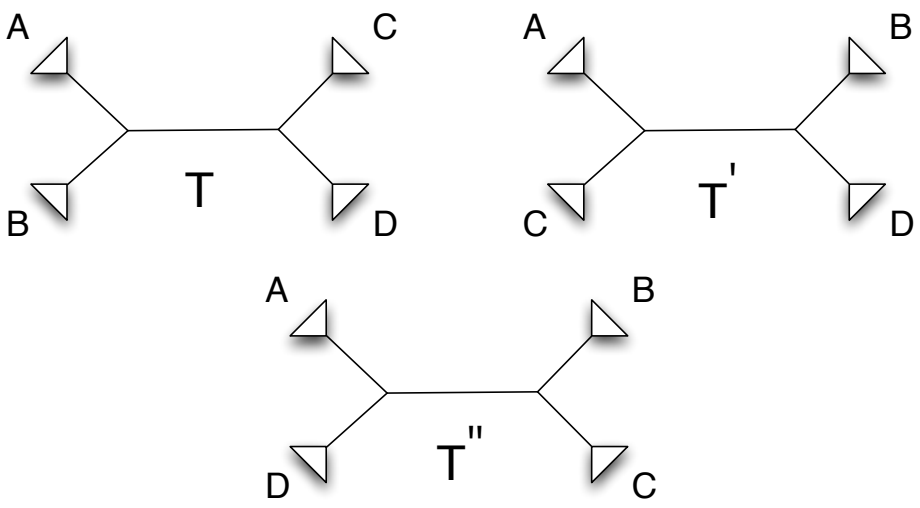


Figure 1.2: Three trees T, T', T'' that are nearest-neighbor interchanges of each other.

Let $T \in \mathcal{T}_n$ and let $Q(T)$ denote the set of $\binom{n}{4}$ quartets of T. Then $Q(T_1) = Q(T_2)$ if and only if T_1 and T_2 are isomorphic [23].

1.3 Distances on trees

While leaf-labeled trees are topologically interesting objects in and of themselves, in phylogenetics it is desirable to associate lengths with the edges of trees. Such lengths may correspond to time (in years), or to the number of mutations (usually an estimate based on a statistical model). This gives rise to a matrix of pairwise distances on X.

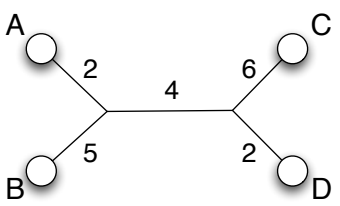
We now make this more precise.

Definition 1.3.1. A *dissimilarity* on X is a map $D : X \times X \to \mathbb{R}$ such that $D_{ij} = D_{ji}$ and $D_{ii} = 0$ for all $i \in X$.

For notational convenience, we will occasionally write $D(i, j)$ to mean D_{ij}. Let $w : E(T) \to \mathbb{R}^{\geq 0}$ be a function that assigns a non-negative weight to each edge of T, and let P_{ij}^T denote the path between i and j. (We occasionally write $P_T(i, j)$ to mean the same thing). Then w naturally gives rise to the dissimilarity values

$$D_{ij} = \sum_{e \in P_{ij}^T} w_e.$$

Such a dissimilarity is said to be *T-additive*. If it is T-additive for some T, we say it is



$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 0 & 7 & 12 & 8 \\ 7 & 0 & 15 & 11 \\ 12 & 15 & 0 & 8 \\ 8 & 11 & 8 & 0 \end{pmatrix} \end{array}$$

Figure 1.3: A weighted X-tree on 4 taxa and its corresponding tree-additive dissimilarity.

tree-additive. More succinctly, let S_T be the $\binom{n}{2} \times |E|$ matrix given by

$$(S_T)_{ij,e} = \begin{cases} 1 & e \in P_{ij}^T, \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Consider $w = (w_e)$ as a vector of length $|E|$. Then D is tree-additive if $D = S_T w$ for some non-negative w.

Here's still another way. Given a split $A|B$, let the *split pseudometric* $\sigma_{A|B}$ be the dissimilarity given by

$$\sigma_{A|B}(i,j) = \begin{cases} 1 & |\{i,j\} \cap A| = 1 \\ 0, & \text{otherwise.} \end{cases}$$

(We can extend this in the natural way to the case when A, B are disjoint clades). Given an edge $e \in E(T)$, let σ_e denote $\sigma_{A|B}$ where $A|B$ is the split corresponding to e. Then D is T-additive if and only if $D = \sum_{e \in E(T)} w_e \sigma_e$ for some non-negative weighting $w : E(T) \to \mathbb{R}^{\geq 0}$.

Although real-world data is expected to arise from a rooted tree, distance-based reconstruction methods cannot detect the location of the root. This is because the distances are assumed to arise from a time-reversible process, so the location of the root does not affect the resulting dissimilarity. Such reconstruction methods thus attempt to determine the unrooted tree topology.

A classic theorem [56] gives a precise combinatorial characterization of the space of tree-additive dissimilarities.

Theorem 1.3.2 (Four-point condition)**.** *A positive dissimilarity D is tree-additive for a tree with nonnegative edge weights if and only if, for all $i, j, k, l \in X$,*

$$D_{ij} + D_{kl} \leq \max\{D_{ik} + D_{jl}, D_{il} + D_{jk}\}. \tag{1.2}$$

One can check that D is tree-additive and, if it is, reconstruct the underlying tree in $O(n^2)$ time. Real world data is never this nice, however. One reason is that evolution – particularly at the microbial level – does not always proceed in a purely tree-like fashion. Instead, horizontal gene transfers, hybridization and reticulation cause conflicting signals in

the underlying data and are thus better modeled by a more general structure known as a *splits network* [51]. Several techniques have been developed for recovering an underlying network from a dissimilarity [8, 43] but they have been slow to be adopted by biologists, perhaps because their visualization does not look very tree-like. In Chapter 2 we generalize Theorem 1.3.2 to metrics arising from *PQ-* and *PC-trees*, combinatorial objects that interpolate between traditional trees and split networks. It is our hope that these objects, which can be used to model non-treelike data while still looking visually like a tree, will prove more useful to biologists.

Another reason metrics are rarely tree-additive is because the space of tree-additive dissimilarities has measure zero in $\mathbb{R}^{\binom{n}{2}}$, so the presence of noise almost surely perturbs the dissimilarity to be non-tree-additive. So Theorem 1.3.2 also motivates the development of algorithms that, given a dissimilarity D, return the tree topology T such that T "best" explains D. Broadly speaking, there are two distinct challenges to this problem. The first is theoretical: How should we define "best-fit tree"? Typically, one defines a scoring function $\phi : \mathcal{T}_X \times \mathbb{R}^{\binom{n}{2}} \to \mathbb{R}$ such that $\phi(T, D)$ measures how good the tree topology T explains the data D. We then choose $\arg\max_T \phi(T, D)$, the tree with the best score. For example, we might assume that the data evolved according to a probabilistic model. In this setting ϕ could be the likelihood function, and $\arg\max_T \phi(T, D)$ is the maximum likelihood estimate of the tree topology. In minimum evolution (ME) methods, which we study in Chapter 3, $\phi(T, D)$ represents the sum of the branch lengths of the tree topology, assuming D came from T; in this case we wish to choose the tree that minimizes $\phi(\cdot, D)$.

Let $\mathcal{T}_n = \mathcal{T}_{[n]}$ be the set of trivalent X-trees on n leaves. A classic result [56] states $|\mathcal{T}_n| = \frac{(2n-4)!}{2^{n-2}(n-2)!} = (2n-5)(2n-7)(2n-9)\cdots 3 \cdot 1$, which grows super exponentially in n. Hence on sets of taxa of meaningful size it is not usually feasible to consider every possible tree topology. This leads to the second principle difficulty: once we have chosen a function ϕ, how can we rapidly compute the maximizing tree? Answering this algorithmic question is essential if the theory is to be applied to real-world data. In practice it is usually impossible to find the maximizing tree in polynomial time, so we often satisfy ourselves with approximation algorithms. Such algorithms in turn require their own theoretical results guaranteeing their accuracy and effectiveness, and in Chapter 3 we discuss the implications of our theoretical results for the accuracy of approximating algorithms.

To quantify their accuracy, we need a few definitions. A *distance-based reconstruction algorithm* is an algorithm that takes a dissimilarity as input and produces an unrooted tree topology as output. Throughout, we use the word "algorithm" and "method" interchangeably.

Definition 1.3.3. A reconstruction method is *consistent* if, when given a T-additive dissimilarity as input, it returns T.

The term *consistent* refers to statistical consistency: As more data is gathered, we expect the observed pairwise distances D to converge to the true distances $\hat{D}$. The statement that

ϕ is consistent is precisely the claim that we recover the correct tree topology T in the limit when $D = \hat{D}$ is T-additive.

Consistency is a weak guarantee, and we would like to know how effective our reconstruction method is when our data is not tree-additive. The l_∞ distance between two dissimilarities $D, \hat{D}$ is defined to be

$$||D - \hat{D}||_\infty = \max_{i,j} |D_{ij} - \hat{D}_{ij}|.$$

Definition 1.3.4. A tree reconstruction method has l_∞ radius α if, for each T-additive dissimilarity $\hat{D} = \sum_{e \in P_{ij}^T} w_e \sigma_e$ with minimum branch length $w_{\min} = \min_e w_e$, and dissimilarity D with $||D - \hat{D}||_\infty < \alpha w_{min}$, the method returns T when given D as input.

In the literature, this is also known as the method's *safety radius* [1]. For example, let

$$D = \begin{pmatrix} 0 & 6.4 & 12.8 & 8 \\ 6.4 & 0 & 14.3 & 11.3 \\ 12.8 & 14.3 & 0 & 8.5 \\ 8 & 11.3 & 8.5 & 0 \end{pmatrix}$$

Then $||D - \hat{D}||_\infty = 0.8$, where $\hat{D}$ is the dissimilarity from Figure 1.3. Since the minimum branch length of the tree underlying $\hat{D}$ is 2, a reconstruction algorithm with l_∞ radius greater than $\frac{0.8}{2}$ will return T when given D as input.

Any method with positive l_∞ radius is necessarily statistically consistent. No reconstruction method can have l_∞ radius greater than $\frac{1}{2}$. This is because there is a dissimilarity D and distinct trees T_1, T_2 with corresponding tree-additive dissimilarities $\hat{D}_1, \hat{D}_2$ and identical minimum branch length $w_{\min}$ such that $||D - \hat{D}_1||_\infty = ||D - \hat{D}_2||_\infty = \frac{1}{2} w_{\min}$. For example,

$$\hat{D}_1 = \begin{pmatrix} 0 & 4 & 6 & 6 \\ 4 & 0 & 6 & 6 \\ 6 & 6 & 0 & 4 \\ 6 & 6 & 4 & 0 \end{pmatrix}, \quad \hat{D}_2 = \begin{pmatrix} 0 & 6 & 4 & 6 \\ 6 & 0 & 6 & 4 \\ 4 & 6 & 0 & 6 \\ 6 & 4 & 6 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 5 & 5 & 6 \\ 5 & 0 & 6 & 5 \\ 5 & 6 & 0 & 5 \\ 6 & 5 & 5 & 0 \end{pmatrix}.$$

If a reconstruction algorithm had l_∞ radius $> \frac{1}{2}$, upon receiving D as input it would have to return both T_1 and T_2, a contradiction.

In Chapter 3, we will examine a class of reconstruction methods known as *minimum evolution* methods which are linear in the elements of D. Our main theorem is that there is a unique method with l_∞ radius $\frac{1}{2}$. This maximally robust method, known as *balanced minimum evolution (BME)*, was discovered by Pauplin 2000 [58]. It is now known [36] to be the theoretical underpinning behind the neighbor-joining algorithm [62], one of the most widely-used distance-based reconstruction algorithms of all time. Although the relationship between accuracy guarantees made by a greedy algorithm and its theoretical counterpart are not well understood, our result strongly suggests that any greedy distance-based algorithm should be based upon BME.

In the final part of Chapter 3, we investigate the minimum evolution reconstruction problem for Kalmanson metrics. We show that the traveling salesman linear form is the only consistent method, and prove it has l_∞ radius $\frac{n-3}{2}$ on n taxa. We also define generalizations of the BME and traveling salesman linear form and investigate their robustness.

Chapter 2

Affine and Projective Tree Metric Theorems

This chapter is based on joint work with Lior Pachter and Matan Harel that is scheduled to appear in [48].

2.1 Introduction

As discussed in the introduction, dissimilarity maps derived from data are rarely exact tree metrics for two reasons. First, some evolutionary mechanisms may not be realizable on trees. Second, even when evolution occurs in a tree-like fashion, noise and high variance can produce signals that conflict with those from the underlying topology. In these cases, fitting the data to a tree might result in losing valuable information.

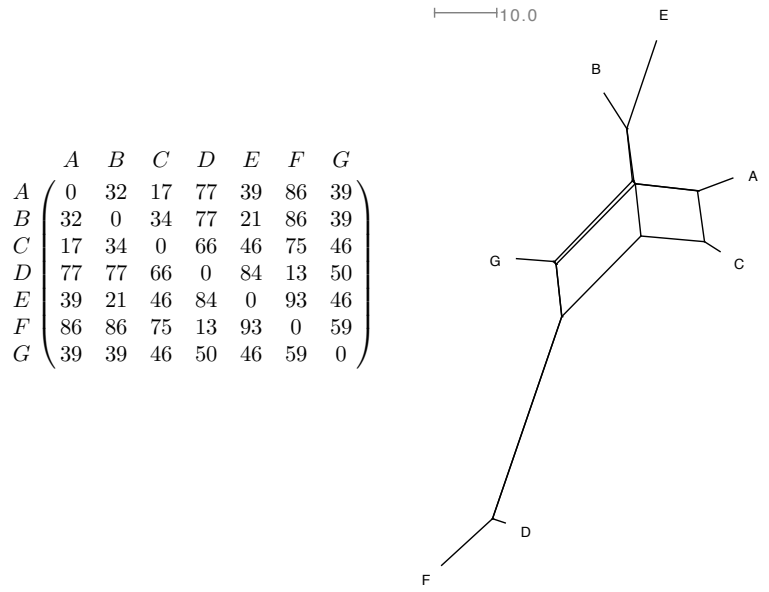
Trees can be defined by pairwise compatible split systems, so it is natural to generalize them by considering other collections of splits. One natural class to consider is the following.

Definition 2.1.1. A *circular ordering* $\mathcal{C} = \{x_1, \ldots, x_n\}$ is a bijection between X and the vertices of a convex n-gon P_n such that x_i and x_{i+1} map to adjacent vertices of P_n (where $x_{n+1} := x_1$). Let $S_{i,j}$ denote the split $\{x_i, x_{i+1}, \ldots, x_{j-1}\}|\{x_j, x_{j+1}, \ldots, x_{i-1}\}$ and let $\mathcal{S}(\mathcal{C}) = \{S_{i,j}|i < j\}$. We say a split S is *circular* with respect to a circular ordering $\mathcal{C}$ if $S \in \mathcal{S}(\mathcal{C})$, and a split system $\mathcal{S}$ is circular if $\mathcal{S} \subseteq \mathcal{S}(\mathcal{C})$ for some circular ordering $\mathcal{C}$.

Consider a planar embedding of an X-tree. Reading the taxa clockwise gives a circular ordering $\mathcal{C}$, and every split $S \in \mathcal{S}(T)$ is compatible with $\mathcal{C}$. The splits of T are thus compatible with each of the 2^{n-2} circular orderings of T. Given a set $\mathcal{E}$ of circular orderings, let $\mathcal{S}(\mathcal{E}) = \cap_{\mathcal{C}\in\mathcal{E}}\mathcal{S}(\mathcal{C})$ be the system of splits that are circular with respect to each ordering in $\mathcal{E}$. The split system associated with a trivalent X-tree arises in this way from a family $\mathcal{E}$ of 2^{n-2} circular orderings [63], and a tree metric is obtained by associating non-negative weights to each split in the system. *Kalmanson metrics*, which were first introduced in the

study of traveling salesmen problems where they provide a class of metrics for which the optimal tour can be identified in polynomial time [45], correspond to the case when $|\mathcal{E}| = 1$.

Kalmanson metrics can be visualized using *split networks* [43]. We do not provide a definition, but show an example in Figure 2.1 (drawn using the software SplitsTree4). The



$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{ccccccc} A & B & C & D & E & F & G \end{array} \\ \begin{pmatrix} 0 & 32 & 17 & 77 & 39 & 86 & 39 \\ 32 & 0 & 34 & 77 & 21 & 86 & 39 \\ 17 & 34 & 0 & 66 & 46 & 75 & 46 \\ 77 & 77 & 66 & 0 & 84 & 13 & 50 \\ 39 & 21 & 46 & 84 & 0 & 93 & 46 \\ 86 & 86 & 75 & 13 & 93 & 0 & 59 \\ 39 & 39 & 46 & 50 & 46 & 59 & 0 \end{pmatrix} \end{array}$$

Figure 2.1: A Kalmanson metric (left) visualized as a split network (right).

neighbor-net [8] and MC-net [28] algorithms provide a way to construct circular split systems from dissimilarity maps, but despite having a number of useful properties [9, 52], they have not been widely adopted in the phylogenetics community. This is likely because split networks (such as in Figure 2.1) fail to reveal the "treeness" of the data. More specifically, the internal nodes of the split network do not correspond to meaningful "ancestors" as do the internal nodes in X-trees. Other approaches to visualizing "treeness," e.g. [72], do reveal the extent of signal conflicting with a tree in data, but do not produce the splits underlying the discordance.

We propose that PQ- and PC-trees, first developed in the context of the consecutive ones problem [5, 41] and for graph planarity testing [40, 65], are convenient structures that interpolate between X-trees and circular split systems. The main result of this chapter is Theorem 2.4.4. Given a Kalmanson metric, the theorem shows how to construct a "best fit" PC-tree which realizes the metric and captures the"treeness" of it.

Another (expository) goal of this chapter is to organize, for the first time, existing results on PC-trees, their cousins PQ-trees, and corresponding metrics (Theorem 2.4.16 in Section 2.4) into a single unified framework. As a prelude, we illustrate the types of results

we derive using a classic theorem relating rooted X-trees to special set systems that encode information about shared ancestry.

Definition 2.1.2. A *hierarchy* $\mathcal{H}$ *over a set* X *is a collection of subsets of* X *such that:*

(i) $X \in \mathcal{H}$*, and* $\{x\} \in \mathcal{H}$ *for all* $x \in X$*,*

(ii) $A \cap B \in \{\emptyset, A, B\}$ *for all* $A, B \in \mathcal{H}$*.*

The requirement that each $\{x\} \in \mathcal{H}$ is not part of the usual definition of hierarchy but its inclusion here will simplify the statements of later results.

Proposition 2.1.3. *There is a natural bijection between hierarchies over* X *and rooted phylogenetic* X*-trees.*

Proposition 2.1.3, which we will prove constructively in Section 2, is an elementary but classic result and has been discovered repeatedly in a variety of contexts [25, 38]. For example, in computer science, hierarchies are known as *laminar families* where they play an important role in the development of recursive algorithms represented by rooted trees [29]. Hierarchies are also important because they are the combinatorial structures that underlie *ultrametrics*.

Definition 2.1.4. An *ultrametric* is a symmetric function $D : X \times X \to \mathbb{R}$ such that

$$D(x, y) \leq \max\{D(x, z), D(y, z)\} \quad \forall x, y, z \in X.$$

Definition 2.1.5. An *indexed hierarchy* is a hierarchy $\mathcal{H}$ with a non-negative function $f : \mathcal{H} \to \mathbb{R}_{\geq 0}$ such that for all $A, B \in \mathcal{H}$, $A \subset B \Rightarrow f(A) \leq f(B)$.

The extension of Proposition 2.1.3 to metrics, proved in [44], exhibits a bijection between ultrametrics and inexed hierarchies. The proposition is an instance of a *tree metric theorem* that associates a class of combinatorial objects (in this case, rooted X-trees) with a class of metrics (in this case, ultrametrics). Our results organize other tree metric theorems that have been discovered (in some cases independently and multiple times) in the contexts of biology, mathematics and computer science.

In particular, we investigate relaxations of Definitions 1.1.1 and 2.1.2 for which there exist analogies of Theorem 1.3.2 and Proposition 2.1.3. For example, hierarchies are special cases of *pyramids* [20], which can be indexed to produce *strong Robinsonian matrices* [59]. Proposition 2.4.10 (originally proved in [20]) states that these objects correspond to each other mimicking the correspondence between hierarchies and ultrametrics.

In discussing tree metric theorems we adopt the nomenclature of Andreas Dress who distinguishes two types of objects and theorems: the *affine* and the *projective* [21]. Roughly speaking, these correspond to "rooted" and "unrooted" statements respectively, and we use these terms interchangeably. At every step we provide maps for transitioning between the two

worlds. For example, a hierarchy is an affine concept whose projective analog is a pairwise compatible split system. Similarly, unrooted X-trees are the projective equivalents of rooted X-trees, and tree metrics are the projective equivalents of ultrametrics. We'll see that *Kalmanson metrics* are to tree metrics as *Robinsonian matrices* [59] are to ultrametrics, and circular split systems are to pairwise compatible split systems as pyramids are to hierarchies. We use PQ-trees [5] and their projective analogs PC-trees [65] to link all of these results. This will also illustrate the general fact (which will show up again in Chapter 4) that when considering a question about unrooted trees, it is often fruitful to examine a rooted version.

We begin by proving Proposition 2.1.3, both for completeness and to introduce some of the notation that we use. A rooted X-tree $\mathcal{T} = (V, E)$ has a natural partial ordering on its vertices: For $u, v \in V$, we say $u \preccurlyeq v$ if $u = v$ or if v lies on the unique path from u to the root. Given $v \in V$, let $H_v = \{x \in X | x \preccurlyeq v\}$ and $\alpha(\mathcal{T}) = \{H_v | v \in V\}$.

Proposition 2.1.6. *The map α is a bijection from rooted phylogenetic X-trees to hierarchies over X.*

Proof. Let (T, ϕ) be a rooted phylogenetic X-tree with root r. $H_r = X$ and $H_{\phi(x)} = \{x\}$ for all $x \in X$, so $\mathcal{H}$ satisfies (1) of Definition 2.1.2. Consider any two $H_u, H_v \in \alpha(\mathcal{T})$. If $u \preccurlyeq v$ then $H_u \cap H_v = H_u$, if $v \preccurlyeq u$ then $H_u \cap H_v = H_v$, and otherwise $H_u \cap H_v = \emptyset$. So each pair of elements in $\alpha(\mathcal{T})$ satisfies (2) of Definition 2.1.2, and $\alpha(\mathcal{T})$ is a hierarchy.

For the reverse direction, let $\mathcal{H}$ be a hierarchy over X. Let $T = (V, E)$ be the digraph with $V = \{v_A | A \in \mathcal{H}\}$ and with edges denoting minimal inclusion: T has an edge from v_B to v_A if and only if $A \subset B$ and there does not exist $C \in \mathcal{H}$ such that $A \subsetneq C \subsetneq B$. We will show that T is a tree. First note that by downward induction on $|C|$ each vertex v_C with $C \neq X$ is connected to v_X and has at least one parent. Now suppose v_A, v_B are distinct parents of v_C. Then $A \cap B \neq \emptyset$, so without loss of generality by the hierarchy condition $A \subset B$. But then $C \subset A \subset B$, a contradiction. Thus T is connected and has one fewer edge than vertices, so T is a tree with root v_X. Define the map $\phi : X \to V$ by $\phi(x) = v_x$. This is a bijection from X to the leaves, so $\mathcal{T} = (T, \phi)$ is a rooted X-tree with $\alpha(\mathcal{T}) = \mathcal{H}$. □

The above proposition gives a characterization of rooted X-trees in terms of collections of subsets of X. We turn now to the projective analogue of rooted X-trees. Recall removing an edge e from a projective X-tree gives an X-split S_e, and let $\beta(\mathcal{T}) = \{S_e | e \in E(T)\}$. Then Theorem 1.2.2 shows β is a bijection from the set of projective X-trees to the set of pairwise compatible split systems over X.

Finally we show that pairwise compatible split systems and hierarchies are in bijection. Fix $r \in X$ and let $\mathcal{S}$ be a set of pairwise compatible splits of X. The *unrooting map* γ_r sends a split S to the component of S that does not contain r. The *rooting map* δ_r sends a set $A \subseteq X \setminus \{r\}$ to the split $\delta_r(A) = A | X \setminus A$. If $\mathcal{S}$ is a split system, let $\gamma_r(\mathcal{S}) = \{\gamma_r(S) | S \in \mathcal{S}\}$.

Proposition 2.1.7. *$\mathcal{S}$ is a pairwise compatible split system over X if and only if $\gamma_r(\mathcal{S})$ is a hierarchy over $X \setminus \{r\}$.*

Proof. Choose $S_1, S_2 \in \mathcal{S}$, with $S_i = A_i|B_i$. If S_1, S_2 are compatible, then without loss of generality $A_1 \cap A_2 = \emptyset$. If $r \in B_1, B_2$, then $\gamma_r(S_i) = A_i$, so $\gamma_r(S_1) \cap \gamma_r(S_2) = \emptyset$. If $r \in A_1$, then $r \notin A_2$ since $A_1 \cap A_2 = \emptyset$, and therefore $r \in B_2$. In this case $\gamma_r(S_1) = X \setminus A_1$ and $\gamma_r(S_2) = A_2$, so $\gamma_r(S_1) \cap \gamma_r(S_2) = A_2 = \gamma_r(S_2)$, again satisfying the hierarchy condition. The final case follows by symmetry. Conversely, suppose $\gamma_r(\mathcal{S})$ is a hierarchy. Then for any two $S_1, S_2 \in \mathcal{S}$, we have $\gamma_r(S_1) \cap \gamma_r(S_2) \in \{\emptyset, \gamma_r(S_1), \gamma_r(S_2)\}$. Checking the cases as above shows that S_1 and S_2 must then be compatible splits. □

Define the map κ_r to take an affine $X \setminus \{r\}$-tree $\mathcal{T}$ to the projective X-tree obtained by attaching a vertex with label r to the root of $\mathcal{T}$. The inverse map λ_r takes a projective X-tree $\mathcal{T}$ to the affine X-tree as follows: Let v be the vertex of $\mathcal{T}$ labeled r. Then $\lambda_r(\mathcal{T})$ is obtained by rooting at the neighbor of v, and then deleting v.

Proposition 2.1.8. *If AT and H are the sets of all affine trees and hierarchies over $X \setminus \{r\}$, respectively, and PT and PSS are the sets of all projective trees and pairwise compatible splits systems over X, respectively, then the following diagram commutes:*

$$\begin{array}{ccc} AT & \underset{\lambda_r}{\overset{\kappa_r}{\rightleftarrows}} & PT \\ \uparrow\downarrow{\scriptstyle \alpha} & & \uparrow\downarrow{\scriptstyle \beta} \\ H & \underset{\gamma_r}{\overset{\delta_r}{\rightleftarrows}} & PSS \end{array}$$

Each arrow is a bijection; the unlabeled arrows are the inverses of the maps going in the other direction.

2.2 PQ-trees

We start our generalization of Proposition 2.1.8 with a generalization of rooted X-trees.

Definition 2.2.1. A *PQ-tree* over X is a rooted phylogenetic X-tree in which every vertex comes equipped with a linear ordering on its children. Every internal vertex of degree 3 or less is labeled a P-vertex, and every internal vertex of degree 4 or more is labeled either as a P-vertex or a Q-vertex. We say two PQ-trees $\mathcal{T}_1, \mathcal{T}_2$ are equivalent (we write $\mathcal{T}_1 \sim \mathcal{T}_2$) if one can be obtained from the other by a series of moves consisting of:

(i) Permuting the ordering on the children of a P-vertex,

(ii) Reversing the ordering on the children of a Q-vertex.

We draw a PQ-tree by representing P-vertices as circles and Q-vertices as squares, and ordering the children of a vertex from left to right as per the corresponding linear order (see Figure 2.2). For any PQ-tree $\mathcal{T}$ over X, define the *frontier* of the tree as the linear ordering on

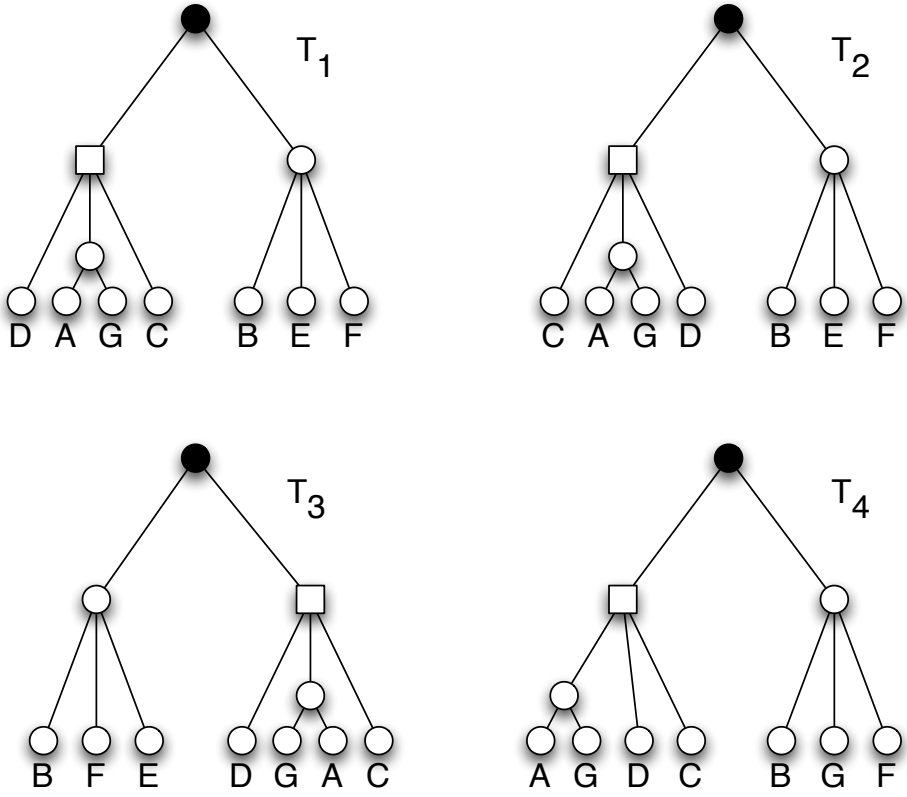


Figure 2.2: Four PQ trees. T_1, T_2 and T_3 are equivalent to each other but T_4 is different from the other three.

X derived from reading the leaves of T from left to right. Let $con(\mathcal{T}) = \{frontier(\mathcal{T}')|\mathcal{T}' \sim \mathcal{T}\}$ be the set of all linear orderings $\prec$ that are consistent with the PQ structure of $\mathcal{T}$. We say I is an *interval* with respect to $\prec$ if there exist $a, b \in X$ such that $I = \{t|a \preceq t \preceq b\}$. Define α to be the map that sends the PQ-tree $\mathcal{T}$ to the set of all $I \subseteq X$ such that I is an interval with respect to every linear ordering in $con(\mathcal{T})$.

Lemma 2.2.2. *$\alpha(\mathcal{T})$ is a hierarchy if and only if every vertex of $\mathcal{T}$ is a P-vertex. If so, let $\mathcal{T}'$ be the corresponding normal affine X-tree. Then $\alpha(\mathcal{T}) = \alpha(\mathcal{T}')$, where $\alpha(\mathcal{T}')$ is the hierarchy constructed in Proposition 2.1.6.*

Proof. Let v be an internal vertex of T, $\{c_1, c_2, \ldots, c_n\}$ the set of its children, and recall $H_v = \{x \in X|v_x \preceq v\}$ is the set of all the elements x such that the path from v_x to the root includes v. Now $H_{c_1} \cup \cdots \cup H_{c_n}$ is in $\alpha(\mathcal{T})$, and if every vertex of T is a P-vertex then every element of $\alpha(\mathcal{T})$ will be of this form. In this case $\alpha(\mathcal{T})$ is identical to the hierarchy constructed in Proposition 2.1.6. Now suppose $\mathcal{T}$ has a Q-vertex v. Then $n \geq 3$ and both

$A = H_{c_1} \cup H_{c_2}$ and $B = H_{c_2} \cup H_{c_3}$ are in $\alpha(\mathcal{T})$, but $A \cap B = H_{c_2}$ is nonempty, so $\alpha(\mathcal{T})$ is not a hierarchy. □

This shows that the map α on PQ-trees agrees with the α in Proposition 2.1.8. It also shows that usual affine X-trees are precisely PQ-trees with all P-vertices. Since PQ-trees do not necessarily give rise to hierarchies, we seek a different combinatorial characterization of them.

Definition 2.2.3. A collection of subsets $\mathcal{P}$ of X is a *prepyramid* if

(i) $X \in \mathcal{P}$ and $\{x\} \in \mathcal{P}$ for all $x \in X$,

(ii) There exists a linear ordering $\prec$ on X such that every $A \in \mathcal{P}$ is an interval with respect to $\prec$.

$\mathcal{P}$ is a *pyramid* if, in addition, it is closed under intersection.

If $\mathcal{T}$ is a PQ-tree then $\alpha(\mathcal{T})$ is a prepyramid with respect to any $\prec \in frontier(\mathcal{T})$.

Definition 2.2.4. Two subsets A, B of X are *compatible* if $A \cap B \in \{\emptyset, A, B\}$. Otherwise they are *incompatible*. A *rooted family* over X is a collection of sets $\mathcal{F}$ such that if $A, B \in \mathcal{F}$ are incompatible, then $A \cap B, A \setminus B, B \setminus A$ and $A \cup B$ are in $\mathcal{F}$.

We are now ready to state the main result of this section.

Proposition 2.2.5. *The map α is a bijection from PQ-trees to prepyramids that are rooted families.*

Proof. If $\mathcal{T}$ is a PQ-tree, $\alpha(\mathcal{T})$ is obviously a prepyramid. We now show it is also a rooted family. Trivially $X \in \alpha(\mathcal{T})$ and $\{x\} \in \alpha(\mathcal{T})$ for all $x \in X$. Let A, B be incompatible sets in $\mathcal{F} := \alpha(\mathcal{T})$ and $\prec \in con(A)$ a linear ordering. We can write $A = \{t | x_A \preceq t \preceq y_A\}$ and $B = \{t | x_B \preceq t \preceq y_B\}$ for some $x_A, x_B, y_A, y_B \in X$, and by incompatibility we can assume $x_A \prec x_B \preceq y_A \prec y_B$. Then $A \cap B = \{t | x_B \preceq t \preceq y_A\}$, $A \cup B = \{t | x_A \preceq t \preceq y_B\}$, $A \setminus B = \{t | x_A \preceq t \prec x_B\}$ and $B \setminus A = \{t | y_A \prec t \preceq y_B\}$. Since each of these four sets is an interval with respect to $\prec$ for every $\prec \in con(\mathcal{T})$ they are each in $\alpha(\mathcal{T})$.

It remains to show that if $\mathcal{F}$ is a collection of subsets of X which is a prepyramid and a rooted family, then there exists a unique PQ-tree $\mathcal{T}$ such that $\alpha(\mathcal{T}) = \mathcal{F}$. Let $\mathcal{F}' \subseteq \mathcal{F}$ consist of the sets in $\mathcal{F}$ that are compatible with all of $\mathcal{F}$. Then the elements of $\mathcal{F}'$ are pairwise compatible, so $\mathcal{F}'$ is a hierarchy and corresponds to a tree $\mathcal{T}'$, as in Proposition 2.1.6. Consider the vertices v_C in $\mathcal{T}'$ corresponding to subsets of the form $C = A \cup B$ with $A, B \in \mathcal{F}$ incompatible, and mark those vertices as Q-vertices. For each such v_C there is a natural ordering on its children $v_{C_1}, v_{C_2}, \ldots, v_{C_n}$ as follows: $v_{C_i} < v_{C_j}$ if the labels of the leaves in the subtree rooted at v_{C_i} are all $\prec$ the labels of the leaves of the subtree rooted at v_{C_j}, where $\prec$ is the order from the prepyramid condition. Ordering the children of the Q-vertices of T' in this way, we obtain a PQ-tree $\mathcal{T}$.

We will use induction on $|\mathcal{F}|$ to show $\alpha(\mathcal{T}) = \mathcal{F}$ and that $\mathcal{T}$ is the only such PQ-tree for which this is true. First, suppose $\mathcal{F}'$ contains a set $A \neq X$, $|A| > 1$. Define $\mathcal{F}_1 = \{C \in \mathcal{F} | A \subseteq C \text{ or } A \cap C = \emptyset\}$ and $\mathcal{F}_2 = \{C \in \mathcal{F} | A \supseteq C\}$. $\mathcal{F}_1 \cap \mathcal{F}_2 = A$, and because A is compatible with everything, $\mathcal{F}_1 \cup \mathcal{F}_2 = \mathcal{F}$. Now $\mathcal{F}_1$ is a prepyramid over $(X \setminus A) \cup \{A\}$ and $\mathcal{F}_2$ is a prepyramid over A, and both are rooted families with $|\mathcal{F}_i| < |\mathcal{F}|$. Then the inductive hypothesis shows there are PQ-trees $\mathcal{T}_1, \mathcal{T}_2$ such that $\alpha(\mathcal{T}_i) = \mathcal{F}_i$ for $i = 1, 2$. Now $\mathcal{T}_1$ has a leaf corresponding to A, and the root of $\mathcal{T}_2$ also corresponds to A. Graft $\mathcal{T}_2$ onto $\mathcal{T}_1$ by identifying the root of $\mathcal{T}_2$ with the leaf A in $\mathcal{T}_1$; this gives a PQ-tree $\mathcal{T}$ with $\alpha(\mathcal{T}) = \alpha(\mathcal{T}_1) \cup \alpha(\mathcal{T}_2) = \mathcal{F}$. Conversely, if $\mathcal{T}$ is a PQ-tree with $\alpha(\mathcal{T}) = \mathcal{F}$, $\mathcal{T}$ must have a node v such that the subtree $\mathcal{T}_1$ with root v satisfies $\alpha(\mathcal{T}_1) = \mathcal{F}_1$, and the tree $\mathcal{T}_2$ obtained by replacing the subtree $\mathcal{T}_1$ with the vertex v and label A gives $\alpha(\mathcal{T}_2) = \mathcal{F}_2$. The inductive hypothesis gives the uniqueness of $\mathcal{T}_1$ and $\mathcal{T}_2$, which in turn implies the uniqueness of $\mathcal{T}$.

Now suppose no such set A exists, so $\mathcal{F}' = \{\{x_1\}, \ldots, \{x_n\}, X\}$. If $\mathcal{F} = \mathcal{F}'$ then $\mathcal{F}$ is a hierarchy, and the PQ-tree of depth one whose root is a P-vertex is the unique tree such that $\alpha(\mathcal{T}) = \mathcal{F}$. We now consider the final case, where $\mathcal{F}$ contains sets other than X and the $\{x_i\}$, and every element in $\mathcal{F}$ except for these is incompatible with another element of $\mathcal{F}$. Without loss of generality assume $x_1 \prec x_2 \prec \cdots \prec x_n$, where $\prec$ is the ordering under which $\mathcal{F}$ is a prepyramid. Then every element of $\mathcal{F}$ is of the form $x_{[i:j]} := \{x_i, x_{i+1}, \cdots, x_j\}$ for $1 \leq i \leq j \leq n$. Let $\mathcal{T}$ be the PQ-tree of depth 1 with a Q-vertex root and children x_i. Since $\alpha(\mathcal{T}) = \{x_{[i:j]} | 1 \leq i \leq j \leq n\}$ it is clear $\alpha(\mathcal{T}) \supseteq \mathcal{F}$.

We must show $\mathcal{F} = \alpha(\mathcal{T})$, or equivalently, that $x_{[i:j]} \in \mathcal{F}$ for all $i < j$. We will prove this by induction on n. It's obvious for $n = 3$, so suppose $n \geq 4$. Let $A \neq X$ be a set in $\mathcal{F}$ that is maximal under inclusion. By assumption $|A| > 1$ and A is incompatible with some $B \in \mathcal{F}$. By the rooted family condition $A \cup B \in \mathcal{F}$ so we must have $A \cup B = X$ by the maximality of A. Without loss of generality we may write $A = x_{[1:j]}$, $B = x_{[i:n]}$ with $i \leq j$. We claim $j = n-1$. For by the rooted family condition, $\mathcal{F}$ contains $A \setminus B = x_{[1:i-1]}$, $A \cap B = x_{[i:j]}$ and $B \setminus A = x_{[j+1:n]}$. If $j \neq n-1$, $B \setminus A$ is incompatible with some other set $C \in \mathcal{F}$. C must be of the form $C = x_{[k_1:k_2]}$ with $k_1 < j + 1 \leq k_2 < n$. If $k_1 = 1$ then $C \supset A$ contradicting the maximality of A. Then $k_1 \neq 1$, so A and C are incompatible and $A \cup C = x_{[1:k_2]} \in \mathcal{F}$, contradicting the maximality of A. Thus $j = n-1$ and $x_{[1:n-1]} \in \mathcal{F}$. By the same reasoning, $x_{[2:n]} \in \mathcal{F}$.

Now let $\mathcal{G} = \{A \in \mathcal{F} | A \subseteq x_{[1:n-1]}\}$. Since $x_{[1:n-1]} \in \mathcal{G}$, $\mathcal{G}$ is a prepyramid over $X \setminus \{x_n\}$. It is also a rooted family. We claim that every element $A \in \mathcal{G}$ with $A \neq \{x_i\}, A \neq X \setminus \{x_n\}$ is incompatible with some other element of $\mathcal{G}$. To see this, write $A = x_{[i:j]}$, $j \leq n-1$ and suppose $j \neq n-1$. By our initial assumption, A is incompatible with some set $B = x_{[i':j']} \in \mathcal{F}$. If $j' \neq n$ then $B \in \mathcal{G}$ gives the incompatible set. Otherwise B and $x_{[1:n-1]}$ are incompatible, so $B \cap x_{[1:n-1]} = x_{[i':n-1]}$ is in $\mathcal{G}$ and is incompatible with A. Finally, suppose $j = n-1$. Then A is incompatible with $x_{[n-1:n]}$, so $A \cup x_{[n-1:n]} = x_{[i:n]} \in \mathcal{F}$. This must be incompatible with some other set $B \in \mathcal{F}$. We can write $B = x_{[i':j']}$, $i \leq j' \leq n-1$, so $B \in \mathcal{G}$ is incompatible with A and the claim is proved.

It follows that $\mathcal{G}$ is a rooted prepyramid over $X \setminus \{x_n\}$, and every element in $\mathcal{G}$ other

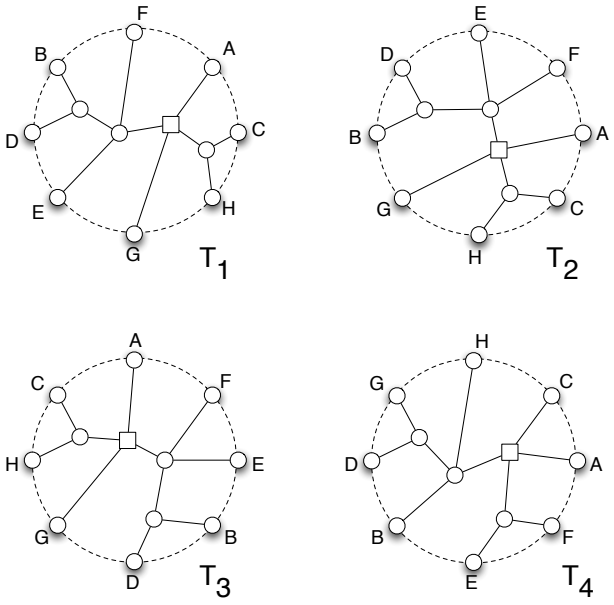


Figure 2.3: Four PC-trees. T_1, T_2 and T_3 are equivalent to each other but T_4 is different from the other three.

than X and the $\{x_i\}$ is incompatible with some other element of $\mathcal{G}$. By our inductive hypothesis, $x_{[i:j]} \in \mathcal{G}$ for each $1 \leq i \leq j \leq n-1$. We showed that $x_{[n-1:n]} \in \mathcal{F}$, and for each i the sets $x_{[i:n-1]}$ and $x_{[n-1:n]}$ are incompatible, so their union $x_{[i:n]}$ is in $\mathcal{F}$. Thus $\mathcal{F} = \{x_{[i:j]} | 1 \leq i \leq j \leq n\}$. □

2.3 PC-trees

In this section we develop the theory behind the projective equivalent of PQ-trees, characterizing them both geometrically and combinatorially.

Definition 2.3.1. A *PC-tree* over X is a phylogenetic X-tree where each internal vertex comes equipped with a circular ordering of its neighbors. Each internal vertex of degree less than 4 is labeled a P-vertex, and each other vertex is labeled either a P-vertex or a C-vertex (Figure 2.3). Two PC-trees $\mathcal{T}_1, \mathcal{T}_2$ are said to be *equivalent* (we write $\mathcal{T}_1 \sim \mathcal{T}_2$) if one can be obtained from the other by a series of the following moves:

(i) Permuting the circular ordering of the neighbors of a P-vertex,

(ii) Reversing the circular ordering on the neighbors of a C-vertex.

For a PC-tree $\mathcal{T}$, let *frontier*($\mathcal{T}$) be the circular ordering given by reading the taxa in either a clockwise or counterclockwise direction. Let $con(\mathcal{T}) = \{frontier(\mathcal{T}')|\mathcal{T}' \sim \mathcal{T}\}$ and let $\beta(\mathcal{T}) = \cap_{\mathcal{C}\in con(\mathcal{T})}\mathcal{S}(\mathcal{C})$ be the set of splits compatible with every circular ordering of $\mathcal{T}$.

Definition 2.3.2. A split system $\mathcal{S}$ is an *unrooted split family over* X if, for each pair of incompatible splits $S_1 = A_1|B_1, S_2 = A_2|B_2$ in $\mathcal{S}$, the splits $A_1 \cap A_2|B_1 \cup B_2, A_1 \cap B_2|A_2 \cup B_1, A_2 \cap B_1|A_1 \cup B_2$, and $B_1 \cap B_2|A_1 \cup A_2$ are all in $\mathcal{S}$ as well.

This is just the projective analog of a rooted split family (Definition 2.2.4), and trivially $\gamma_r(\mathcal{S})$ is a rooted family if and only if $\mathcal{S}$ is an unrooted split family.

Following the proof in Proposition 2.2.5 that $\alpha(\mathcal{T})$ is a rooted family for every PQ-tree $\mathcal{T}$, one can show $\beta(\mathcal{T})$ is an unrooted family for every PC-tree $\mathcal{T}$. We now generalize the map γ_r to PQ- and PC-trees.

Definition 2.3.3. The *unrooting map* κ_r sends a PQ-tree $\mathcal{T}$ over $X \setminus \{r\}$ to the PC-tree $\kappa_r(\mathcal{T})$ as follows: Attach a vertex labeled r to the root of $\mathcal{T}$. If vertex v in $\mathcal{T}$ has children $\{v_1, v_2, \ldots, v_k\}$ with linear ordering $v_1 \prec v_2 \prec \ldots \prec v_k$ and parent w, in $\kappa_r(\mathcal{T})$ the vertex has the same neighbors with circular ordering $\{v_1, \ldots, v_k, w\}$.

The *rooting map* λ_r sends a PC-tree $\mathcal{T}$ over X to the PQ-tree over $X \setminus \{r\}$ obtained by rooting at the vertex adjacent to r, deleting r, and replacing each C vertex with a Q vertex. Let v be such a vertex with a circular ordering $\mathcal{C} = \{v_1, \ldots, v_m\}$; we may assume the path from v to the root passes through v_m. Then in $\lambda_r(\mathcal{T})$, vertex v has parent v_m and children $v_1, \ldots, v_{m-1}$ with linear ordering $v_1 \prec v_2 \prec \ldots \prec v_{m-1}$.

Since κ_r and λ_r are inverses, this immediately gives the following:

Proposition 2.3.4. *κ_r is a bijection from PQ-trees over $X \setminus \{r\}$ to PC-trees over X.*

Recall that if S is a split of X, the map γ_r takes S to the component of S not containing r.

Proposition 2.3.5. *Let CUF be the set of circular split systems that are unrooted families over X, and let PRF be the set of prepyramids that are rooted families over $X \setminus \{r\}$. Then the map γ_r is a bijection from CUF to PRF.*

Proof. δ_r and γ_r are inverses, so it suffices to show that $\gamma_r(CUF) \subseteq PRF$ and $\delta_r(PRF) \subseteq CUF$. Let $\mathcal{S}$ be a circular unrooted split family with circular ordering $\{x_1, \ldots, x_n\}$ and suppose $r = x_i$. Then $\gamma_r(\mathcal{S})$ is a prepyramid with respect to the linear ordering on $X \setminus \{r\}$ given by $x_{i+1} \prec x_{i+2} \prec \cdots \prec x_{i-1}$.

Next, consider $\mathcal{S} \in CUF$ and two incompatible sets $G, H \in \gamma_r(\mathcal{S})$, with $G = \gamma_r(S_1)$ and $H = \gamma_r(S_2)$. Assume $S_1 = A_1|B_1$, $S_2 = A_2|B_2$ are compatible with $A_1 \cap A_2 = \emptyset$. If $r \in A_1$, then $r \in B_2$ and $G \cap H = (X \setminus A_1) \cap A_2 = A_2 = H$, contradicting the incompatibility of G and H. The other cases produce similar contradictions, so S_1 and S_2 must be incompatible. Thus the splits $A_1 \cap A_2|B_1 \cup B_2, A_1 \cap B_2|A_2 \cup B_1, A_2 \cap B_1|A_1 \cup B_2$ and $B_1 \cap B_2|A_1 \cup A_2$ are all in $\mathcal{S}$. Assume without loss of generality that $G = A_1, H = A_2$. Then $\gamma_r(\{A_1 \cap A_2|B_1 \cup B_2\}) =$

$A_1 \cap A_2 = G \cap H \in \gamma_r(\mathcal{S})$, and similarly $G \cup H, G \setminus H, H \setminus G$ are all in $\gamma_r(\mathcal{S})$, so $\gamma_r(\mathcal{S})$ is a rooted family.

Conversely, let $\mathcal{F}$ be a prepyramidal rooted family over $X \setminus \{r\}$ with linear ordering $\prec$. $\delta_r(\mathcal{F})$ is circular and contains all the trivial splits, as $\{x\}|X \setminus \{x\} = \gamma_r(\{x\})$ and $\{r\}|X \setminus \{r\} = \gamma_r(X \setminus \{r\})$. The above argument reverses to show $\delta_r(G), \delta_r(H) \in \delta_r(\mathcal{F})$ are incompatible as split systems only if G and H are incompatible as sets. In this case $\delta_r(G \cap H), \delta_r(G \cup H), \delta_r(G \setminus H)$ and $\delta_r(H \setminus G)$ are in $\delta_r(\mathcal{F})$ so $\delta_r(\mathcal{F})$ is an unrooted family. Thus, $\delta_r(PRF) = CUF$ and $\gamma_r(CUF) = PRF$ as required. □

The above propositions combine to show that the map $\kappa_r \circ \alpha^{-1} \circ \gamma_r$ is a bijection from circular split systems that are unrooted families to PC-trees, and that in fact for a PC-tree $\mathcal{T}$, $\kappa_r \circ \alpha^{-1} \circ \gamma_r(\mathcal{T})$ is precisely the split system arising from $\mathcal{T}$ in the natural way. An example is shown in Figure 2.4. This gives:

Proposition 2.3.6. *The following diagram commutes:*

$$\begin{array}{ccc} PQ & \underset{\lambda_r}{\overset{\kappa_r}{\rightleftarrows}} & PC \\ \uparrow\downarrow \alpha & & \uparrow\downarrow \beta \\ PRF & \underset{\gamma_r}{\overset{\delta_r}{\rightleftarrows}} & CUF \end{array}$$

where PQ is the set of PQ-trees over $X \setminus \{r\}$, PRF is the set of prepyramids that are rooted families over $X \setminus \{r\}$, PC is the set of PC-trees over X, and CUF is the set of circular unrooted families over X.

We next give a geometric characterization of those split systems that come from PC-trees. We can represent a circular ordering $\mathcal{C}$ by an n-gon P whose sides are labeled with the elements of X. Each diagonal of P divides the set X into two nonempty pieces and thus gives rise to an X-split. This split lies in $\mathcal{C}$, and every split in $\mathcal{C}$ arises in this way; the edges of P correspond to the trivial splits. Hence the set of $\mathcal{C}$-compatible split systems is in bijection with the power set of diagonals of P. Two splits S_1, S_2 in $\mathcal{C}$ are incompatible if and only if the corresponding diagonals of P intersect, and $\mathcal{S}$ is a rooted family if and only if, for each such pair of splits, $\mathcal{S}$ also contains the four splits corresponding to the edges of the quadrilateral whose diagonals are S_1, S_2.

Definition 2.3.7. A *dissection* of an n-gon P is a (possibly empty) set of nonintersecting diagonals of P.

A dissection divides P into smaller regions that are also convex polygons. If our dissection has $n-3$ diagonals then it is just a triangulation of P.

Proposition 2.3.8. *Let P be the n-gon corresponding to $\mathcal{C}$. Let D be a dissection of P into subpolygons $P_1, \ldots, P_m$ and let χ be a red-blue coloring of the set of P_i that paints each*

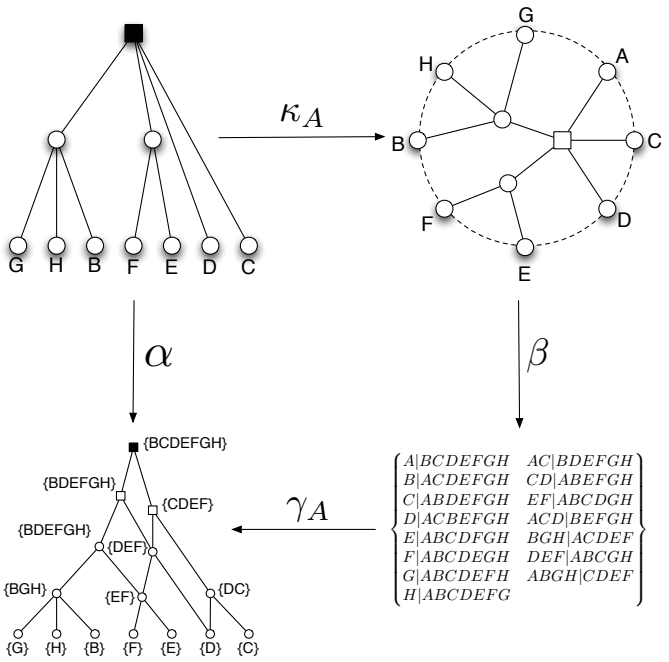


Figure 2.4: An instance of Proposition 2.3.6.

triangle blue. Let ζ be the map that sends a pair (D, χ) to the split system $\mathcal{S}$ of diagonals of P that arise either in the dissection D, or that are a diagonal of any of the red P_i. Then $\zeta \circ \beta^{-1}$ is a bijection onto the set of PC-trees.

Proof. Given a dissection D of P, let $\mathcal{T}$ be the graph dual of D. Then $\mathcal{T}$ has a leaf for each edge of P and an internal vertex for each P_i in D. Two vertices of $\mathcal{T}$ are neighbors if and only if the corresponding subpolygons P_i, P_j share an edge in D. Label a vertex in $\mathcal{T}$ a C-vertex if it corresponds to a red P_i; otherwise, label it a P-vertex.

Let $S \in \mathcal{S}$. If S is the boundary between two subpolygons, by construction the edge in $\mathcal{T}$ that joins the interior vertices corresponding to these subpolygons induces S. Otherwise S is a diagonal of some red P_i. Suppose P_i is an m-gon whose sides determine the intervals $A_1, A_2, \ldots, A_m$ of X, in clockwise order. Then S corresponds to a split of the form $A_iA_{i+1}\ldots A_{j-1}|A_jA_{j+1}\ldots A_{i-1}$. The C-vertex of P corresponding to P_i has edges that give rise to the splits $A_i|X \setminus A_i$, in clockwise order. Then $S \in \beta(\mathcal{T})$.

This shows $\mathcal{S} \subseteq \beta(\mathcal{T})$. Conversely, every split arising from a C-vertex appears as a diagonal in the corresponding P_i, and by construction every other split of $\mathcal{T}$ is in the dissection D, so $\beta(\mathcal{T}) = \mathcal{S}$. This argument is reversible, completing the proposition. □

The number of dissections of an n-gon is equal to s_{n+1}, where s_n is the nth *Schröder number* [66]. These numbers also count bracketings of strings and rooted planar trees [67].

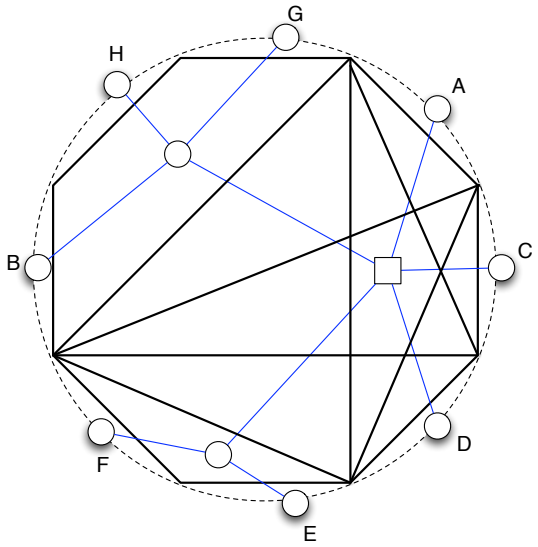


Figure 2.5: The PC-tree from Figure 2.4 (in blue) and its split system, represented as a polygon with diagonals.

Let $s_{n+1}(k)$ denote the number of dissections of an n-gon containing precisely k subpolygons with ≥ 4 edges. Then we have shown that the number of PC-trees compatible with a given circular ordering is the q-analog $\sum_{k\geq 0} s_{n+1}(k)q^k$, evaluated at $q = 2$. (Note this is different from the q-analog defined in [4]).

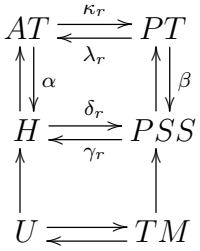
Example 2.3.9. Figure 2.5 depicts the PC-tree $\mathcal{T}$ from Figure 2.4, in blue, overlaid atop the polygon and diagonals corresponding to its split system. The subpolygon isomorphic to K_5 corresponds to the C-vertex. ◇

2.4 Metrics Realized by PC-trees

In this section we extend the tree metric theorem (Theorem 1.3.2) to classify metrics on PQ- and PC-trees. Recall the connection between trees and their representations as hierarchies and split systems:

$$\begin{array}{ccc} AT & \underset{\lambda_r}{\overset{\kappa_r}{\rightleftarrows}} & PT \\ \uparrow\downarrow\,\alpha & & \uparrow\downarrow\,\beta \\ H & \underset{\gamma_r}{\overset{\delta_r}{\rightleftarrows}} & PSS \end{array}$$

This correspondence can be extended to metrics [63]:



Here U are ultrametrics and TM are tree metrics. The tree metric theorem is proved by diagram chasing: Starting with a tree metric, the Gromov product is applied (see Definition 2.4.5 below), resulting in an ultrametric. A (unique) hierarchy representing the ultrametric can be obtained and then the PSS corresponding to the hierarchy is derived by the unrooting map δ_r. This weighted PSS represents the tree metric.

We construct a PC-tree that best realizes a Kalmanson metric by a similar approach, constructing an analog of the above diagram with suitable PQ- and PC-tree counterparts (Theorem 2.4.16). The extension requires some care, because the weighted PC-trees representing a Kalmanson metric may require extra zero splits. A key result (Theorem 2.4.4) is that there is a unique PC-tree that minimally represents any Kalmanson metric.

Definition 2.4.1. A dissimilarity map D is *Kalmanson* if there is a circular ordering $\{x_1, \ldots, x_n\}$ such that for for all $i < j < k < l$,

$$\max\{D(x_i, x_j) + D(x_k, x_l), D(x_l, x_i) + D(x_j, x_k)\} \leq D(x_i, x_k) + D(x_j, x_l). \tag{2.1}$$

Let $\mathcal{T}$ be a projective X-tree and $\mathcal{C}$ a circular ordering obtained by reading the taxa of $\mathcal{T}$ clockwise. If D is $\mathcal{T}$-additive then D is Kalmanson with respect to $\mathcal{C}$. Additionally, in this case one actually has equality in (2.1) for each $i < j < k < l$. Kalmanson metrics are thus generalizations of tree metrics obtained by relaxing the equality conditions of the four-point theorem. The following theorem, proved in [12], gives the Kalmanson metric equivalent of the four-point condition. We now make this precise.

Definition 2.4.2. Let A, B be disjoint nonempty subsets of X. Then $\sigma_{A|B}$ is the dissimilarity given by

$$\sigma_{A|B}(i,j) = \begin{cases} 1 & |\{i,j\} \cap A| = 1 \\ 0 & \text{otherwise} \end{cases}$$

When $A|B$ is an X-split, we call this a *split pseudometric*.

In this context, if T is an X-tree with an edge e, then σ_e is the split pseudometric corresponding to the split induced by e.

Theorem 2.4.3. *A metric D satisfies the Kalmanson condition if and only if there exists a circular split system $\mathcal{S}$ and weight function $w : \mathcal{S} \to \mathbb{R}^+$ such that $D = \sum_{S\in\mathcal{S}} w_S\sigma_S$. If it does, the decomposition is unique.*

Proof. Suppose $D = \sum_{S\in\mathcal{S}} w_S\sigma_S$ for some split system $\mathcal{S}$ that is compatible with respect to a circular ordering $\mathcal{C} = \{x_1, x_2, \ldots, x_n\}$. Choose $i < j < k < l$ and $S = A|B \in \mathcal{S}$. One can check that

$$D_S(x_i, x_k) + D_S(x_j, x_l) - D_S(x_i, x_j) - D_S(x_k, x_l) = \begin{cases} 2 & x_i, x_j \in A, x_k, x_l \in B, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

so D satisfies the Kalmanson condition.

Conversely, assume D is Kalmanson with respect to the circular ordering $\{x_1, \ldots, x_n\}$. Define

$$2\alpha(i, j) = D(x_i, x_j) + D(x_{i-1}, x_{j-1}) - D(x_i, x_{j-1}) - D(x_{i-1}, x_j).$$

The Kalmanson condition shows this is non-negative.

Recall that $S_{i,j} := \{x_i, x_{i+1}, \ldots, x_{j-1}\}|\{x_j, x_{j+1}, \ldots, x_{i-1}\}$. The system $\mathcal{S} = \{S_{i,j}\}_{i<j}$ is clearly circular. We claim

$$D = \sum_{i<j} \alpha(i, j)\sigma_{S_{i,j}}. \tag{2.3}$$

To see this, rewrite the right hand side of (2.3), expanding the $\alpha(i, j)$ and grouping together the coefficients of each $D(x_i, x_j)$. This gives $D = \sum_{i<j} D(x_i, x_j)c_{i,j}$ with

$$2c_{i,j} = \sigma_{S_{i,j}} + \sigma_{S_{i+1,j+1}} - \sigma_{S_{i+1,j}} - \sigma_{S_{i,j+1}}. \tag{2.4}$$

Now $c_{i,j}(x_k, x_l) = \delta_{ik}\delta_{jl}$. This proves the correctness of (2.3) and thus shows that D comes from a weighted circular split system.

For a circular ordering $\mathcal{C}$ there are $\binom{n}{2}$ splits in $\mathcal{S}$ and by (2.4) the dimension of metrics that are Kalmanson with respect to $\mathcal{C}$ is also $\binom{n}{2}$, so for a fixed circular ordering the weighting is unique. Now suppose D is Kalmanson with respect to two distinct circular orderings $\mathcal{C}_1, \mathcal{C}_2$. Let $\mathcal{S}_i$ be the split system given by $\mathcal{C}_i$, and consider the decomposition $D = \sum_{k<l} \alpha(k, l)\sigma_{S_{k,l}}$ with respect to $\mathcal{C}_1$. If $S_{i,j}$ is circular with respect to $\mathcal{C}_1$ but not with respect to $\mathcal{C}_2$, then without loss of generality there exists some k, l with $i < k < j < l$ such that $\{x_i, x_k, x_j, x_l\}$ is cyclic with respect to $\mathcal{C}_1$ and $\{x_i, x_j, x_k, x_l\}$ is cyclic with respect to $\mathcal{C}_2$. This implies

$$D(x_i, x_j) + D(x_k, x_l) \geq D(x_i, x_k) + D(x_j, x_l) \geq D(x_i, x_j) + D(x_k, x_l),$$

where the first inequality comes from the Kalmanson condition on $\mathcal{C}_1$ and the second comes from the Kalmanson condition on $\mathcal{C}_2$. So we have equality, and by (2.2),

$$0 = D(x_i, x_j) + D(x_k, x_l) - D(x_i, x_k) - D(x_k, x_l) = 2 \sum_{\substack{S=A|B\in\mathcal{S} \\ ik\in A, jl\in B}} w_S \geq w(S_{i,j}),$$

where the inequality follows since $S_{i,j}$ is in the summand. So $\alpha(i,j) = w(S_{i,j}) = 0$, and the only nonzero terms in the decomposition of D with respect to $\mathcal{C}_1$ correspond to splits in $\mathcal{S}_1$ which are also splits in $\mathcal{S}_2$. This shows the decomposition of D is unique, and thus the map ν from weighted circular split systems to Kalmanson metrics given by

$$\nu(\mathcal{S}, w)(x,y) = \sum_{S \in \mathcal{S}} w_S D_S(x,y)$$

is a bijection. □

Let $\xi = \nu^{-1}$ be the map that takes a Kalmanson metric to the weighted circular split system that describes it, and let D be Kalmanson with $\xi(D) = (\mathcal{S}, w)$. We want to find a PC-tree $\mathcal{T}$ such that $\beta(\mathcal{T}) = \mathcal{S}$ as this would provide a nice encapsulation of the "treeness" of our metric, but by Proposition 2.3.6 such a tree exists if and only if $\mathcal{S}$ is an unrooted family, which is not necessarily the case. There is, however, a canonical best choice.

Theorem 2.4.4. *Let D be a Kalmanson metric. There is a unique PC-tree $\mathcal{T}$ and weighting function $w : \beta(\mathcal{T}) \to \mathbb{R}_{\geq 0}$ such that the weighted circular split system $(\beta(\mathcal{T}), w)$ gives rise to D, and such that the number of zero weights $|\{S \in \beta(\mathcal{T}) | w_S = 0\}|$ is minimal.*

Proof. Let the *closure* of a pair of splits $A_1|B_1, A_2|B_2$ be the set $\{A_1 \cap A_2 | B_1 \cup B_2, A_1 \cap B_2 | A_2 \cup B_1, A_2 \cap B_1 | A_1 \cup B_2, B_1 \cap B_2 | A_1 \cup A_2\}$. A split system is said to be closed if it contains the closure of each pair of its splits. Define

$$\bar{\mathcal{S}} = \bigcap_{\substack{\mathcal{S}^* \text{ closed, circular} \\ \mathcal{S}^* \supseteq \mathcal{S}}} \mathcal{S}^*. \tag{2.5}$$

Note the set of all splits is closed and contains $\mathcal{S}$, so the intersection is over a nonzero number of sets. An easy lemma shows that the intersection of two closed split systems is another closed split system, so $\bar{\mathcal{S}}$ is a closed split system containing $\mathcal{S}$. By construction $\bar{\mathcal{S}}$ is a circular split system and an unrooted family, and if $\mathcal{S}'$ is an unrooted family with $\mathcal{S}' \supseteq \mathcal{S}$, then $\mathcal{S}'$ appears in the intersection (2.5) so $\mathcal{S}' \supseteq \bar{\mathcal{S}}$. By Proposition 2.3.6 there is a unique PC-tree $\mathcal{T}$ with $\beta(\mathcal{T}) = \bar{\mathcal{S}}$. We have shown that if $\mathcal{T}'$ is another PC-tree with $\beta(\mathcal{T}') \supseteq \mathcal{S}$, then $\beta(\mathcal{T}') \supset \beta(\mathcal{T})$, so in a well-defined sense $\mathcal{T}$ is the "best-fit" PC-tree for D. Let $\xi(D) = (\mathcal{S}, w)$ be the weighted circular split system corresponding to D and let $\bar{w}$ be a weighting on $\bar{\mathcal{S}}$ given by extending w as

$$\bar{w}(S) = \begin{cases} w_S & S \in \mathcal{S}, \\ 0 & S \in \bar{\mathcal{S}} \setminus \mathcal{S}. \end{cases}$$

Then $\nu(\bar{\mathcal{S}}, \bar{w}) = D$, so if $(\mathcal{S}', w')$ is a weighted circular split system with $\xi((\mathcal{S}', w')) = D$, then $\mathcal{S}' \supset \bar{\mathcal{S}}$ and $w' = w$ on $\mathcal{S}$, $w' = 0$ on $\mathcal{S}' \setminus \mathcal{S}$. □

Algorithm 2.1 Construction of the unrooted closure

Input: A circularly compatible split system $\mathcal{S}$
Output: The closure $\bar{\mathcal{S}}$ of $\mathcal{S}$
  $\bar{\mathcal{S}} \leftarrow \mathcal{S}$
  while $\bar{\mathcal{S}}$ contains a pair of incompatible splits $A_1|B_1, A_2|B_2$ but not their closure **do**
    Add the closure of $A_1|B_1, A_2|B_2$ to $\bar{\mathcal{S}}$
  end while

We can construct $\bar{\mathcal{S}}$ via Algorithm 2.1. Because X is finite this algorithm must eventually terminate. Let $\mathcal{S}'$ denote its output. Clearly $\mathcal{S}'$ is a closed split system containing $\mathcal{S}$, so $\mathcal{S}' \supseteq \bar{\mathcal{S}}$. Now if $\mathcal{S}' \neq \bar{\mathcal{S}}$, let S be the first split in $\mathcal{S}' \setminus \mathcal{S}$ that is added during the algorithm, and suppose S is added as the closure of a pair of incompatible splits $A_1|B_1$, $A_2|B_2$. Then $\bar{S}$ contains this pair of splits but not their closure, a contradiction. So the algorithm is correct, and its output does not depend on the order in which it adds splits.

In the geometry of Proposition 2.3.8, a pair of splits S_1, S_2 is incompatible if, as diagonals of P, they intersect. Adding in their closure to $\bar{\mathcal{S}}$ corresponds to adding the sides of the quadrilateral with diagonals S_1, S_2. So incompatible splits make the splits graph non-planar. PC-trees can be thought of as a way to somehow present the data in a planar fashion.

We now explore how this construction looks on the affine side.

Definition 2.4.5. Let D be a metric on X and choose $r \in X$. The *Gromov product based at* r is defined by

$$2\phi_r(D)(x,y) = D(x,y) - D(x,r) - D(y,r) \quad \forall x, y \in X \setminus \{r\}. \tag{2.6}$$

The Gromov product is also known as the *Farris transform* [22, 30] in phylogenetics. Let

$$\psi_r(R)(x,y) = 2R(x,y) - R(x,x) - R(y,y).$$

It is easy to check $\psi_r \circ \phi_r(D) = D$.

Definition 2.4.6. A matrix $\mathcal{R}$ is *Robinsonian* over X if there exists a linear ordering $\prec$ of X such that

$$max\{R(x,y), R(y,z)\} \leq R(x,z) \quad \forall x \preceq y \preceq z.$$

$\mathcal{R}$ is a *strong Robinsonian matrix* if, in addition, for all $w \preceq x \preceq y \preceq z$,

$$R(x,y) = R(w,y) \implies R(x,z) = R(w,z), \tag{2.7}$$
$$R(x,y) = R(x,z) \implies R(w,y) = R(w,z). \tag{2.8}$$

In [13] it is shown that if D is Kalmanson then $\phi_r(D)$ is a strong Robinsonian matrix. We give a slightly more precise characterization of the image.

Lemma 2.4.7. *Let D be a Kalmanson dissimilarity map and $R = \phi_r(D)$. Then R is a strong Robinsonian matrix with the following properties:*

(i) $R(x, y) \leq 0$ *for all* $x, y \in X$,

(ii) For every $w \preceq x \preceq y \preceq z$,

$$R(x, y) + R(w, z) \leq R(x, z) + R(w, y). \tag{2.9}$$

Furthermore, ϕ_r is a bijection from Kalmanson dissimilarities to the space of these matrices.

Proof. Suppose D is Kalmanson with respect to the order $\{x_1, x_2, \ldots, x_n, r\}$ and let $R = \phi_r(D)$. From the definition of the Gromov product (2.6) and the Kalmanson condition (2.1), it is immediate that $\phi_r(D)$ satisfies the above conditions with linear ordering $x_1 \prec x_2 \prec \ldots \prec x_n$. We claim that this implies R is a strong Robinsonian matrix. For $w \preceq x \preceq y \preceq z$,

$$\begin{aligned} 2(R(x, z) - R(x, y)) &= D(x, z) + D(y, r) - D(x, y) - D(z, r) \geq 0, \\ 2(R(w, y) - R(w, z)) &= D(w, y) + D(z, r) - D(w, z) - D(y, r) \geq 0, \end{aligned}$$

so $R(x, z) \geq R(x, y)$ and $R(x, z) \geq R(y, z)$, which shows R is Robinsonian. Now assume $R(x, y) = R(x, z)$. Then (2.9) gives $R(w, z) \leq R(w, y)$, and since R is Robinsonian we also have the reverse inequality, so $R(w, z) = R(w, y)$. Similarly if $R(x, y) = R(w, y)$ then $R(x, z) = R(w, z)$, so R is strong.

Conversely, let R be a strong Robinsonian matrix satisfying (2.9). Then $\psi_r(R)$ clearly satisfies the Kalmanson conditions. Also,

$$\psi_r(x, y) = (R(x, y) - R(x, x)) + (R(x, y) - R(y, y)) \geq 0,$$

and $\psi_r(x, x) = 0$ for all $x, y \in X$. So $\psi_r(R)$ is a Kalmanson dissimilarity. The maps ϕ_r and ψ_r are inverses, completing the proof. □

Therefore the image of ϕ_r consists of negative strong Robinsonian matrices satisfying a kind of four-point condition (2.9).

Next we define the affine analogue of weighted circular split systems.

Definition 2.4.8. Let $\mathcal{P}$ be a pyramid or a prepyramid. A function $f : \mathcal{P} \to \mathbb{R}$ is an *indexing function* if $A \subset B \implies f(A) < f(B)$ for all $A, B \in \mathcal{P}$. We call $(\mathcal{P}, f)$ an *indexed pyramid* or an *indexed prepyramid* respectively.

Definition 2.4.9. A subset $A \subseteq X$ is *maximally linked* [3] with respect to a Robinsonian matrix R if there exists $d \in \mathbb{R}$ such that $\mathcal{R}(x, y) \leq d$ for all $x, y \in A$, and A is maximal in this way. If A is such a set, define the *diameter* of A to be $diam(A) = \max_{x,y \in A} R(x, y)$.

Let $\mathcal{M}(\mathcal{R})$ denote the set of maximally linked sets with respect to Robinsonian matrix $\mathcal{R}$ and define the function $f : \mathcal{M}(\mathcal{R}) \to \mathbb{R}$ by $f(A) = diam(A)$.

Proposition 2.4.10. *The map $\tau : \mathcal{R} \to (\mathcal{M}(\mathcal{R}), f)$ is a bijection from Robinsonian matrices to indexed prepyramids, and from strong Robinsonian matrices to indexed pyramids.*

Proof. Suppose $A \in \mathcal{M}(\mathcal{R})$ for $\mathcal{R}$ Robinsonian and let $a, b \in A$ be the leftmost and rightmost points in A. Then for every $a \preceq x \prec y \preceq b$, $\mathcal{R}(x, y) \leq R(a, b)$, so $diam(A) = R(a, b)$. This shows $x \in A$ for all $a \preceq x \preceq b$, so every set in $\mathcal{M}(R)$ is an interval. Now suppose $A, B \in \mathcal{M}(\mathcal{R})$ with $A \subset B$. Let $A = [x_1, y_1], B = [x_2, y_2]$. Then $x_2 \preceq x_1 \preceq y_1 \preceq y_2$, and

$$f(A) = diam(A) = \mathcal{R}(x_1, y_1) < \mathcal{R}(x_2, y_2) = diam(B) = f(B),$$

where the inequality follows from the fact that A is a maximally-linked set. So f is an index and $\tau(\mathcal{R})$ is an indexed prepyramid.

Conversely, consider the map μ from indexed prepyramids to matrices given by

$$\mu(\mathcal{P}, f)(x, y) = \min_{\substack{A \in \mathcal{P} \\ x, y \in A}} f(A).$$

Let $\mathcal{R} = \mu(\mathcal{P}, f)$. Given $x \preceq y \preceq z$, let $E = \{A \in \mathcal{P} | x, y \in A\}$, $F = \{A \in \mathcal{P} | x, z \in A\}$. Then $F \subseteq E$, so

$$\mathcal{R}(x, y) = \min_{A \in E} f(A) \leq \min_{A \in F} f(A) = \mathcal{R}(x, z).$$

Similarly $\mathcal{R}(y, z) \leq \mathcal{R}(x, z)$, so $\mathcal{R}$ is Robinsonian. It is easy to check that $\mathcal{P}$ consists precisely of the sets that are maximally linked with respect to R, so τ and μ are inverses.

Now suppose $\mathcal{R}$ is a strong Robinsonian matrix. We must show $\tau(\mathcal{R})$ is closed under intersection. Let $A = [a_1, b_1], B = [a_2, b_2]$ be sets in $\mathcal{P}$, suppose $a_1 \prec a_2 \preceq b_1 \prec b_2$ and let $C = A \cap B = [a_2, b_1]$. We will show C is a maximally linked set with diameter $\mathcal{R}(a_2, b_1)$. If $x \succ b_1$, the Robinsonian condition gives $\mathcal{R}(a_2, x) \geq R(a_2, b_1)$. If there was equality then by the strong Robinsonian condition $\mathcal{R}(a_1, x) = R(a_1, b_1)$ and $x \in A$, a contradiction. Similarly, there is no $x \prec a_2$ with $R(x, b_1) = R(y, a_1)$, so $C \in \mathcal{P}$ and $\mathcal{P}$ is closed under intersection.

Conversely, suppose $(\mathcal{P}, f)$ is an indexed pyramid and let $R = \mu((\mathcal{P}, f))$. Because $\mathcal{P}$ is closed under intersection, for each $A \subseteq X$ there is a unique $\bar{A} \in \mathcal{P}$ such that $A \subseteq \bar{A}$, and $\bar{A} \subseteq B$ for all $A \subseteq B \in \mathcal{P}$. This follows immediately from taking $\bar{A} = \bigcap_{A \subseteq B \in \mathcal{P}} B$. So now, suppose $w \prec x \prec y \prec z$ and $R(x, y) = R(x, z)$. The set $A := \overline{\{x, y\}} \cap \overline{\{w, y\}}$ is in $\mathcal{P}$ since $\mathcal{P}$ is closed under intersection. $f(\overline{\{x, y\}}) = f(\overline{\{x, z\}})$ which implies $z \in \{x, y\}$. Now $x, y \in A$ implies $\overline{\{x, y\}} \in A$, so $z \in A$. But then $z \in \overline{\{w, y\}}$ which gives $R(w, z) = R(w, y)$. A similar argument shows $R(x, y) = R(w, y) \implies R(x, z) = R(w, z)$, so R is a strong Robinsonian matrix. □

Given two elements $A, B \in \mathcal{P}$ we say B is a *predecessor* of A if $A \subset B$ and there does not exist $C \in \mathcal{P}$ such that $A \subsetneq C \subsetneq B$.

Lemma 2.4.11. *Let $\mathcal{P}$ be a pyramid. Then each set in $\mathcal{P}$ has at most two predecessors.*

Proof. Suppose there is an $A = [a, b] \in \mathcal{P}$ with three distinct predecessors $B_i = [a_i, b_i]$, $i = 1, 2, 3$. Because $\mathcal{P}$ is closed under intersection $B_i \cap B_j = A$ so either $a_i = a$ or $b_i = b$. By the pigeonhole principle, two of the B_is must share an endpoint, so assume $a_1 = a_2 = a$. Then either $B_1 \subset B_2$ or $B_2 \subset B_1$ contradicting the fact that each B_i is a predecessor of A. □

For a set $A \in \mathcal{P}$, let $P_A = \{P_i\}$ denote the predecessors of A. If $(\mathcal{P}, f)$ is an indexed pyramid, we define the map $w : \mathcal{P} \to \mathbb{R}$ as the unique function satisfying

$$w(A) = \begin{cases} -f(A) & \text{if } |P_A| = 0, \\ -f(A) + f(P_1) & \text{if } |P_A| = 1, \\ -f(A) + f(P_1) + f(P_2) - f(\overline{P_1 \cup P_2}) & \text{if } |P_A| = 2. \end{cases} \tag{2.10}$$

By Lemma 2.4.11 this is well-defined.

Proposition 2.4.12. *Let R be a negative strong Robinsonian matrix satisfying the Robinsonian four-point condition, and take $\tau(R) = (\mathcal{P}, f)$. Then f is negative, and $w(A) \geq 0$ for all $A \in \mathcal{P}$. Furthermore, every such indexed pyramid lies in the image of τ.*

Proof. Let R be a negative Robinsonian matrix satisfying (2.9). Clearly (2.10) holds for $|P_A| = 0$ because f is negative, and holds for $|P_A| = 1$ because $P_1 \supset A \implies f(P_1) > f(A)$. So now assume $A = [x, y]$ has two predecessors. By the argument in Lemma 2.4.11 these must be of the form $B_1 = [w, y]$ and $B_w = [x, z]$ for some $w \prec x \prec y \prec z$, so

$$\begin{aligned} w(A) &= -f([x, y]) + f([w, y]) + f([w, z]) - f(\overline{\{w, z\}}) \\ &= -R(x, y) + R(w, y) + R(w, z) - R(w, z) \\ &\geq 0, \end{aligned}$$

because R satisfies the Robinsonian four-point condition. This argument is reversible, so we see τ really is a bijection. □

The requirement that $w(A) \geq 0$ for A with two predecessors (2.10) is thus a kind of four-point property for pyramids, and we will refer to it as such later.

Let η_r be the map sending $(\mathcal{P}, f)$ to the weighted circular split system $(\mathcal{S}, w')$ given by $\mathcal{S} = \{\delta_r(A) | A \in \mathcal{P}\}$, $w'(\delta_r(A)) = w(A)$.

Proposition 2.4.13. *If D is a Kalmanson metric, then $\nu \circ \eta_r \circ \tau \circ \phi_r$ is the identity map.*

Proof. Let $D' = \nu \circ \eta_r \circ \tau \circ \phi_r(D)$ and for $A \in \mathcal{P}$, let $O_A = \{B \in \mathcal{P} | A \subseteq B\}$ be the sets over A. A split $\delta_r(A)$ separates $x, y \in X \setminus \{r\}$ if $x \in A$ or $y \in A$ but not both. So the split pseudometric $D_{\delta_r(A)}(x, y)$ is 1 if and only if $A \in O_{\{x\}} \setminus O_{\{x,y\}}$ or $A \in O_{\{y\}} \setminus O_{\{x,y\}}$. Then

$$D'(x, y) = \sum_{A \in O_{\{x\}}} w(A) - \sum_{A \in O_{\{x,y\}}} w(A) + \sum_{A \in O_{\{y\}}} w(A) - \sum_{A \in O_{\{x,y\}}} w(A). \tag{2.11}$$

Now $O_A = O_{\bar{A}}$, so by an easy induction

$$\sum_{B \in O_A} w(B) = \sum_{B \in O_{\bar{A}}} w(B) = -f(\bar{A}),$$

and $D'(x,y) = 2f\left(\overline{\{x,y\}}\right) - f(\overline{\{x\}}) - f(\overline{\{y\}})$. Since $f(\bar{A}) = diam(A)$, by the definition of the Gromov transform,

$$D'(x,y) = D(x,y) - D(x,r) - D(y,r) + D(x,r) + D(y,r) = D(x,y).$$

To compute $D'(x,r)$ we note x and r are separated by a split $\delta_r(A)$ if and only if $x \in A$, so

$$D'(x,r) = \sum_{A \in O_{\overline{\{x\}}}} w(A) = -f(\overline{\{x\}}) = -\phi_r(D)(x,x) = D(x,r).$$

□

Let $\mathcal{R}$ be a Robinsonian matrix over X. If the prepyramid $\mathcal{P}$ in $\tau(\mathcal{R})$ is a rooted set family, Proposition 2.2.5 shows there exists a PQ-tree $\mathcal{T}$ such that $\alpha(\mathcal{T}) = \mathcal{P}$. Unfortunately this is usually not the case, so we seek instead to find a "best fit" tree. Analogous to the projective case, we define the closure of a pair of sets A, B as the set $\{A \cup B, A \cap B, A \setminus B, B \setminus A\}$ and call a collection of sets closed if it contains the pairwise closure of all its elements. We define the *rooted closure* $\bar{\mathcal{P}}$ of $\mathcal{P}$ via

$$\bar{\mathcal{P}} = \bigcap_{\substack{P^* \text{ closed} \\ P^* \supseteq P}} P^*.$$

Then $\bar{\mathcal{P}}$ is closed and contains $\mathcal{P}$, and lies inside any other closed $\mathcal{P}^*$ containing $\mathcal{P}$.Let θ be the closure map sending $\mathcal{P}$ to $\bar{\mathcal{P}}$. We then have the affine analog of Theorem 2.4.4:

Lemma 2.4.14. *The PQ-tree $\mathcal{T}$ with $\alpha(\mathcal{T}) = \theta(\mathcal{P})$ is the unique tree with $\alpha(\mathcal{T}) \supseteq \mathcal{P}$ that minimizes $|\alpha(\mathcal{T})|$.*

Analog 2.2 can be considered the affine analog of Algorithm 2.1. It remains to show that θ

Algorithm 2.2 Construction of the rooted closure

Input: A prepyramid $\mathcal{P}$
Output: The closure $\bar{\mathcal{P}}$ of $\mathcal{P}$
  $\bar{\mathcal{P}} \leftarrow \mathcal{P}$
  while $\bar{\mathcal{P}}$ contains a pair of incompatible sets A, B **do**
    $\bar{\mathcal{P}} \leftarrow \bar{H} \cup \{A \cup B, A \cap B, A \setminus B, B \setminus A\}$
  end while

commutes with the rest of the diagram. Let D be a Kalmanson metric, $\mathcal{S}$ the corresponding

split system and $\mathcal{P}$ the associated indexed pyramid. There is not necessarily a bijection between the intervals in $\mathcal{P}$ and the splits in $\mathcal{S}$; this can be seen, for example, because $\mathcal{P}$ is closed under intersection while $\mathcal{S}$ need not be. The sets in $\mathcal{P}$ that are not in $\mathcal{S}$ will get assigned weight zero by the map η_r, which is why the lower rectangle commutes, but the maps θ and ι forget about the weights so it is not clear that $\theta \circ \delta_r = \iota \circ \eta_r$. Fortunately, for pyramids that arise from Kalmanson metrics this bijection holds.

Lemma 2.4.15. *$\delta_r \circ \theta \circ \tau \circ \phi_r(D) = \iota \circ \xi(D)$ for all Kalmanson metrics D.*

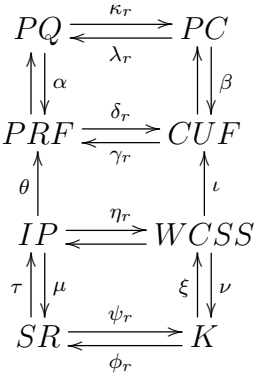
Proof. Let $(\mathcal{P}, f) = \tau \circ \phi_r(D)$ and let $(\mathcal{S}, w)$ the corresponding weighted split system. Suppose $A \in \mathcal{P}$ but $\delta_r(A) \notin \mathcal{S}$, or equivalently $f(A) = 0$. $A = [a, b]$ is an interval with respect to the Robinsonian metric. If $c \succ b$ then $M(a, c) > M(a, b)$ because $c \notin A$, so

$$0 < M(a,c) - M(a,b) = f(a,c) - f(a,b) = -\sum_{\substack{B \in \mathcal{P} \\ a,c \in B}} w(B) + \sum_{\substack{B \in \mathcal{P} \\ a,b \in B}} w(B).$$

The first summand is a subset of the second, so there exists $B \in \mathcal{P}$ with $a, b \in B$, $c \notin B$ and $w(B) > 0$. Letting c be the smallest element with $c \succ b$ shows there exists a set $B = [x, b] \in \mathcal{P}$ with $w(B) > 0$ and $x \prec a$. Similarly there exists a set $C = [a, y] \in \mathcal{P}$ with $w(C) > 0$ and $y \succ b$. So $A = B \cap C$ for sets $B, C \in \mathcal{P}$ that correspond to splits of positive weight in $\mathcal{S}$, and thus $\delta_r(A) \in \iota \circ \eta(\mathcal{S})$. This completes the proof. □

We are now ready to state our final result that summarizes the bijections described above. Let PC be the set of all PC-trees, CUF the set of all circular, unrooted split families, $WCSS$ the set of all weighted circular split systems, and K the set of all Kalmanson metrics, all over X. Let PQ be the set of all PQ-trees, PRF the set of pyramidal rooted families, IP the set of negative indexed pyramids satisfying the pyramidal four-point condition, and SR the set of negative strong Robinsonian matrices satisfying the Robinsonian four-point condition, all over $X \setminus \{r\}$.

Theorem 2.4.16. *The following diagram commutes:*



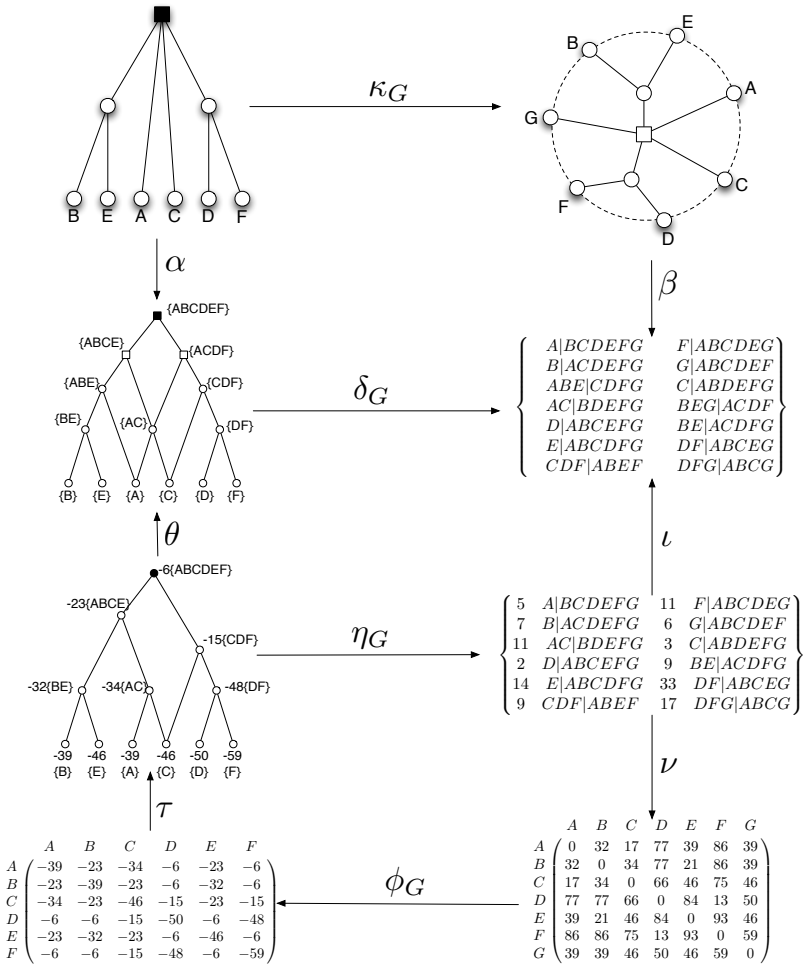


Figure 2.6: An example illustrating Theorem 2.4.16.

This gives a way of constructing the best-fit PC-tree for a given Kalmanson metric D:

$$\mathcal{T} = \kappa_r \circ \alpha^{-1} \circ \theta \circ \tau \circ \phi_r(D).$$

An example illustrating Theorem 2.4.16 is shown in Figure 2.6. The PC-tree in the upper right reveals the tree structure in the Kalmanson metric (see also Figure 2.1).

Chapter 3

Robustness of Linear Reconstruction Methods

3.1 Introduction

The fundamental challenge in distance-based phylogenetics is: Given a dissimilarity D on X, how can we reconstruct the ancestral history of X? As discussed in Chapter 1, such data can arise from a number of places including multiple alignment of homologous regions of the genomes of the taxa. Theorem 1.3.2 provides a starting point by recognizing tree-additive dissimilarities, but data is almost always noisy, and as the space of tree-additive dissimilarities has measure zero in the space of dissimilarities, almost all perturbations will result in D being not tree-additive. If this noise is sufficiently small, however, we may still maintain hope of recovering T.

In this chapter we will investigate distance-based methods that are linear in the entries of D, and will give several robustness and uniqueness results. We will outline the relevance to neighbor joining, a popular distance-based phylogenetic reconstruction algorithm, and will conclude by giving a similar robustness result for algorithms that attempt to reconstruct full circular split systems from Kalmanson metrics.

The minimum evolution (ME) approach to phylogenetic reconstruction is based on the following idea: Given a matrix D of pairwise distances between a set of n taxa, find the tree that explains D with as little evolution as possible [47, 60]. Traditionally, it employs the following general procedure:

(i) For each tree topology T, find the branch lengths of T assuming D comes from T.

(ii) Use the branch lengths to compute the length l_T of the tree T.

(iii) Choose the tree $\hat{T} = \arg\min_T l_T$ with minimum length.

There is some ambiguity in how to use negative branch lengths to compute the length of the tree. Kidd and Sgaramella-Zonta [47] proposed summing the absolute value of the edge

lengths, while Swofford et. al. [71] suggested summing only positive edge lengths. For reasons that will become clear later, we will generalize the approach of Rzhetsky and Nei [61] in calculating the length of the tree by summing all the edges, with sign.

The effectiveness of this method then depends crucially on how we select the branch lengths for a given tree T. If $w : E(T) \to \mathbb{R}$ is an edge weighting on T, let $\hat{D}$ be the corresponding T-additive dissimilarity given by $\hat{D}_{ij} = \sum_{e \in P_{ij}^T} w_e$. One classic approach, first proposed by Cavalli-Sforza and Edwards [11] and Fitch and Margoliash [34], chooses the edge lengths that minimize the sum of squares

$$\sum_{i<j} (D_{ij} - \hat{D}_{ij})^2.$$

This is known as ordinary least squares (OLS), and we refer to the corresponding ME method as OLS+ME. If we know the variances V_{ij} of the D_{ij}, then the variance-minimizing estimate of the edge lengths is given by the T-additive $\hat{D}$ that minimizes the weighted least squares (WLS)

$$\sum_{i<j} V_{ij}^{-1} (D_{ij} - \hat{D}_{ij})^2. \tag{3.1}$$

WLS assumes the D_{ij} are uncorrelated. Generalized least squares (GLS) gets rid of this assumption and seeks to minimize

$$\sum_{ij,kl} V_{ij,kl}^{-1} (D_{ij} - \hat{D}_{ij})(D_{kl} - \hat{D}_{kl}),$$

where V^{-1} is the inverse of the variance-covariance matrix of the D_{ij}. In GLS, computing the optimal edge weights is equivalent to minimizing a quadratic form, and the solution is then linear in the elements of D. The formula is

$$\hat{l}_T = (S_T^t V^{-1} S_T)^{-1} S_T^t V^{-1} D, \tag{3.2}$$

where V is the $\binom{n}{2} \times \binom{n}{2}$ variance-covariance matrix and S_T is the $\binom{n}{2} \times |E|$ matrix given in (1.1).

We briefly mention that GLS has the following statistical interpretation. Suppose $D_{ij} = \hat{D}_{ij} + \epsilon_{ij}$, where D is the observed dissimilarity, $\hat{D}_{ij}$ is the true dissimilarity and the ϵ_{ij} are error terms that are normally distributed with mean zero and covariance matrix V. Then $\hat{l}_e$ is the linear unbiased estimator for the length of edge e with minimal variance [55]. Under GLS, the total length of the tree is a linear form in the coefficients of D given by $l_T = \mathbf{1}^t \hat{l}_T$, where $\mathbf{1}$ is the vector of ones of length $|E|$. If D actually is T-additive, $D = S_T E_T$ for some positive vector E_T, and then by (3.2) a GLS+ME method on T will estimate the length of D to be $\mathbf{1}E_T$, which is the correct length.

In 2000, Pauplin proposed a certain linear form that computed the length of the tree directly, bypassing the need to calculate WLS-predicted edge lengths [58]. This suggests the

following definition. Given a T-additive dissimilarity $D = \sum_{e \in E(T)} w_e \sigma_e$, let $\text{len}(D) = \sum_e w_e$ be the length of the underlying tree. Consider the space of all dissimilarity maps on X as a subset of $\mathbb{R}^{\binom{n}{2}}$, and let $\mathcal{L}$ be the space of linear forms on $\mathbb{R}^{\binom{n}{2}}$. Unless otherwise specified, throughout the rest of the chapter we will assume our trees are trivalent.

Definition 3.1.1. A *minimum evolution (ME) method* is a reconstruction method that sends a dissimilarity D to $\arg\min_T \phi(T, D)$, where ϕ is a map $\phi : \mathcal{T}_X \to \mathcal{L}$ such that if D is T-additive,

$$(\phi(T))(D) = \text{len}(D) \tag{3.3}$$

(For notational convenience, in what follows we frequently write $\phi(T, D)$ to mean $(\phi(T))(D)$. We also use the phrase "minimum evolution method" to refer to the function ϕ itself).

Note that an ME method doesn't attempt to calculate the edge lengths; it cares only about the total length of the tree. Thus, unlike GLS, there is no good statistical interpretation for it. We are interested in determining which ϕ have the best performance. To analyze this, we will use concepts defined in Definitions 1.3.3 and 1.3.4.

Recall that a reconstruction method is *statistically consistent* if, when given a T-additive dissimilarity as input, it always returns the underlying tree T. For ME methods, this means $\arg\min_{T^*} \phi(T^*, D) = T$ for all $T \in \mathcal{T}_X$ and T-additive D. OLS+ME is consistent [61], and while this property does hold for WLS methods other than OLS [17], it does not hold for all WLS (and therefore all GLS) ME methods [35].

3.2 Balanced Minimum Evolution

Let $T \in \mathcal{T}_X$, and let i be a leaf, e the adjacent pendant edge and $A|B$ the $X \setminus \{i\}$ split such that P^T_{ab} is adjacent to e for each $a \in A, b \in B$. Let $n_A = |A|, n_B = |B|$. Then OLS+ME assigns edge e the length

$$\frac{1}{2}\left(\frac{1}{n_A}\sum_{a\in A} D_{ai} + \frac{1}{n_B}\sum_{b\in B} D_{bi} - \frac{1}{n_A n_B}\sum_{a\in A}\sum_{b\in B} D_{ab}\right).$$

This weights the clades A, B unequally. To remedy this, Pauplin proposed [58] calculating the length of a pendant edge by the formula

$$\frac{1}{2}\left(\sum_{a\in A} D_{ai} + \sum_{b\in B} D_{bi} - \sum_{a\in A}\sum_{b\in B} D_{ab}\right).$$

Similarly, if e is an interior edge with clades A, B, C, D positioned as in T in Figure 1.2, Pauplin suggested the length of e be

$$\frac{1}{2}\left(\sum_{a,c} D_{ac} + \sum_{a,d} D_{ad} + \sum_{b,c} D_{bc} + \sum_{b,d} D_{b,d} - \sum_{a,b} D_{ab} - \sum_{c,d} D_{cd}\right).$$

Summing these edge weight formulas over all edges gives a simple formula for the length of the tree:

Definition 3.2.1. Balanced Minimum Evolution (BME) is the minimum evolution method given by

$$\phi_{BME}(T, D) = \sum_{i,j} 2^{1-|P_{ij}^T|} D_{ij}. \tag{3.4}$$

Pauplin's original motivation for BME was that it permits direct calculation of the length of the tree without computing individual edge lengths. In 2008, Mihaescu and Pachter discovered that BME is actually a WLS+ME method with diagonal variance matrix $V_{ij} = 2^{-|P_{ij}^T|}$ [55]. In that paper the authors also show that these variances, which grow exponentially in the number of edges between a pair of taxa, do in fact roughly model the relationship between variance and the distance that arises from a Markov model of nucleotide substitution on a tree.

It is natural to ask about the robustness of BME. Like OLS+ME, BME is consistent [18]. It also satisfies something much stronger.

Definition 3.2.2. A dissimilarity D is *quartet consistent* with T if, for all quartets $(ij : kl)$ induced by T,

$$D_{ij} + D_{kl} < \min\{D_{ik} + D_{jl}, D_{il} + D_{jk}\}.$$

In this language, the four-point condition (Theorem 1.3.2) says that if D is T-additive, then D is quartet consistent with T. It is thus natural to make the following definition:

Definition 3.2.3. A reconstruction method is *quartet consistent* if it returns T when given an input that is quartet consistent with T.

In an unpublished work in 2009 [53], Mihaescu showed

Theorem 3.2.4. *BME is quartet consistent.*

We present his proof for completeness.

Proof. The statement holds trivially for $|X| = 4$. Let D be a dissimilarity on a set X of minimum size such that D is quartet consistent with T, but $\phi_{BME}(T, D) > \phi_{BME}(T', D)$ for some T'. Let D/a denote the restriction of D to the set $X \setminus \{a\}$, and let T/a denote the topology obtained by removing the pendant edge corresponding to a in T and contracting the resulting degree 2 node.

Let a, b form a cherry in T'. Then

$$2\phi_{BME}(T', D) = \phi_{BME}(T'/a, D/a,) + \phi_{BME}(T'/b, D/b) + D_{ab}.$$

If D, T are quartet consistent, then so are D/a and T/a. Since D, T, T' form a minimum counterexample, we also have that

$$\phi_{BME}(T/a, D/a) < \phi_{BME}(T'/a, D/a), \qquad \phi_{BME}(T/b, D/b) < \phi_{BME}(T'/b, D/b).$$

Thus the theorem is an immediate consequence of the following:

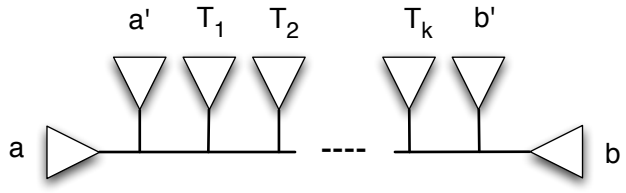


Figure 3.1: Tree used in the proof of Theorem 3.2.4.

Lemma 3.2.5. *Given a T quartet consistent D, the following holds for any pair $a, b \in X$:*

$$2\phi_{BME}(T, D) \leq \phi_{BME}(T/a, D/a) + \phi_{BME}(T/b, D/b) + D_{ab}. \tag{3.5}$$

Suppose a, b are positioned according to the configuration in Figure 3.2, where a', b', $T_1, \ldots, T_k$ are subtrees of T. We extend the definition of D to disjoint clades as $D_{AB} = \sum_{a\in A, b\in B} D_{ab}$. Then (3.5) is equivalent to

$$\begin{aligned} D_{aa'}/2 + D_{bb'}/2 + \sum_i D_{aT_i} 2^{-i-1} + \sum_i D_{bT_i} 2^{i-k-2} \\ \leq D_{ab}(1 - 2^{-k-1}) + D_{a'b'} 2^{-k-1} + \sum_i D_{a'T_i} 2^{-i-1} + \sum_i D_{b'T_i} 2^{i-k-2}. \end{aligned}$$

Taking $\delta_{xy}(z) = D_{xz} + D_{yz} - D_{xy}$, the above inequality is equivalent to

$$\frac{D_{aa'} + D_{bb'}}{2^{k+1}} + \sum_i \frac{\delta_{a'T_i}(a)}{2^{i+1}} + \sum_i \frac{\delta_{b'T_i}(b)}{2^{k+2-i}} \leq D_{ab}(1 - 2^{-k-1}) + \frac{D_{a'b'}}{2^{k+1}}.$$

By quartet consistency $D_{aa'} + D_{bb'} \leq D_{ab} + D_{a'b'}$, so it suffices to prove

$$\sum_i \frac{\delta_{a'T_i}(a)}{2^{i+1}} + \sum_i \frac{\delta_{b'T_i}(b)}{2^{k+2-i}} \leq D_{ab}(1 - 2^{-k}).$$

Split the left hand side as

$$\begin{aligned} \sum_i \frac{\delta_{a'T_i}(a)}{2^{i+1}} + \sum_i \frac{\delta_{b'T_i}(b)}{2^{k+2-i}} = \sum_{i\leq k/2} \frac{\delta_{a'T_i}(a) + \delta_{b'T_i}(b)}{2^{k+2-i}} + \sum_{i>k/2} \frac{\delta_{a'T_i}(a) + \delta_{b'T_i}(b)}{2^{i+1}} \\ + \sum_{i\leq k/2} (\delta_{a'T_i}(a) + \delta_{b'T_{k+1-i}}(b)) \left(\frac{1}{2^{i+1}} - \frac{1}{2^{k+2-i}} \right). \end{aligned}$$

It thus suffices to prove the two inequalities

$$\begin{aligned} \delta_{a'T_i}(a) + \delta_{b'T_i}(b) \leq 2D_{ab}, \\ \delta_{a'T_i}(a) + \delta_{b'T_{k+1-i}}(b) \leq 2D_{ab} \quad \text{for } i \leq k/2 \end{aligned}$$

The first inequality is equivalent to

$$D_{a'a} + D_{b'b} + D_{aT_i} + D_{bT_i} \leq 2D_{ab} + D_{b'T_i} + D_{a'T_i},$$

which itself follows from

$$D_{b'b} + D_{aT_i} \leq D_{ab} + D_{b'T_i},$$
$$D_{a'a} + D_{bT_i} \leq D_{ab} + D_{a'T_i}.$$

The second inequality is equivalent to

$$D_{a'a} + D_{b'b} + D_{aT_i} + D_{bT_{k+1-i}} \leq 2D_{ab} + D_{b'T_{k+1-i}} + D_{a'T_i}.$$

This follows from

$$D_{b'b} + D_{aT_{k+1-i}} \leq D_{ab} + D_{b'T_{k+1-i}},$$
$$D_{a'a} + D_{bT_i} \leq D_{ab} + D_{a'T_i},$$
$$D_{bT_{k+1-i}} + D_{aT_i} \leq D_{aT_{k+1-i}} + D_{bT_i}.$$

□

The set of dissimilarities that are quartet consistent with T is generally strictly greater than the set of dissimilarities that are within l_∞ radius $\frac{1}{2}$ of D. Thus Theorem 3.2.4 immediately implies BME has l_∞ radius $\frac{1}{2}$. This was also discovered independently in 2010 by Pardi, Guillemot and Gascuel [57].

Other ME methods besides OLS have been shown to be consistent [17], but quantitative calculations of their robustness are otherwise lacking. In 2005 Willson showed OLS+ME has l_∞ radius $< \frac{1}{4}$ [73], and this was improved to 0 by Pardi, Guillemot and Gascuel [57]. To our knowledge, aside from the aforementioned results on BME, these are the only such computations in the literature. Such calculations seem somewhat ad hoc, and the possibility remained of there being many other ME methods with optimal l_∞ radius. Our main result in this section is that this is not the case.

Theorem 3.2.6. *BME is the only ME method with l_∞ radius $\frac{1}{2}$.*

Thus, by at least one measure of robustness, BME is strictly better than any other distance-based reconstruction method that is linear in the elements of D. This theorem also gives an alternate definition of BME – it is the unique form with optimal l_∞ radius.

Theorem 3.2.6 will follow from a more general result later in the chapter, so we defer the proof for now.

3.3 Neighbor-Joining

Finding $\arg\min_T \phi_{BME}(T, D)$ is NP-hard to approximate [33], so BME is seldom used directly in practice. This raises the question: How can we rapidly compute an approximation to the BME-minimizing tree? One natural approach is to construct a greedy implementation.

First introduced in [62], the neighbor-joining algorithm has historically been very important in phylogenetic reconstruction, and currently has over 28,000 citations on Google Scholar. It constructs a tree in the following agglomerative way:

(1) Given a distance matrix $D : X \times X \to \mathbb{R}$, compute the Q-criterion

$$Q_D(i, j) = (n - 2)D_{ij} - \sum_{k \neq i} D_{ik} - \sum_{k \neq j} D_{jk}.$$

(2) Select a pair (a, b) of taxa that minimize Q_D. If there are more than three taxa, replace this pair by a leaf ab and construct a new dissimilarity given by $D'_{i,ab} = \frac{1}{2}(D_{ia} + D_{ib} - D_{ab})$ and $D'_{ij} = D_{ij}$ for $i, j \neq ab$.

(3) Repeat until there are three taxa remaining.

Although it was discovered in the 1980s, theoretical questions about the effectiveness of neighbor-joining and what the Q-criterion is actually measuring have taken several decades to play out. The algorithm was originally motivated by the result [62, 70] that if D is a T-additive tree metric, then a pair (a, b) minimizing Q_D is a cherry in T. This shows that neighbor-joining is consistent. The next major robustness result came in a celebrated 1999 paper, where Atteson settled a long-standing conjecture and showed neighbor-joining has l_∞ radius $\frac{1}{2}$ [1].

A method with a large l_∞ radius may still perform poorly in practice, since if the underlying tree has but one single short edge, small perturbations in the data can lead to an incorrect reconstruction. A stronger measure of robustness is a method's *edge radius*.

Definition 3.3.1. Let $\hat{D} = \sum_{e \in E(T)} w_e \sigma_e$ be a T-additive dissimilarity and chose an edge e. A reconstruction method has *edge radius* α if, when given as input a dissimilarity D satisfying $||D - \hat{D}||_\infty < \alpha w_e$, it always returns a tree containing the split induced by e.

In 2009 Mihaescu, Levy and Pachter answered a long-standing conjecture of Atteson and showed neighbor-joining has edge radius $\frac{1}{4}$ [54]. This means that even if the underlying tree has short edges, small noise will still allow it to successfully reconstruct the larger edges of the tree.

Despite these results in understanding the accuracy of neighbor-joining, the question "what is neighbor-joining actually doing?" remain unanswered until 2005. That year Desper and Gascuel provided, for the first time, an explanation of the Q-criterion [19]. Before giving

it, we need to define the BME form for non-trivalent trees [64]. Let T be such an X-tree, and define

$$c_{ij}^T = \prod_{v \in P_{ij}^T} (\deg(v) - 1)^{-1}. \tag{3.6}$$

Then

$$\phi_{BME}(T, D) = \sum_{i<j} c_{ij}^T D_{ij}.$$

This agrees with Pauplin's definition when T is binary.

Theorem 3.3.2. *Consider the star tree on X, and let T_{ab} be the tree obtained by fusing a and b together into a cherry. At each step, neighbor-joining selects the taxa a, b that minimize $\phi_{BME}(T_{ab}, D)$.*

In other words, neighbor-joining is a greedy implementation of BME. This was surprising because neighbor-joining and BME were discovered independently of each other; in fact, Pauplin did not publish his paper on BME until 13 years after neighbor-joining!

In 2005, Bryant [7] showed that the Q-criterion is the only criterion that is:

- linear in the coefficients of D,
- consistent (i.e. that given tree-like data the criterion will select a cherry at each step),
- indifferent to the order of the taxa.

For σ a permutation on X, let σD denote the dissimilarity given by $(\sigma D)_{ij} = D_{\sigma(i)\sigma(j)}$. The condition that Q is indifferent to the order of the taxa means that $Q_{\sigma D}(\sigma(i), \sigma(j)) = Q_D(i, j)$.

Thus, one can consider neighbor-joining as the unique algorithm satisfying a certain set of desirable properties. Theorem 3.2.6 can then be considered an analogous result for BME. So while the relationship between the robustness of an ME method and the robustness of an algorithm may be complex, Theorem 3.2.6 suggests that if such an algorithm is going to be based on a ME method, it's best to choose BME. Indeed, while some distance-based algorithms are known to have a better edge radius than neighbor-joining, they are themselves based on BME [6].

3.4 Generalized Balanced Minimum Evolution

In Definition 3.1.1, we required that a ME method satisfy $\phi(T, D) = \text{len}(D)$ when D is a T-additive dissimilarity. Although this normalization requirement holds for all GLS+ME methods, it seems unnatural in the broader context of linear forms. We now drop this requirement, so in what follows a *general ME method* is just a map $\phi : \mathcal{T}_X \to \mathcal{L}$.

Definition 3.4.1. Let f be a real function on the set of X-splits. For each X-tree T, let $\mathcal{U}_f(T)$ be the set of linear forms l in the entries of D such that $l(\sigma_S) = f_S$ for all splits S of T, and $f(\sigma_S) > f_S$ for all splits $S \notin T$. We say $\phi : \mathcal{T}_X \to \mathcal{L}$ is *f-consistent* if $\phi(T) \in \mathcal{U}_f(T)$ for all $T \in \mathcal{T}_X$.

This is a strict generalization of Definition 3.1.1, as the following lemma shows.

Lemma 3.4.2. *A general ME method satisfies Equation* (3.3) *if and only if it is f-consistent for the constant function $f = 1$.*

Proof. Suppose ϕ is f-consistent for $f = 1$. Then $\phi(T, \sigma_S) = 1$ if $S \in \mathcal{S}(T)$, so for $D = \sum_e w_e\sigma_e$ T-additive, $\phi(T, D) = \sum_e w_e = \text{len}(D)$. Conversely, if ϕ satisfies Equation (3.3) then for each $S \in \mathcal{S}(T)$, taking $D = \sigma_S$ gives $\phi(T, \sigma_S) = 1$. □

The following lemma illustrates why this definition is useful.

Lemma 3.4.3. *A general ME method is statistically consistent if and only if it is f-consistent for some real-valued function f on the set of X-splits.*

Proof. Suppose ϕ is f-consistent and $D = \sum_e w_e\sigma_e$ is T-additive, $w_e > 0$. If T' is a different tree, then

$$\phi(T, D) = \sum_{e \in E(T)} w_e f_{S_e} < \sum_{e \in E(T)} w_e \phi(T', \sigma_e),$$

where S_e is the split corresponding to e. The strict inequality arises from the fact at least one split in T is not in T'. So $T = \arg\min_{\hat{T}} \phi(\hat{T}, D)$, and ϕ is statistically consistent. Conversely, let ϕ be a statistically consistent general ME method, so $\arg\min_{\hat{T}} \phi(\hat{T}, D) = T$ whenever D is T-additive. For any split S, $\arg\min_T \phi(T, \sigma_S)$ must be the set of trees that contain the split S. Let $f_S = \min_T \phi(T, \sigma_S)$. Then $\phi(T, \sigma_S) \geq f_S$ with equality if and only if T contains S, and $\phi(T) \in \mathcal{U}_f(T)$ for all $T \in \mathcal{T}_X$. □

There are some values f for which no f-consistent ME methods exist. For example, let $X = \{a, b, c, d\}$. We have

$$\sigma_{a|bcd} + \sigma_{b|acd} + \sigma_{c|abd} + \sigma_{d|abc} = \sigma_{ab|cd} + \sigma_{ac|bd} + \sigma_{ad|bc}. \tag{3.7}$$

Every X-tree contains the four splits on the left hand side and only one of the splits on the right hand. So for $\phi \in \mathcal{U}_f(T)$, applying $\phi(T)$ to both sides of (3.7) gives

$$\gamma_f := f_{a|bcd} + f_{b|cda} + f_{c|dab} + f_{d|abc} - f_{ab|cd} - f_{ac|bd} - f_{ad|bc} \geq 0. \tag{3.8}$$

If this inequality is not satisfied then $\mathcal{U}_f(T)$ is empty.

We now give an analog of Theorem 3.2.6 for the more general case of f-consistent linear forms.

Theorem 3.4.4. *Let f be a real-valued function on the set of X-splits. Given $T \in \mathcal{T}_X$ and disjoint clades a, b, consider the path P_{ab}^T from a to b, and let $T_1, \ldots, T_m$ be the clades hanging off the path in order. For $I \subset [m]$, let $T_I := \bigcup_{i \in I} T_i$. Define c_{ab}^T by*

$$c_{ab}^T = \frac{1}{2^m} \sum_{I \subset [m]} \left(f\left(aT_I | bT_{[m]\setminus I}\right) - f(abT_I | T_{[m]\setminus I}) \right), \tag{3.9}$$

where $f(X|)$ is defined to be 0. If there exists an f-consistent ME method ϕ with l_∞ radius $\frac{1}{2}$, then $\phi(T, \sigma_{a|b}) = c_{ab}^T$. In particular, $\phi(T, D) = \sum_{i<j} c_{ij}^T D_{ij}$.

Proof of Theorem 3.2.6. The sum in (3.9) consists of 2^m terms with positive sign and $2^m - 1$ terms with negative sign. So when $f = f_0$ is constant, $c_{ij}^T = 2^{1-|P_{ij}^T|} f_0$ and $\phi = f_0 \cdot \phi_{BME}$ is a scalar multiple of BME. We know BME has l_∞ radius $\frac{1}{2}$, and taking $f = 1$ shows immediately that it is the only such ME method. This proves Theorem 3.2.6. □

Proof of Theorem 3.4.4. We will induct on m. Let ϕ be f-consistent, let T be a tree and let l be the linear form associated to T. Suppose T has an internal vertex v such that removing v partitions X into the three pieces A, B, C. Applying l to the equality $2\sigma_{A|B} = \sigma_{A|BC} + \sigma_{B|AC} - \sigma_{C|AB}$ gives

$$2l(\sigma_{A|B}) = f_{A|BC} + f_{B|AC} - f_{C|AB}.$$

Now let A, B, C, D be clades such that T has the split $AB|CD$, and let T' be the NNI of T with the split $AC|BD$, as in Figure 1.2. Let $\hat{D} = \sum_{e \in E(T)} w_e \sigma_e$ be a T-additive dissimilarity and let l (respectively l') be the linear form associated to T (respectively T'). We have

$$l'(\hat{D}) - l(\hat{D}) = \sum_{e \in E(T)} w_e (l'(\sigma_e) - l(\sigma_e)) = w_{AB|CD}(l'(\sigma_{AB|CD}) - f_{AB|CD}), \tag{3.10}$$

since $l'(\sigma_S) = l(\sigma_S)$ for all splits $S \in T$, $S \neq AB|CD$. Applying l' to both sides of the identity

$$\sigma_{AB|CD} = \sigma_{B|ACD} + \sigma_{C|ABD} - \sigma_{AC|BD} + 2\sigma_{A|D}$$

gives

$$l'(\sigma_{AB|CD}) = f_{B|ACD} + f_{C|ABD} - f_{AC|BD} + 2l'(\sigma_{A|D}),$$

and combining with (3.10) yields

$$l'(\hat{D}) - l(\hat{D}) = w_{AB|CD}(2\gamma + 2l'(\sigma_{AB|CD})),$$

where we have defined $\gamma := \frac{1}{2}(f_{B|ACD} + f_{C|ABD} - f_{AC|BD} - f_{AB|CD})$. Let D be a dissimilarity with $||D - \hat{D}||_\infty \leq \alpha w_{min}$. Then

$$\begin{aligned} l'(D) - l(D) &= (l'(D) - l'(\hat{D})) - (l(D) - l(\hat{D})) + (l'(\hat{D}) - l(\hat{D})) \\ &= \sum_{i<j} c_{ij}^{T'} (D - \hat{D})_{ij} - \sum_{i<j} c_{ij}^T (D - \hat{D})_{ij} + w_{AB|CD}(2\gamma + 2l'(\sigma_{AB|CD})). \end{aligned}$$

Since $l'(D) - l(D) \geq 0$, taking $(D - \hat{D})_{ij} = -\alpha w_{min} \text{sgn}(c_{ij}^{T'} - c_{ij}^{T})$ and $w_{AB|CD} = w_{min}$ gives

$$2\gamma + 2l'(\sigma_{A|D}) \geq \alpha \sum_{i<j} \left| c_{ij}^{T'} - c_{ij}^{T} \right| \tag{3.11}$$

We compute

$$\sum_{\substack{i\in A \\ j\in B}} |c_{ij}^{T'} - c_{ij}^{T}| \geq \sum_{\substack{i\in A \\ j\in B}} |c_{ij}^{T'} - c_{ij}^{T}| + \sum_{\substack{i\in A \\ j\in C}} |c_{ij}^{T'} - c_{ij}^{T}| + \sum_{\substack{i\in B \\ j\in D}} |c_{ij}^{T'} - c_{ij}^{T}| + \sum_{\substack{i\in C \\ j\in D}} |c_{ij}^{T'} - c_{ij}^{T}|. \tag{3.12}$$

Now

$$\begin{aligned}
\sum_{\substack{i\in A \\ j\in B}} |c_{ij}^{T'} - c_{ij}^{T}| &\geq |\sum_{\substack{i\in A \\ j\in B}} c_{ij}^{T'} - c_{ij}^{T}| \\
&\geq |l(\sigma_{A|B}) - l'(\sigma_{A|B})| \\
&\geq \frac{1}{2}(f_{A|BCD} + f_{B|ACD} - f_{AB|CD}) - (l'(\sigma_{A|BD}) - l'(\sigma_{A|D})) \\
&\geq \frac{1}{2}(f_{A|BCD} + f_{B|ACD} - f_{AB|CD}) - \frac{1}{2}(f_{A|BCD} + f_{AC|BD} - f_{C|ABD}) + l'(\sigma_{A|D}) \\
&\geq \gamma + l'(\sigma_{A|D}),
\end{aligned}$$

Similar calculations show

$$\begin{aligned}
\sum_{\substack{i\in A \\ j\in C}} |c_{ij}^{T'} - c_{ij}^{T}| &\geq l(\sigma_{A|D}) + \gamma, \\
\sum_{\substack{i\in B \\ j\in D}} |c_{ij}^{T'} - c_{ij}^{T}| &\geq l(\sigma_{A|D}) + \gamma, \\
\sum_{\substack{i\in C \\ j\in D}} |c_{ij}^{T'} - c_{ij}^{T}| &\geq l'(\sigma_{A|D}) + \gamma.
\end{aligned}$$

Combining these with (3.11) and (3.12) yields

$$2\gamma + 2l'(\sigma_{A|D}) \geq 2\alpha(2\gamma + l'(\sigma_{A|D}) + l(\sigma_{A|D})).$$

Taking $\alpha = \frac{1}{2}$ then gives $l'(\sigma_{A|D}) \geq l(\sigma_{A|D})$. Repeating these steps with a T'-additive dissimilarity gives the reverse inequality $l(\sigma_{A|D}) \geq l'(\sigma_{A|D})$, so we must have equality. This implies that we have equality at every step, so

$$c_{ij}^{T} = c_{ij}^{T'} \quad \forall i, j \text{ s.t. } \sigma_{A\cup D|B\cup C}(i,j) = 0. \tag{3.13}$$

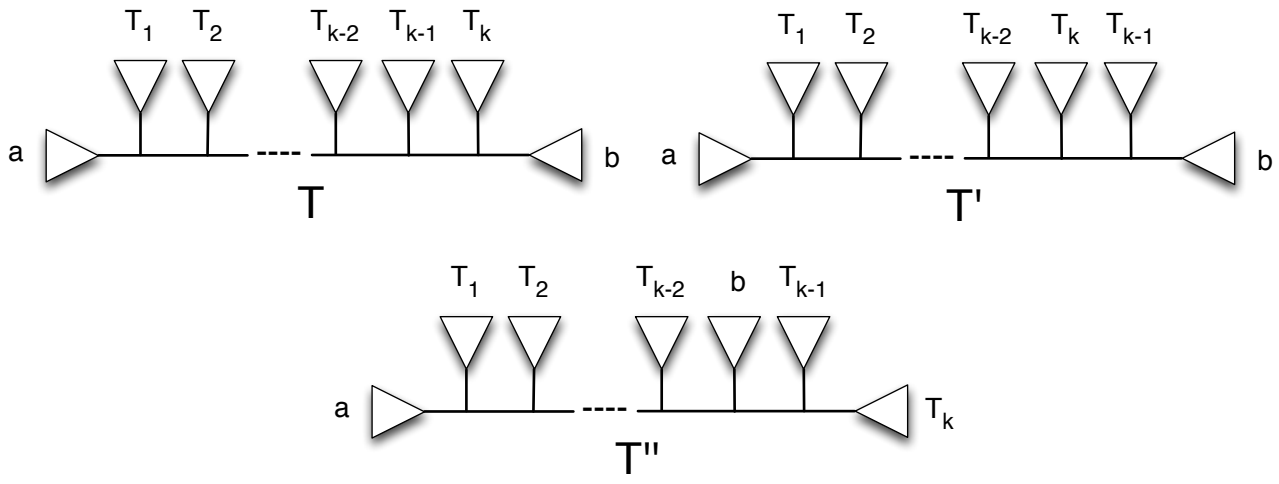


Figure 3.2: Three trees T, T', T'' used in the proof of Theorem 3.4.4.

Now let l, l', l'' be the forms associated to the trees T, T', T'' in Figure 3.2. These trees are NNIs of each other, so by Equation (3.13) we know $l(\sigma_{a|b}) = l'(\sigma_{a|b})$, and similarly $l'(\sigma_{a|T_{k-1}}) = l''(\sigma_{a|T_{k-1}})$, $l''(\sigma_{a|T_k}) = l(\sigma_{a|T_k})$. Hence

$$l(\sigma_{a|b}) - l(\sigma_{a|T_k}) = l'(\sigma_{a|T_{k-1}b}) - l''(\sigma_{a|T_{k-1}T_k}).$$

Adding this to the equation $l(\sigma_{a|b}) + l(\sigma_{a|T_k}) = l(\sigma_{a|bT_k})$ gives

$$2l(\sigma_{a|b}) = l(\sigma_{a|bT_k}) - l''(\sigma_{a|T_{k-1}T_k}) + l'(\sigma_{a|T_{k-1}b}). \tag{3.14}$$

Let $I|J$ be a partition of $[m]$, and suppose $k-1, k \in I$. By the inductive hypothesis, the coefficient of $f_{aT_I|bT_J}$ is $2^{-(m-1)}$ in each of the terms on the right hand side of (3.14), so it has coefficient 2^{-m} in $l(\sigma_{a|b})$. Now suppose $k-1 \in I$, $k \in J$. Then $f_{aT_I|bT_J}$ has coefficient $2^{-(m-1)}$ in $l(\sigma_{a|T_{k-1}b})$ and coefficient 0 in $l'(\sigma_{a|T_{k-1}b})$ and $l(\sigma_{a|T_{k-1}T_k})$, so it has coefficient 2^{-m} in $l(\sigma_{a|b})$. Checking the other cases proves the induction. □

Remark 3.4.5. Our proof did not use the full strength of the hypothesis that ϕ has l_∞ radius $\frac{1}{2}$. Indeed we only used the fact that if $\hat{D}$ is T-additive and $||D - \hat{D}||_\infty < \frac{1}{2}w_{min}$, then $\phi(T, D) < \phi(T', D)$ for T, T' separated by an NNI. This means we have shown a slightly stronger statement than that given in Theorem 3.4.4.

Example 3.4.6. Let $X = \{a, b, c, d\}$ and let T be the trivalent tree with split $ab|cd$. Then

$$\begin{aligned}
\phi_f(T, D) = &\frac{1}{2}(f_{a|bcd} + f_{b|acd} - f_{ab|cd})D_{ab} \\
&+ \frac{1}{2}(f_{c|abd} + f_{d|abc} - f_{cd|ab})D_{cd} \\
&+ \frac{1}{4}(f_{a|bcd} - f_{b|acd} + f_{c|abd} - f_{d|abc} + f_{ab|cd} - f_{ac|bd} + f_{ad|bc})D_{ac} \\
&+ \frac{1}{4}(f_{a|bcd} - f_{b|acd} - f_{c|abd} + f_{d|abc} + f_{ab|cd} + f_{ac|bd} - f_{ad|bc})D_{ad} \\
&+ \frac{1}{4}(-f_{a|bcd} + f_{b|acd} + f_{c|abd} - f_{d|abc} + f_{ab|cd} + f_{ac|bd} - f_{ad|bc})D_{bc} \\
&+ \frac{1}{4}(-f_{a|bcd} + f_{b|acd} - f_{c|abd} + f_{d|abc} + f_{ab|cd} - f_{ac|bd} + f_{ad|bc})D_{bd}
\end{aligned}$$

◇

For a fixed f let ϕ_f be the ME method given by the coefficients in (3.9). We call ϕ_f *BME-like*. Let $n = 4$, $X = \{a, b, c, d\}$, T the X-tree with split $ab|cd$ in Example 3.4.6, T' the tree with split $ac|bd$ and T'' the tree with split $ad|bc$. Then

$$\begin{aligned}
\phi_f(T', D) - \phi_f(T, D) &= \frac{\gamma_f}{4}\left(-D_{ab} + D_{ac} + D_{bd} - D_{cd}\right), \\
\phi_f(T'', D) - \phi_f(T, D) &= \frac{\gamma_f}{4}\left(-D_{ab} + D_{ad} + D_{bc} - D_{cd}\right),
\end{aligned}$$

where γ_f is the quantity defined in (3.8). If $\gamma \leq 0$ then ϕ_f is not even statistically consistent. When $\gamma_f > 0$, these expressions are positive when D is T quartet consistent, and ϕ_f is quartet consistent and has l_∞ radius $\frac{1}{2}$. For $n = 4$ every dissimilarity is T quartet consistent for one of the three tree topologies, so specifying how a method acts on quartet consistent dissimilarities defines the method over all space. Thus $\arg\min_T \phi_{BME}(T, D) = \arg\min_T \phi_f(T, D)$ and, while the methods return different scores for the same dissimilarity, they are identical in their reconstruction topology. We conjecture that this is always true.

Conjecture 3.4.7. *If ϕ_f has l_∞ radius $\frac{1}{2}$, then $\arg\min_T \phi_f(T, D) = \arg\min_T \phi_{BME}(T, D)$ for all T, D.*

Assume $X = [n]$ and let $\tau \in \mathcal{S}_n$ be a permutation on X. Define $\tau(T)$ to be the tree obtained by permuting the leaves of T according to τ, and $\tau(D)$ the dissimilarity given by $(\tau(D))_{ij} = D_{\tau(i)\tau(j)}$. We say ϕ is *permutation-invariant* if $\phi(T, D) = \phi(\tau(T), \tau(D))$ for all $\tau \in S_n$.

Lemma 3.4.8. *A BME-like method ϕ_f is permutation invariant if and only there are numbers $c_1, c_2, \ldots, c_{n-1}$ such that $c_k = c_{n-k}$ for all k, and $f_{A|B} = c_{|A|}$ for every split $A|B$.*

Proof. Fix f throughout. Let T, T' be the trees in Figure 1.2 where A, B, C, D are clades with $|B| = |C|$ and $T' = \sigma(T)$ with $\tau \in S_n$ satisfying $\tau(B) = C$, $\tau(C) = B$ and $\tau(i) = i$ for all $i \notin B, C$. Recall that for two clades A, B, $\sigma_{A|B} = \sum_{a\in A, b\in B} \sigma_{ab}$. Because ϕ is permutation invariant $\phi_f(T, \sigma_{A|B}) = \phi_f(T', \sigma_{A|C})$, so by Theorem 3.4.4,

$$f_{A|BCD} + f_{B|ACD} - f_{AB|CD} = f_{A|BCD} + f_{C|ABD} - f_{AC|BD},$$

and

$$f_{B|ACD} - f_{C|ABD} = f_{AB|CD} - f_{AC|BD}. \tag{3.15}$$

The same argument with $D = \sigma_{B|D}$ gives $f_{B|ACD} - f_{C|ABD} = f_{AC|BD} - f_{AB|CD}$, which contradicts (3.15) unless $f_{B|ACD} = f_{C|ABD}$.

This shows there are numbers $c_1, c_2, \ldots, c_{\lfloor n/2 \rfloor}$ such that $f_{A|B} = c_k$ for $|A| = k$. Conversely, suppose such numbers exist and let $a, b \in X$. Let $T_1, \ldots, T_k$ be the subtrees hanging off the path from a to b, in order. Then from the form of ϕ_f the value $\phi_f(T, \sigma_{ab})$ depends only on the sequence $|T_1|, |T_2|, \ldots, |T_k|$; let us call this sequence S^T_{ab}. We can then talk about $\phi_f(S)$ where S is any sequence of positive integers that sums to $|X| - 2$.

To show ϕ is invariant under permutations, it suffices to show $\phi_f(T, D) = \phi_f(\tau(T), \tau(D))$ when D is of the form $D = \sigma_{ab}$ for $a, b \in X$ and τ a transposition. This follows from additivity, and from the fact that transpositions generate S_n. Suppose $\tau = (i\ j)$. By examining the three cases of $|\{i, j\} \cap \{a, b\}|$, we see $S^T_{ab} = S^{\tau(T)}_{\tau(a)\tau(b)}$, so $\phi_f(T, D) = \phi_f(\tau(T), \tau(D))$. Thus ϕ_f is permutation invariant. □

3.5 A Geometric Interpretation

We can interpret ME methods geometrically. Let $\mathcal{P}_f$ be the polytope in $\mathbb{R}^{\binom{n}{2}}$ given by

$$\mathcal{P}_f = \{x | x \cdot \sigma_S \geq f_S \ \forall \text{ splits } S, \text{ with equality if } S \text{ is trivial}\}. \tag{3.16}$$

$\mathcal{P}_f$ is cut out by n equations and $2^{n-1} - n - 1$ inequalities. We show first that $\mathcal{P}_f$ is bounded. Let $x \in \mathcal{P}_f$. For every split $S = A|B$ of X,

$$f_{A|B} \leq \sum_{i\in A}\sum_{j\in B} x_{ij} = \sum_{i\in A}\sum_{j\in X} x_{ij} - 2\sum_{\{i,j\}\in\binom{A}{2}} x_{ij} \qquad = \sum_{i\in A} f_{i|X\setminus\{i\}} - 2\sum_{\{i,j\}\in\binom{A}{2}} x_{ij},$$

so

$$2\sum_{\{i,j\}\in\binom{A}{2}} x_{ij} \leq \sum_{i\in A} f_{i|X\setminus\{i\}} - f_{A|B}.$$

Taking $A = \{i, j\}$ gives

$$2x_{ij} \leq f_{i|X\setminus\{i\}} + f_{j|X\setminus\{j\}} - f_{\{i,j\}|X\setminus\{i,j\}},$$

bounding x_{ij} from above. Since

$$x_{ij} = f_{\{i,j\}|X\setminus\{i,j\}} - \sum_{k\neq i,j} x_{ik},$$

x_{ij} is bounded from below, so $\mathcal{P}_f$ is bounded. This justifies our use of the word "polytope."

Note $\mathcal{U}_f(T)$ is a face of $\mathcal{P}_f$ for each $T \in \mathcal{T}_X$, and thus is itself a polytope. A statistically consistent ME method corresponds to picking a point $\phi(T) \in \mathcal{U}_f(T)$ on each of these faces. We then construct the subpolytope $\mathcal{P}' = \text{conv}\{\phi(T)|T \in \mathcal{T}_X\}$ of $\mathcal{P}_f$ and let $\mathcal{F}$ be the normal fan of $\mathcal{P}'$. The maximal cones of $\mathcal{F}$ partition the space; each one corresponds to a vertex of $\mathcal{P}'$ and thus to an X-tree. Let C_T denote the cone corresponding to tree T. Given a dissimilarity D, our method then returns the T such that $D \in C_T$. In this geometric framework, ϕ is statistically consistent if and only if the space of T-additive dissimilarities lies inside C_T for each T. It is quartet consistent if and only if the polyhedral cone of T quartet consistent dissimilarities lies inside C_T for each T. When $f = 1$, $\mathcal{P}_f$ is known as the *BME polytope*. The BME polytope has been studied [39], but open questions remain. It would be interesting to study $\mathcal{P}_f$ for nonconstant f.

The BME-like form ϕ_f is only statistically consistent if

$$\phi_f(T, \sigma_S) > f_S \quad \forall (T, S) \; s.t. \; T \in \mathcal{T}_X, \quad S \notin T. \tag{3.17}$$

Let $N = 2^{n-1} - 1$. The set of all X-splits has cardinality N, so we may think of f as a point in $\mathbb{R}^N$. Each $\phi_f(T, \sigma_S)$ is an linear combination of the coordinates of this point, so for each T and $S \notin T$, (3.17) describes a hyperplane in this space. The set of all such hyperplanes cuts out a polyhedral cone in $\mathbb{R}^N$, and ϕ_f is statistically consistent if and only if it lies in this cone.

3.6 Robustness of Traveling Salesman for Kalmanson Metrics

In previous sections we've discussed minimum evolution methods for reconstructing the tree underlying a dissimilarity. We now suppose that our dissimilarity D is Kalmanson and arises from a full circular split system. So $D = \sum_{S\in\mathcal{C}} w_S \sigma_S$, where $\mathcal{C}$ is a circular ordering and $S \in \mathcal{C}$ means S is a split compatible with $\mathcal{C}$. We define the *length* of D to be $\text{len}(D) = \sum_{S\in\mathcal{C}} w_S$.

Definition 3.6.1. Let $\mathcal{C}_X$ denote the space of circular orderings on X. A *general minimum evolution method* is a map $\phi : \mathcal{C}_X \to \mathcal{L}$. A *minimum evolution method* is a map $\phi : \mathcal{C}_X \to \mathcal{L}$ such that if D is $\mathcal{C}$-additive, $\phi(\mathcal{C}, D) = \text{len}(D)$.

As before, we often write $\phi(\mathcal{C}, D)$ to represent $(\phi(\mathcal{C}))(D)$. We say ϕ is *consistent* if $\arg\min_{\mathcal{C}^*} \phi(\mathcal{C}^*, D) = \mathcal{C}$ whenever D is $\mathcal{C}$-additive. Let f be a real-valued function on the set of X-splits. Define $\mathcal{U}_f(\mathcal{C}) = \{l \in \mathcal{L} | l(\sigma_S) \geq f_S\}$ with equality if and only if $S \in \mathcal{C}$. The following is analogous to Lemma 3.4.3:

Lemma 3.6.2. *ϕ is consistent if and only if $\phi(\mathcal{C}) \in \mathcal{U}_f(\mathcal{C})$ for all $\mathcal{C}$.*

The space $\mathcal{U}_f(\mathcal{C})$ consists of at most a single point.

Proposition 3.6.3. *Let $\mathcal{C}$ be a circular ordering. For each pair A, B of disjoint intervals of $\mathcal{C}$, let C, D be the (possibly empty) intervals such that $\{A, C, B, D\}$ partitions X and is cyclic with respect to $\mathcal{C}$. Let*

$$2c^{\mathcal{C}}_{AB} = f_{AC|BD} + f_{AD|BC} - f_{B|ACD} - f_{D|ABC}, \tag{3.18}$$

where we take $f_{\emptyset|X} = 0$. Then $\mathcal{U}_f(\mathcal{C})$ is either empty or consists of the single linear form l such that $l(\sigma_{A|B}) = c^{\mathcal{C}}_{AB}$. In particular, $l(D) = \sum_{i<j} c^{\mathcal{C}}_{ij} D_{ij}$.

We call this form ϕ^{TSP}_f or sometimes just ϕ_f when there is no chance of confusion with the form ϕ^{BME}_f.

Proof. Let $l \in \mathcal{U}_f(\mathcal{C})$. Apply l to the equality

$$2\sigma_{A|B} = \sigma_{AC|BD} + \sigma_{AD|BC} - \sigma_{B|ACD} - \sigma_{D|ABC}.$$

□

The fact that $\dim \mathcal{U}_f(\mathcal{C}) = 0$ is not surprising, since the condition $l(\sigma_S) = f_S$ for all $S \in \mathcal{C}$ imposes $\binom{n}{2}$ linear conditions on the coordinates of the point $l \in \mathbb{R}^{\binom{n}{2}}$. If it is nonempty, $\mathcal{U}_f$ is a vertex of the polytope $\mathcal{P}_f$ defined in (3.16).

Note that the requirement $\phi(\mathcal{C}, D) = \text{len}(D)$ is equivalent to $f = 1$. In this case $\phi(\mathcal{C}, D) = \frac{1}{2}\sum_{i=1}^{n} \sigma_{\mathcal{C}(i)\mathcal{C}(i+1)}$. We denote this by ϕ_{TSP} and call it the *traveling salesman linear form*, since for a fixed D, finding $\arg\min_{\mathcal{C}} \phi(\mathcal{C}, D)$ is the traveling salesman problem. Finding the TSP-minimizing tour is NP-hard [46].

There is a relationship between circular ME methods and tree ME methods. In [64], Semple and Steel show that if $\mathcal{C}(T)$ denotes the set of circular orderings of T, then

$$\phi_{BME}(T) = \frac{1}{|\mathcal{C}(T)|} \sum_{\mathcal{C} \in \mathcal{C}(T)} \phi_{TSP}(\mathcal{C}, D). \tag{3.19}$$

Thus the BME form is the average of the circularly compatible traveling salesmen forms.

Let ϕ^{BME}_f be the form defined in Theorem 3.4.4. The same relationship holds between TSP- and BME-like forms.

Theorem 3.6.4.

$$\phi^{BME}_f(T) = \frac{1}{|\mathcal{C}(T)|} \sum_{\mathcal{C} \in \mathcal{C}(T)} \phi^{TSP}_f(\mathcal{C}, D).$$

We do not give a proof here, but note that Equation 3.19 is the special case of this when $f = 1$.

We will now calculate the l_∞ radius of the TSP form. The following are analogous to Definition 3.2.2 and Definition 3.2.3.

Definition 3.6.5. A dissimilarity D is *quartet consistent with* $\mathcal{C}$ if, for each (i, j, k, l) cyclic with respect to $\mathcal{C}$,

$$D_{ik} + D_{jl} > \max\{D_{ij} + D_{kl}, D_{il} + D_{jk}\}.$$

A circular ordering reconstruction method is quartet consistent if it returns $\mathcal{C}$ when the input is $\mathcal{C}$ quartet consistent.

Theorem 3.6.6. *ϕ_{TSP} is quartet consistent.*

Proof. Suppose not. Then there are circular orderings $\mathcal{C}, \mathcal{C}'$ and a $\mathcal{C}$ quartet consistent dissimilarity D such that $\phi_{TSP}(\mathcal{C}', D) < \phi_{TSP}(\hat{\mathcal{C}}, D)$ for all $\hat{\mathcal{C}} \neq \mathcal{C}'$. Suppose $\mathcal{C} = (1, 2, \ldots, n)$ and consider the vertices $\{1, 2, \ldots, n\}$ arranged in order around a circle. Connect two vertices by an edge if they are adjacent in $\mathcal{C}'$. Since $\mathcal{C}' \neq \mathcal{C}$ there must be a pair of intersecting chords ik and jl. This means there are $i < j < k < l$ and a, b such that $i = \mathcal{C}'(a), k = \mathcal{C}'(a+1), j = \mathcal{C}'(b), l = \mathcal{C}'(b+1)$. Let $\mathcal{C}''$ be the circular ordering obtained by inverting the elements in $\mathcal{C}'$ between k and j,

$$\mathcal{C}'' = \{\mathcal{C}'(1), \mathcal{C}'(2), \ldots, \mathcal{C}'(a), \mathcal{C}'(b), \mathcal{C}'(b-1), \ldots, \mathcal{C}'(a+1), \mathcal{C}'(b+1), \mathcal{C}'(b+2), \ldots, \mathcal{C}'(n)\}.$$

Then

$$2(f(\mathcal{C}') - f(\mathcal{C}'')) = D_{ik} + D_{jl} - D_{il} - D_{jk} > 0$$

since D is quartet consistent, a contradiction. □

As with BME, here quartet consistency immediately implies that if $\hat{D}$ is $\mathcal{C}$-additive, our method will return $\mathcal{C}$ when $||D - \hat{D}||_\infty < \frac{1}{2} w_{min}$. Unlike BME, however, this isn't optimal. In fact:

Theorem 3.6.7. *ϕ_{TSP} has l_∞ radius $\frac{n-3}{2}$, and this is the best possible.*

Proof. For notational simplicity suppose $X = \{1, \ldots, n\}$, $\mathcal{C} = \{1, \ldots, n\}$ and $\mathcal{S}$ is the corresponding split system. Let $\eta_{\mathcal{S}}(i, j)$ be the number of splits in $\mathcal{S}$ separating i and j. For any circular ordering $\mathcal{C}'$ define

$$f(\mathcal{C}') = \sum_{i=1}^{n} \eta_{\mathcal{S}}(\mathcal{C}'(i), \mathcal{C}'(i+1)). \tag{3.20}$$

Let $l(\mathcal{C}', D) = \frac{1}{2} \sum_i D(\mathcal{C}'(i), \mathcal{C}'(i+1))$ be the consistent linear form corresponding to the split system associated with $\mathcal{C}'$.

Lemma 3.6.8.

$$l(\mathcal{C}', \hat{D}) - l(\mathcal{C}, \hat{D}) \geq \frac{w_{min}}{2}(f(\mathcal{C}') - f(\mathcal{C})). \tag{3.21}$$

Note $\eta_{\mathcal{S}}(i, i+1) = n-1$, so $f(\mathcal{C}) = n(n-1)$.

We'll prove this by induction on $f(\mathcal{C}')$. For fixed i the expression $\eta_{\mathcal{S}}(i,j)$ is minimized for $j \neq i$ when $j = i \pm 1$, so $f(\mathcal{C}')$ is minimized when $\mathcal{C}' = \mathcal{C}$ in which case (3.21) obviously holds.

Now suppose $\mathcal{C}' \neq \mathcal{C}$. Consider the vertices $\{1, 2, \ldots, n\}$ arranged in order around a circle (as per $\mathcal{C}$), and connect two vertices if they are adjacent in $\mathcal{C}'$. Since $\mathcal{C}' \neq \mathcal{C}$ there must be a pair of intersecting chords ik and jl. This means there are $i < j < k < l$ and a, b such that $i = \mathcal{C}'(a), k = \mathcal{C}'(a+1), j = \mathcal{C}'(b), l = \mathcal{C}'(b+1)$. Let $\mathcal{C}''$ be the circular ordering obtained by inverting the elements in $\mathcal{C}'$ between k and j, i.e.

$$\mathcal{C}'' = \{\mathcal{C}'(1), \mathcal{C}'(2), \ldots, \mathcal{C}'(a), \mathcal{C}'(b), \mathcal{C}'(b-1), \ldots, \mathcal{C}'(a+1), \mathcal{C}'(b+1), \mathcal{C}'(b+2), \ldots, \mathcal{C}'(n)\}.$$

Note

$$f(\mathcal{C}') - f(\mathcal{C}'') = \eta_{\mathcal{S}}(i,k) + \eta_{\mathcal{S}}(j,l) - \eta_{\mathcal{S}}(i,j) - \eta_{\mathcal{S}}(k,l) = 2\eta_{\mathcal{S}}(ij,kl) > 0$$

where by abuse of notation $\eta_{\mathcal{S}}(ij,kl)$ is the number of splits $A|B \in \mathcal{S}$ with $i,j \in A$, $k,l \in B$. This shows we can assume the inductive hypothesis for $\mathcal{C}''$. Now

$$l(\mathcal{C}'', D) - l(\mathcal{C}', D) = \frac{1}{2}\left(D(i,k) + D(j,l) - D(i,j) - D(k,l)\right),$$

so

$$\begin{aligned}
l(\mathcal{C}', \hat{D}) - l(\mathcal{C}, \hat{D}) &= (l(\mathcal{C}', \hat{D}) - l(\mathcal{C}'', \hat{D})) + (l(\mathcal{C}'', \hat{D}) - l(\mathcal{C}, \hat{D})) \\
&\geq \frac{1}{2}\left(\hat{D}(i,k) + \hat{D}(j,l) - \hat{D}(i,j) - \hat{D}(k,l)\right) + \frac{w_{min}}{2}(f(\mathcal{C}'') - f(\mathcal{C})) \\
&\geq \sum_{\substack{A|B \in \mathcal{S} \\ i,j \in A, k,l \in B}} w_S - \frac{w_{min}}{2}\left(f(\mathcal{C}') - f(\mathcal{C}'')\right) + \frac{w_{min}}{2}\left(f(\mathcal{C}') - f(\mathcal{C})\right) \\
&\geq w_{min}\eta_{\mathcal{S}}(ij,kl) - w_{min}\eta_{\mathcal{S}}(ij,kl) + \frac{w_{min}}{2}(f(\mathcal{C}') - f(\mathcal{C})) \\
&\geq \frac{w_{min}}{2}(f(\mathcal{C}') - f(\mathcal{C})),
\end{aligned}$$

and the claim is proved.

Now take $\alpha = \frac{n-3}{2}$ and let D be a dissimilarity map with $|||D - \hat{D}||_\infty \leq \alpha w_{min}$. There are $2a$ terms in $l(\mathcal{C}', D) - l(\mathcal{C}, D)$, where a is the number of adjacencies in $\mathcal{C}'$ that are not in $\mathcal{C}$. Now for any i, by definition $\mathcal{C}'(i)$ and $\mathcal{C}'(i+1)$ are adjacent in $\mathcal{C}'$. If they're adjacent in $\mathcal{C}$ as well then $\eta(\mathcal{C}'(i), \mathcal{C}'(i+1)) - \eta(i, i+1) = 0$. If they're not adjacent in $\mathcal{C}$ then $\eta(\mathcal{C}'(i), \mathcal{C}'(i+1)) - \eta(i, i+1) \geq 2(n-2) - (n-1) = n-3$, so

$$\begin{aligned}
f(\mathcal{C}') - f(\mathcal{C}) &= \sum_{1 \leq i \leq n} \eta_{\mathcal{C}}(\mathcal{C}'(i), \mathcal{C}'(i+1)) - \eta_{\mathcal{C}}(i, i+1) \\
&\geq (n-3) \sum_{\substack{1 \leq i \leq n \\ \mathcal{C}'(i), \mathcal{C}'(i+1) \text{ not adjacent in } \mathcal{C}}} 1 \\
&\geq (n-3)a.
\end{aligned}$$

Thus

$$\begin{aligned} l(\mathcal{C}', D) - l(\mathcal{C}, D) &\geq l(\mathcal{C}', \hat{D}) - l(\mathcal{C}, \hat{D}) - \alpha w_{min} a \\ &\geq \frac{w_{min}}{2}(f(\mathcal{C}') - f(\mathcal{C})) - \frac{n-3}{2} w_{min} a \\ &\geq 0, \end{aligned}$$

and the proof is complete. □

For $f = 1$, the faces of the polytope $\mathcal{U}_f(\mathcal{C})$ for circular split systems is of dimension 1 while the polytope $\mathcal{U}_f(T)$ is of dimension $\frac{(n-2)(n-3)}{2}$, so we might naively expect that the optimal l_∞ radius of TSP-like forms would be $< \frac{1}{2}$. But each T has an NNI T' such that $|\mathcal{S}(T) \setminus \mathcal{S}(T')| = 1$. However, if $\mathcal{S}$ and $\mathcal{S}'$ are sets of splits consistent with distinct circular orderings $\mathcal{C}$ and $\mathcal{C}'$, then $|\mathcal{S} \setminus \mathcal{S}'| \geq n - 3$.

Conjecture 3.6.9. *Let $\mathcal{E} = \{\mathcal{S}_i\}$ be a collection of circular split systems closed under permutation (that is, $\sigma\mathcal{S}_i \in \mathcal{E}$ for all $\sigma \in S_n$), and let $\phi : \{\mathcal{S}_i\} \to \mathcal{L}$ be a corresponding minimum evolution method. Then the l_∞ radius of ϕ is $\frac{1}{2} \min_{i \neq j} |\mathcal{S}_i \setminus \mathcal{S}_j|$.*

Chapter 4

The Maximum Agreement Subtree Conjecture

4.1 Introduction

Suppose we have some set of host taxa and a set of corresponding parasites with a bijection between the two sets. We would like to determine if some of the parasites and hosts evolved together, and if so, which ones. One obvious thing to look for is for subtrees that are common to the two trees, as we expect such a shared structure to be a strong indicator of coevolution. This leads to the following definition.

Definition 4.1.1. Given an X-tree T and and a subset $Y \subseteq X$, let $T|_Y$ denote the phylogenetic Y-tree obtained by restricting T to the leaves of Y, and then contracting every edge whose vertices are both of degree 2. Throughout, the *size* of the tree will refer to the number of its leaves. Given two X-trees T_1, T_2, a subset $Y \subseteq X$ and a Y-tree T, we say T is an *agreement subtree* of T_1, T_2 if $T_1|_Y = T_2|_Y = T$. We say T is a *maximum agreement subtree* if there is no agreement subtree of larger size.

Given two phylogenetic X-trees T_1, T_2, can we find a maximum agreement subtree? This problem was first posed in the mid 1980s [32, 37] and a number of polynomial-time algorithms have since been developed that do precisely this [69]. In this chapter we consider the following combinatorial question. Let $MAST(T_1, T_2)$ be the size of a maximum agreement subtree, and let $f(n)$ denote the smallest number such that if $|X| \geq f(n)$, $MAST(T_1, T_2) \geq n$ for any $T_1, T_2 \in \mathcal{T}_X$. What is the growth rate of $f(n)$? Knowing a lower bound for the size of a maximum agreement subtree has implications for determining the significance of large common subtrees in coevolution. In particular, if we expect two trees to have a large agreement subtree, we will consider the appearance of such a tree as much less convincing evidence for coevolution than we otherwise might.

We start by investigating $f(n)$ for small values of n. We trivially have $f(3) = 3$. A case analysis [42, 68] shows:

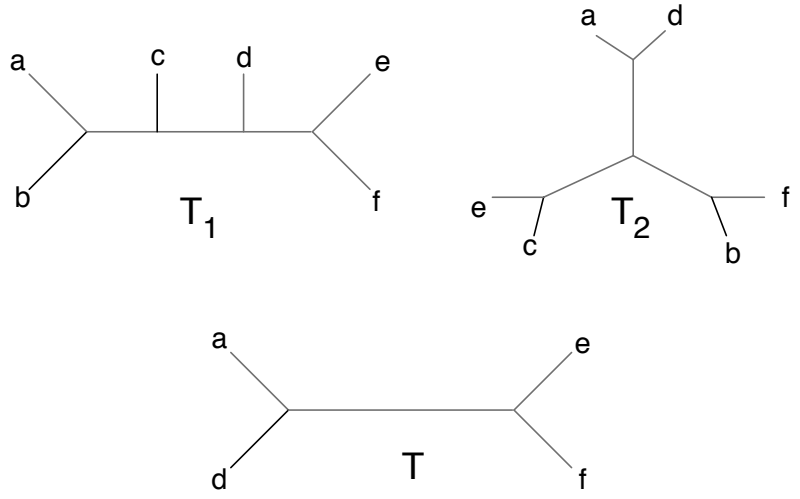


Figure 4.1: Two trees T_1, T_2 and a maximum agreement subtree T.

Lemma 4.1.2. $f(4) = 6$.

Proof. First, note that if T_1 and T_2 have disjoint cherries, these four taxa make up a shared quartet. So T_1 and T_2 must each have precisely two cherries and are thus caterpillar. Each T_i then has a split $A_i|B_i$ with $|A_i| = |B_i| = 3$. WLOG $A_1 \cap A_2 \geq 2$, so $B_1 \cap B_2 \geq 2$ and the trees share a quartet. □

With a more involved case analysis, one can show $f(5) = 13$ [42]. Larger values of $f(n)$ are unknown, and it is not even obvious that $f(n)$ is finite for all n. However, the following Ramsey-theoretic argument in [68] shows it is.

Proposition 4.1.3. *$f(n)$ exists for all $n \geq 3$.*

Proof. Ramsey theory states that there exists a minimum number $r_k(m, n)$ such that if $|X| \geq r_k(m, n)$, every red-blue coloring of the unordered k-tuples of X contains either a red set of size m or a blue set of size n, where a set is red if all its k-tuples are red, and blue if all its k-tuples are blue. So given an $|X| \geq r_4(6, n)$, color each 4-tuple $Z \subset X$ red if $T_1|_Z = T_2|_Z$ and blue otherwise. If $Y \subset X$ is a blue set of size 6 then $T_1|_Y$ and $T_2|_Y$ have no quartets in common, contradicting Lemma 4.1.2. So X contains a red set Y of size n. Then $T_1|_Y$ and $T_2|_Y$ agree on every quartet, which implies that the trees are equal [2, 14]. □

This proof shows $f(n) \leq r_4(6, n)$. Bounds in [15, 26] show that for some constant c,

$$f(n) \leq 2^{2^{cn^4 \log n}}.$$

In fact, we can do much better. In [49], Kubicki, Kubicka and McMorris showed there are constants c_1, c_2 such that

$$c_1^n < f(n) < c_2^{c_2^n}. \tag{4.1}$$

These were the first quantitative bounds for $f(n)$. We will sketch their arguments here. Both require the following easy lemma.

Lemma 4.1.4. *Let $T \in \mathcal{T}_X$ and let $Y \subset X$ be a subset of maximum size such that $T|_Y$ is caterpillar. Then $|Y| = \operatorname{diam}(T) + 1$.*

Proof. Let $d = \operatorname{diam}(T)$. Choose $x, y \in X$ such that $|P_T(x, y)| = d$ and let $A_1, A_2, \ldots, A_{d-1}$ be the clades hanging off the path P_{xy}^T in order. For each i, select $a_i \in A_i$ and let $Y = \{x, y, a_1, \ldots, a_{d-1}\}$. Then $T|_Y$ is a caterpillar with $|Y| = \operatorname{diam}(T) + 1$. Conversely, suppose $T|_Y$ is caterpillar and choose $x, y \in Y$ such that $|P_{T|_Y}(x, y)| = |Y| - 1$ is maximal. Then

$$\operatorname{diam}(T) \geq |P_T(x, y)| \geq |P_{T|_Y}(x, y)| = |Y| - 1.$$

□

The following theorem, proved by Erdös and Szekeres [27], can be thought of as the first investigation of $f(n)$.

Theorem 4.1.5 (Erdös-Szekeres)**.** *Let S be a permutation of $[N]$, where $N \geq (n-1)(m-1)+1$. Then S either has an increasing subsequence of length n or a decreasing subsequence of length m.*

Proof. Let $S = (a_1, a_2, \ldots, a_N)$ and for each i, let (l_i, g_i) be the ordered pair such that g_i is the length of the maximum increasing subsequence and l_i the length of the maximum decreasing subsequence that ends in a_i. Now for $i < j$, if $a_i < a_j$ then $g_i < g_j$, else $a_i > a_j$ and $l_j > l_i$. So $(l_i, g_i) \neq (l_j, g_j)$ for $i \neq j$. Since the pairs (l_i, g_i) are distinct and there are $(n-1)(m-1)+1$ of them, there must be some pair with either $g_i > n-1$, in which case there is an increasing subsequence of length n, or $l_i > m-1$, in which case there is a decreasing subsequence of length m. □

We now give the proof of (4.1). Let T be a caterpillar tree on $X = [N]$. We associate to T a permutation of N by choosing two taxa $x, y \in X$ with $|P_T(x, y)| = N + 1$. The path from x to y then gives a linear ordering (and hence a permutation) on X, where we take $c_1 = x$, $c_N = y$ and c_i is the leaf $2i$ edges away from x. (There are actually four permutations we can get in this way, depending on our choice of x and y). Suppose T_1, T_2 are caterpillars, and assume a permutation associated to T_1 is the identity. By Theorem 4.1.5, T_2 has either an increasing or decreasing subsequence Y of length $> \sqrt{N}$. Then $T_1|_Y = T_2|_Y$. This shows that when T_1, T_2 are both caterpillar, $MAST(T_1, T_2) > \sqrt{N}$.

So for $T_1, T_2 \in \mathcal{T}_X$, we can find a $Y_1 \subset X$, $|Y| = \log|X|$ such that $T_1|_{Y_1}$ is caterpillar. We can then find a $Y_2 \subset Y_1$, $|Y_2| = \log|Y_1|$ such that $(T_2|_{Y_1})|_{Y_2} = T_2|_{Y_2}$ is caterpillar. By Theorem 4.1.5 there exists $Y_3 \subset Y_2$, $|Y_3| > \sqrt{Y_2}$ such that $T_1|_{Y_3} = T_2|_{Y_3}$. Thus $MAST(T_1, T_2) > \sqrt{\log\log|X|}$ and

$$f(n) < 2^{2^{n^2}}. \tag{4.2}$$

For the lower bound, let $|X| = 2^n$ and let $T_1, T_2 \in T_X$ with T_1 caterpillar, T_2 a balanced binary tree. Any maximum agreement subtree of T_1, T_2 is caterpillar, so by Lemma 4.1.4 we must have $MAST(T_1, T_2) \leq \text{diam}(T_2) + 1 = 2n$. Thus $2^{n/2} \leq f(n)$.

The upper and lower bounds in (4.1) are far apart. Kubicka, Kubicki and McMorris conjectured the lower bound was the correct one [50]:

Conjecture 4.1.6. *There exists a constant c such that $f(n) < c^n$.*

The upper bound (4.2) was improved in [68] to $f(n) < c^{c^n}$, but until now this was the previously best-known result. As in [50], the proof in [68] proceeds by using Theorem 4.1.5 to find a caterpillar subtree common to both trees. In this paper we will take a different approach that will result in the first-known bound less than a double exponential. We will show

Theorem 4.1.7. *If $|X| \geq (3n)^n$, there exists a $Y \subset X$, $|Y| = n$ such that $T_1|_Y, T_2|_Y$ are caterpillar and $T_1|_Y = T_2|_Y$.*

Define $f_C(n)$ to be the smallest number such that if $|X| \geq f_C(n)$ and T_1, T_2 are caterpillar X-trees, then $MAST(T_1, T_2) \geq n$. As $f_C(n)$ is used in the derivation of both the upper and lower bound of (4.1), it is of some interest to obtain as exact a value for $f_C(n)$ as possible. Theorem 4.1.5 immediately gives $f_C(n) \leq n^2 - 2n + 2$, and Humphries [42] showed $f_C(n) \geq n^2 - 2n - 2$. We modify the proof of Theorem 4.1.5 to improve the upper bound and obtain

Theorem 4.1.8.

$$n^2 - 2n - 2 \leq f_C(n) \leq n^2 - 2n - 1.$$

4.2 Proof

In this section we prove Theorem 4.1.7. Recall $\mathcal{R}_X$ is the set of rooted binary trees. Given a tree $T \in \mathcal{R}_X$ and a subset $Y \subset X$, we define $T|_Y$ to be the binary rooted Y-tree obtained by restricting T to the elements in Y as well as the root. We say a pair $T_1, T_2 \in \mathcal{R}_X$ is (m, n)-*similar* if either of the following is true:

(i) T_1 and T_2 have a common caterpillar subtree of size n,

(ii) T_1 and T_2 have a common rooted subtree of size m.

Let $g(m,n)$ be the smallest number such that if $|X| \geq g(m,n)$, T_1 and T_2 are (m,n)-similar. Note that for $Y \subset X$, if $T_1|_Y, T_2|_Y$ are (m,n)-similar, then so are T_1, T_2. Also, $f(n) \leq g(n,n)$.

Lemma 4.2.1. *Suppose there are disjoint subsets $Y, Z \subset X$ such that for $i = 1, 2$, the root of $T_i|_{Y \cup Z}$ induces a split $A_i|B_i$ with $Y \subseteq A_i, Z \subseteq B_i$. Suppose further that $|Y| \geq g(m_1, n)$, $|Z| \geq g(m_2, n)$ for some m_1, m_2. Then T_1, T_2 are (m_1+m_2, n)-similar. In particular, if there exists $Y \subset X$, $z \in X$ with $|Y| \geq g(m-1, n)$ such that the roots of both trees split Y and z, then T_1, T_2 are (m,n)-similar.*

Proof. If $T_1|_Y, T_2|_Y$ have a common unrooted n-element subtree then so do T_1, T_2. Otherwise, because $T_1|_Y$ and $T_2|_Y$ are (m_1, n)-similar, there exists a $Y' \subseteq Y$ of size m_1 such that $T_1|_{Y'} = T_2|_{Y'}$. Similarly, if $T_1|_Z$ and $T_2|_Z$ have a common unrooted subtree of size n, there is a $Z' \subseteq Z$ of size m_2 such that $T_1|_{Z'} = T_2|_{Z'}$. Then $T_1|_{Y' \cup Z'} = T_2|_{Y' \cup Z}$, since the root induces the split $Y'|Z'$ in both subtrees. Thus T_1, T_2 either have an unrooted agreement subtree of size n or a rooted agreement subtree of size $|Y' \cup Z'| = m_1 + m_2$.

The second half of the lemma follows immediately. □

Lemma 4.2.1 will allow us to bound $g(m,n)$ by induction on m. In particular, we will show

Proposition 4.2.2.

$$g(m,n) < 3ng(m-1,n). \tag{4.3}$$

Proof. Given $T_1, T_2 \in \mathcal{R}_X$, let $|X| = N$ and suppose T_1, T_2 are not (m,n)-similar. We will find N large enough to get a contradiction. Let $A_i|B_i$ be the split given by removing the root of T_i. Suppose first that $A_1 \cap A_2, A_1 \cap B_2, A_2 \cap B_1, A_2 \cap B_2$ are all nonempty. Then WLOG $A_1 \cap A_2 \geq \frac{N}{4}$. If $N \geq 4g(m-1,n)$ then $T_1|_{A_1 \cap A_2}$ and $T_2|_{A_1 \cap A_2}$ are $(m-1,n)$-similar. Since $B_1 \cap B_2$ is nonempty, by Lemma 4.2.1, T_1, T_2 are (m,n)-similar.

So now suppose one of the intersections $A_1 \cap A_2, A_1 \cap B_2, B_1 \cap A_2, B_1 \cap B_2$ is empty. Then there is a partition of X into three nonempty pieces A, B, C such that the split in T_1 induced by the root is $AC|B$ and the split in T_2 induced by the root is $AB|C$. We must have $|B|, |C| < g(m,n)$ or else by Lemma 4.2.1, T_1, T_2 are (m,n)-similar. Choose $b \in B$ and $c \in C$ and consider the path in T_1 from r to c. Suppose this path contains k internal vertices and let $B_1, B_2, \ldots, B_k = B$ denote the clades hanging off this path, in order. Note $|B_i| < g(m-1,n)$ for all i. For if not, then taking $Y = B_i, Z = \{c\}$ and applying Lemma 4.2.1 to $T_1|_{Y \cup Z}, T_2|_{Y \cup Z}$ shows T_1, T_2 are (m,n)-similar. Similarly, consider the path in T_2 from the root to b and let $C = C_1, C_2, \cdots, C_l$ be the clades hanging off in order. Then each $|C_i| < g(m-1,n)$.

We now choose taxa $y_1, y_2, \cdots, y_n$ inductively as follows. Let $y_1 = c$, $r_1 = k+1$, $s_i = 1$. Given r_i, s_i, y_i, let $X_{i+1} = C_{[s_i+1:l]} \cap B_{[1:r_i-1]}$. If X_{i+1} is nonempty let s_{i+1} be the smallest number such that $s_{i+1} > s_i$ and $C_{s_{i+1}} \cap B_{[1:r_i-1]} \neq \emptyset$, and let r_{i+1} be the index such that $y_{i+1} \in B_{r_{i+1}}$. So $y_{i+1} \in B_{r_{i+1}} \cap C_{s_{i+1}}$, and $r_{i+1} < r_i$.

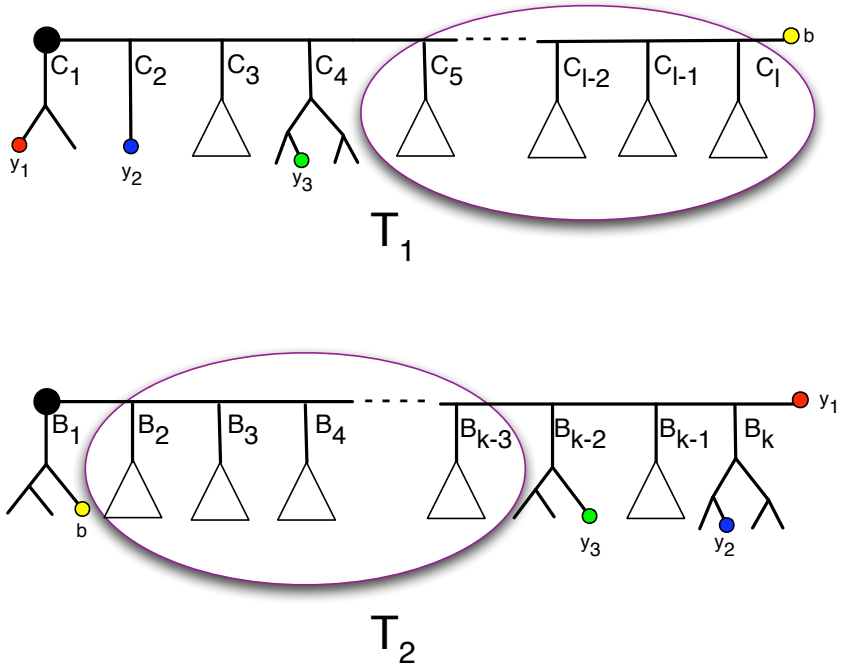


Figure 4.2: An illustration of the construction of the agreement subtree in the proof of Proposition 4.2.2

When is X_i nonempty? $X_i \setminus X_{i+1} \subseteq (B_{[r_i+1:r_{i+1}-1]} \cap C_{[s_{i+1}+1:l]}) \cup B_{r_{i+1}} \cup C_{s_{i+1}}$. Now $|B_{[r_i+1:r_{i+1}-1]} \cap C_{[s_i+1:l]}| < g(m-1,n)$, by applying Lemma 4.2.1 to this set and y_{i+1}. We already know $|C_{s_{i+1}}|, |B_{r_{i+1}}| < g(m-1,n)$, so $|X_i \setminus X_{i+1}| < 3g(m-1,n)$. Thus X_i is nonempty if $|X| \geq 3ig(m-1,n)$. So if $|X| \geq 3ng(m-1,n)$ we get a sequence of taxa $y_1, \cdots, y_n$, where $y_i \in B_{r_i} \cap C_{s_i}$, (r_i) is strictly decreasing and (s_i) is strictly increasing. Let $Y = \{y_1, \cdots, y_n\}$. Then $T_1|_Y, T_2|_Y$ are n-leaf caterpillar trees with $T_1|_Y = T_2|_Y$. This proves (4.3). □

Figure 4.2 shows this construction through $i = 3$. Here y_1, y_2, y_3 have already been computed. $r_3 = k-2$, $s_3 = 4$ and X_4 is the set of taxa that lies in both ovals.

Proof of Theorem 4.1.7. Since $g(2,n) = 2$, Proposition 4.2.2 gives the inductive bound $f(n) \leq g(n,n) < (3n)^n$. □

Remark 4.2.3. With a little care we can obtain the improved bound $f(n) \leq 6g(n/2,n)$ which gives $f(n) < c(3n)^{n/2}$ for a computable constant c. This can likely be improved with a more delicate analysis.

4.3 The Caterpillar Case

We now prove Theorem 4.1.8. Humphries showed [42] that $f_C(n) \geq n^2 - 2n - 3$, so it remains to compute the upper bound.

Proof. Let $|X| = N$ and assume WLOG that when read left to right, T_1 has leaves $1, 2, \ldots, N$. The leaves of T_2, when read left to right, are some permutation $a_1, a_2, \ldots, a_N$ of $[N]$. The size of the maximum agreement subtree is then the length of the longest subsequence $t_1, t_2, \ldots, t_m$ of (a_i) such that either

(a) $t_1, t_2 < t_3 < \ldots < t_{m-2} < t_{m-1}, t_m$, or

(b) $t_1, t_2 > t_3 > \ldots > t_{m-2} > t_{m-1}, t_m$.

Let us call such subsequences good. Let $N = (n-1)(m-1) - 3 = nm - n - m - 2$. For each i, let l_i denote the length of longest decreasing subsequence that ends with a_i, and let g_i denote the length of the longest increasing subsequence that ends with a_i. As in the proof of the Erdös-Szekeres theorem, $(l_i, g_i) \neq (l_j, g_j)$ for $i \neq j$.

So first, assume that there are $m-2$ indices $i_1 < i_2 < \ldots < i_{m-2}$ such that $g_{i_k} = n-1$ for $1 \leq k \leq m-2$. Each a_{i_k} is then the last element in an $n-1$ increasing subsequence $b_1^k, b_2^k, \ldots, b_{n-1}^k = a_{i_k}$. We can assume that for each k, $b_{n-2}^k > a_{k+1}$. For if not, then $b_1^k, b_2^k, \ldots, b_{n-1}^k, a_{i_{k+1}}$ is a good subsequence, and we're done. In particular, this implies $b_{n-2}^1 > a_{i_2}$. We may also assume that b_{n-2}^k appears after $a_{i_{k-1}}$ in the sequence, for if not, then $b_1^k, \ldots, b_{n-2}^k, a_{i_{k-1}}, a_{i_k}$ is a good sequence. This implies $a_{i_{n-2}}$ appears before b_{n-2}^{n-1} in the sequence. But then the subsequence

$$b_{n-2}^1, a_{i_1}, a_{i_2}, \ldots, a_{i_{n-2}}, b_{n-2}^k, a_{i_{n-1}}$$

is good. By a similar argument, if there exist $n-2$ elements a_i with $l_i = m-1$, we're done. Otherwise there are only $(m-3)$ values of i with $g_i = n-1$, and $n-3$ values of i with $l_i = m-1$. This implies $f(n) \leq (n-2)(m-2) + (n-3) + (m-3) + 1 = mn - n - m - 1$. □

Appendix A

Nomenclature and Abbreviations

Table A.1: Nomenclature and abbreviations.

AT	Affine (rooted) X-trees with root r
PT	Projective (unrooted) X-trees
H	Hierarchies over $X \setminus \{r\}$
PSS	Pairwise compatible split systems over X
U	Ultrametrics
TM	Tree metrics
PQ	PQ-trees over $X \setminus \{r\}$
PC	PC-trees over X
PRF	Pyramids that are rooted families over $X \setminus \{r\}$
CUF	Circular split systems that are unrooted families over X
IP	Negative indexed pyramids satisfying the pyramidal four-point condition over $X \setminus \{r\}$
WCSS	Weighted circular compatible split systems over X
SR	Negative strong Robinsonian matrices satisfying the Robinsonian four-point condition over $X \setminus \{r\}$
K	Kalmanson metrics over X

Bibliography

[1] K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.

[2] H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. in Appl. Math.*, 7(3):309–343, 1986.

[3] P. Bertrand and M. Janowitz. Pyramids and weak hierarchies in the ordinal model for clustering. *Discrete Applied Mathematics*, 122:55–81, 2002.

[4] J. Bonin, L. Shapiro, and R. Simion. Some q-analogues of the Schröder numbers arising from combinatorial statistics on lattice paths. *J. Statist. Plann. Inference*, 34(1):35–55, 1993.

[5] K. Booth and G. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976.

[6] M. Bordewich and R. Mihaescu. Accuracy guarantees for phylogeny reconstruction algorithms based on balanced minimum evolution. In V. Moulton and M. Singh, editors, *Algorithms in Bioinformatics*, volume 6293 of *Lecture Notes in Computer Science*, pages 250–261. Springer Berlin / Heidelberg, 2010.

[7] D. Bryant. On the uniqueness of the selection criterion in neighbor-joining. *Journal of Classification*, 22:3–15, 2005.

[8] D. Bryant and V. Moulton. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Molecular Biology And Evolution*, 21:255–265, 2004.

[9] D. Bryant, V. Moulton, and A. Spillner. Consistency of the Neighbor-Net algorithm. *Algorithms for Molecular Biology*, 2:8, 2007.

[10] P. Buneman. The recovery of trees from measures of dissimilarity. In F. Hodson, D. Kendall, and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.

[11] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human genetics*, 19:223–257, 1967.

[12] V. Chepoi and B. Fichet. A note on circular decomposable metrics. *Geometrica Dedicata*, 69:237–240, 1998.

[13] G. Christopher, M. Farach, and M. Trick. The structure of circular decomposable metrics. In *Lecture Notes in Computer Science*, volume 1136, pages 406–418. Springer, New York, 1996.

[14] H. Colonius and H.-H. Schulze. Tree structures for proximity data. *British J. Math. Statist. Psych.*, 34(2):167–180, 1981.

[15] D. Conlon, J. Fox, and B. Sudakov. Hypergraph Ramsey numbers. *J. Amer. Math. Soc.*, 23(1):247–266, 2010.

[16] C. Darwin. Notebook B – Transmutation of Species. 1837–8.

[17] F. Denis and O. Gascuel. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127(1):63–77, Apr. 2003.

[18] R. Desper and O. Gascuel. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21:587–598, 2004.

[19] R. Desper and O. Gascuel. The minimum evolution distance-based approach to phylogenetic inference. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*. Oxford University Press, 2005.

[20] E. Diday. *Multidimensional data analysis*, chapter Orders and overlapping clusters by pyramids, pages 201–234. DWO Press, Leiden, 1986.

[21] A. Dress. Towards a theory of holistic clustering. In *Mathematical Hierarchies and Biology*. DIMACS, 1997.

[22] A. Dress, K. Huber, and V. Moulton. Some uses of the Farris transform in mathematics and phylogenetics– a review. *Annals of Combinatorics*, 11:1–37, 2007.

[23] A. Dress, K. T. Huber, J. Koolen, V. Moulton, and A. Spillner. *Basic Phylogenetic Combinatorics*. Cambridge University Press, 2012.

[24] M. Dunn, A. Terrill, G. Reesnik, R. Foley, and S. Levinson. Structural phylogenetics and reconstruction of ancient language history. *Science*, pages 2072–2075, 2005.

[25] J. Edmonds and R. Giles. A min-max relation for submodular functions on graphs. In *Studies in integer programming (Proc. Workshop, Bonn, 1975)*, pages 185–204. Ann. of Discrete Math., Vol. 1. North-Holland, Amsterdam, 1977.

[26] P. Erdös and R. Rado. Combinatorial theorems on classifications of subsets of a given set. *Proc. London Math. Soc. (3)*, 2:417–439, 1952.

[27] P. Erdös and G. Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[28] C. Eslahchi, M. Habibi, R. Hassanzadeh, and E. Mottaghi. MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method. *BMC Evolutionary Biology*, 10:254, 2010.

[29] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the thirty-fifth annual ACM symposium on the Theory of Computing*, pages 448–455, 2003.

[30] J. Farris. Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106:645–668, 1972.

[31] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):pp. 401–410, 1978.

[32] C. R. Finden and A. D. Gordon. Obtaining common pruned trees. *Journal of Classification*, 2:255–276, 1985.

[33] S. Fiorini and G. Joret. Approximating the balanced minimum evolution problem. *CoRR*, abs/1104.1080, 2011.

[34] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(3760):279–284, 1967.

[35] O. Gascuel, D. Bryant, and F. Denis. Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50(5):pp. 621–627, 2001.

[36] O. Gascuel and M. Steel. Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, Nov. 2006.

[37] A. D. Gordon. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classification*, 3(2):335–348, 1986.

[38] D. Gusfield. Efficient algorithms for inferring evolutionary history. *Networks*, 21:19–28, 1991.

[39] D. C. Haws, T. L. Hodge, and R. Yoshida. Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. *Bull. Math. Biol.*, 73(11):2627–2648, 2011.

[40] W.-L. Hsu. PC-trees vs. PQ-trees. In *Lecture Notes in Computer Science*, volume 2108, pages 207–217, 2001.

[41] W.-L. Hsu and R. McConnell. PC trees and circular-ones arrangements. *Theoretical Computer Science*, 296:99–116, 2003.

[42] P. J. Humphries. *Combinatorial Aspects of Leaf-Labelled Trees*. PhD thesis, University of Canterbury, 2008.

[43] D. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–267, 2005.

[44] C. Jardine, N. Jardine, and R. Sibson. The structure and construction of taxonomic hierarchies. *Mathematical Bioscience*, 1:173–179, 1967.

[45] K. Kalmanson. Edgeconvex circuits and the traveling salesman problem. *Canadian Journal of Mathematics*, 27:1000–1010, 1974.

[46] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, pages 85–103. Plenum, New York, 1972.

[47] K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: concepts and methods. *Am J Hum Genet*, 23(3):235–252, May 1971.

[48] A. Kleinman, M. Harel, and L. Pachter. Affine and projective tree metric theorems. *Annals of Combinatorics*. In press.

[49] E. Kubicka, G. Kubicki, and F. R. McMorris. On agreement subtrees of two binary trees. In *Proceedings of the Twenty-third Southeastern International Conference on Combinatorics, Graph Theory, and Computing (Boca Raton, FL, 1992)*, volume 88, pages 217–224, 1992.

[50] E. Kubicka, G. Kubicki, and F. R. McMorris. An algorithm to find agreement subtrees. *Journal of Classification*, 12:91–99, 1995.

[51] V. Kunin, L. Goldovsky, N. Darzentas, and C. A. Ouzounis. The net of life: reconstructing the microbial phylogenetic network. *Genome Res.*, 15(7):954–959, Jul 2005.

[52] D. Levy and L. Pachter. The neighbor-net algorithm. *Advances in Applied Mathematics*, 47:240–258, 2011.

[53] R. Mihaescu. BME is quartet consistent. Unpublished, 2009.

[54] R. Mihaescu, D. Levy, and L. Pachter. Why neighbor-joining works. *Algorithmica*, 54(1):1–24, 2009.

[55] R. Mihaescu and L. Pachter. Combinatorics of least squares trees. *Proceedings of the National Academy of Sciences*, 105:13206–13211, 2008.

[56] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.

[57] F. Pardi, S. Guillemot, and O. Gascuel. Robustness of phylogenetic inference based on minimal evolution. *Bulletin of Mathematical Biology*, 72:1820–1839, 2010.

[58] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51(1):41–47, 2000.

[59] W. Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16:293–301, 1951.

[60] A. Rzhetsky and M. Nei. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*, 9(5):945, 1992.

[61] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5):1073–1095, 1993.

[62] N. Saitou and M. Nei. The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[63] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[64] C. Semple and M. Steel. Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32:669–680, 2004.

[65] W.-K. Shih and W.-L. Hsu. A new planarity test. *Theoretical Computer Science*, 223:179 – 191, 1999.

[66] R. P. Stanley. *Enumerative Combinatorics. Vol. 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997.

[67] R. P. Stanley. Hipparchus, Plutarch, Schröder, and Hough. *Amer. Math. Monthly*, 104(4):344–350, 1997.

[68] M. Steel and L. A. Székely. An improved bound on the maximum agreement subtree problem. *Appl. Math. Lett.*, 22(11):1778–1780, 2009.

[69] M. Steel and T. Warnow. Kaikoura tree theorems: computing the maximum agreement subtree. *Inform. Process. Lett.*, 48(2):77–82, 1993.

[70] J. Studier and K. Keppler. A note on the neighbor-joining method of saitou and nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[71] D. Swofford, G. Olsen, P. Waddell, and D. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates, 1996.

[72] W. White, S. Hills, R. Gaddam, B. Holland, and D. Penny. Treeness triangles: visualizing the loss of phylogenetic signal. *Molecular Biology and Evolution*, 24:2029–2039, 2007.

[73] S. J. Willson. Minimum evolution using ordinary least-squares is less robust than neighbor-joining. *Bull. Math. Biol.*, 67(2):261–279, 2005.