2015

STATISTICS IN THE BILLERA-HOLMES-VOGTMANN TREESPACE

Grady S. Weyenberg
University of Kentucky, gradysw@gmail.com

STUDENT AGREEMENT:

REVIEW, APPROVAL AND ACCEPTANCE

Grady S. Weyenberg, Student

Dr. Ruriko Yoshida, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

STATISTICS IN THE BILLERA-HOLMES-VOGTMANN TREESPACE

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Arts and Sciences at the University of Kentucky

By

Grady Weyenberg

Lexington, Kentucky

Director: Dr. Ruriko Yoshida, Ph.D., Professor of Statistics

Lexington, Kentucky

ABSTRACT OF DISSERTATION

STATISTICS IN THE BILLERA-HOLMES-VOGTMANN TREESPACE

This dissertation is an effort to adapt two classical non-parametric statistical techniques, kernel density estimation (KDE) and principal components analysis (PCA), to the Billera-Holmes-Vogtmann (BHV) metric space for phylogenetic trees. This adaption gives a more general framework for developing and testing various hypotheses about apparent differences or similarities between sets of phylogenetic trees than currently exists.

For example, while the majority of gene histories found in a clade of organisms are expected to be generated by a common evolutionary process, numerous other coexisting processes (e.g. horizontal gene transfers, gene duplication and subsequent neofunctionalization) will cause some genes to exhibit a history quite distinct from the histories of the majority of genes. Such “outlying” gene trees are considered to be biologically interesting and identifying these genes has become an important problem in phylogenetics.

The R sofware package KDETREES, developed in Chapter 2, contains an implementation of the kernel density estimation method. The primary theoretical difficulty involved in this adaptation concerns the normalizion of the kernel functions in the BHV metric space. This problem is addressed in Chapter 3. In both chapters, the software package is applied to both simulated and empirical datasets to demonstrate the properties of the method.

A few first theoretical steps in adaption of principal components analysis to the BHV space are presented in Chapter 4. It becomes necessary to generalize the notion of a set of perpendicular vectors in Euclidean space to the BHV metric space, but there some ambiguity about how to best proceed. We show that convex hulls are one reasonable approach to the problem. The NYE-PCA-ALGORITHM provides a method of projecting onto arbitrary convex hulls in BHV space, providing the core of a modified PCA-type method.

KEYWORDS: Phylogenetic trees, Non-parametric statistics, Outlier Detection, Kernel Density Estimation, Principal Components Analysis

GRADY WEYENBERG
Student’s Signature

JULY 30, 2015
Date

STATISTICS IN THE BILLERA-HOLMES-VOGTMANN TREESPACE

By

Grady Weyenberg

RURIKO YOSHIDA, PH.D.
Director of Dissertation

CONSTANCE L. WOOD, PH.D.
Director of Graduate Studies

JULY 30, 2015
Date

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Portions of this chapter have been published as Weyenberg and Yoshida [177] and [173].

Abstract

Phylogenetic trees are mathematical objects which summarize the most recent common ancestor relationships between a given set of organisms. There is often a need to quantify the degree of similarity or discordance between two proposed trees. For instance, a person may be interested in knowing whether the phylogenetic trees reconstructed from two distinct sequence alignments are truly different, or if the differences are so minor as to be attributable only to statistical variation. In this chapter we introduce a number of important concepts relating to phylogenetic trees and their reconstruction, and survey the literature of statistical methods designed for use with trees. The most common models of sequence evolution are presented first, followed by a brief description of the process of tree reconstruction. Next, a few methods for defining distances between phylogenetic trees are discussed, culminating in a discussion of the Billera-Holmes-Vogtmann metric space for trees, which is the main subject of study in this dissertation. Finally some existing statistical methods for testing tree congruence, and the reasons for their inadequacy are discussed.

1.1 Phylogenetic Trees

Extensive evidence of interrelatedness between all life on Earth is one of the central findings of the modern biological sciences. Although the origins of the theory of evolution predate the work of Darwin and Wallace, *The Origin of Species* is the first publication which presented a compelling natural mechanism for evolution: descent with modification, guided by the forces of natural selection [33]. The descent with modification theory posits that every pair of organisms share, somewhere in their extended genealogies, a *most recent common ancestor* (MRCA). Thus, a genealogy of any collection of individual organisms and their MRCAs should be organized in a tree-like structure, called a *phylogeny.* An early sketch of such a structure is shown in Figure 1.1.

A *phylogenetic tree* is a specific type of mathematical graph, where the vertices are used to represent individual organisms, species, or possibly larger populations of contemporaneous individuals, called *operational taxonomic units* (OTUs), or simply *taxa.* Edges connect those OTUs which are most closely related through a direct line of descent. Typically, some of the taxa found in a phylogeny are species which are directly observed in the present, and some of the taxa represent hypothetical ancestor populations. When drawn as a tree, the vertices corresponding to contemporary OTUs appear at the tips of the tree, and are therefore called *leaf vertices*, or simply *leaves.* The other vertices, which correspond to (unobserved) MRCAs appear within the tree, and are called *internal.*

Edge weights (also called *branch lengths*) if they are provided, are typically used to describe how closely or distantly related the taxa they connect are to each other. The most common units used for an edge weight are time (or the number of generations) separating the taxa, or the expected number of times that a nucleotide in an ancestor sequence will be substituted for a different base as the population evolves into the descendent sequence. A pair of trees equipped with edge lengths are shown in Figure 1.2. Note that the internal vertices, representing MRCA taxa, have not been explicitly labeled.

Often, phylogenies are assumed to have the additional structure of a *binary tree.* A *tree* is a graph in which the edges do not form any closed cycles, and a tree is *binary* if each vertex has degree at most three, i.e., each vertex has edges connected to at most one ancestor and two child taxa. For example, the tree in Figure 1.1 is not binary, whereas the trees in Figure 1.2 are binary. While a binary tree might at first seem to be a necessary consequence of the canonical framework of descent with modification, we shall see that this is not actually the case and that there are good reasons to

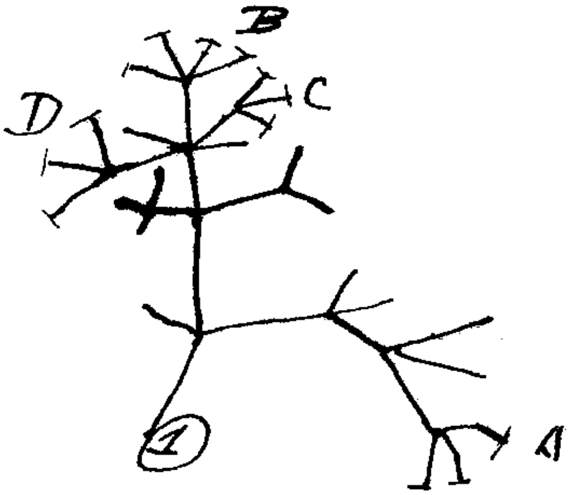


Figure 1.1: An early sketch of a phylogenetic tree, found in Darwin's early notebooks [34].

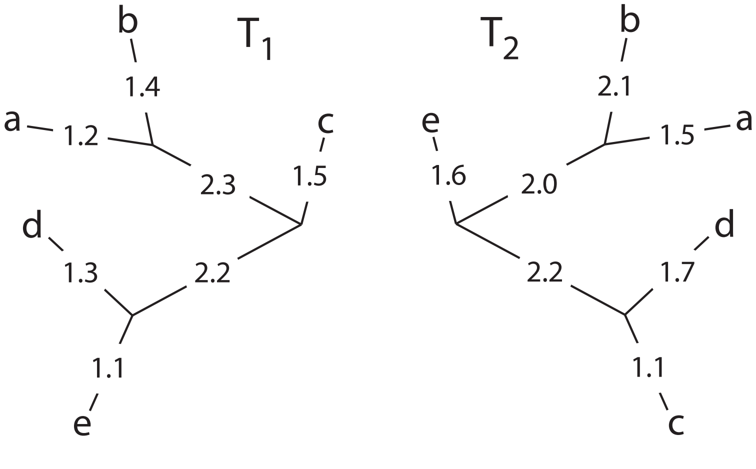


Figure 1.2: Example phylogenetic trees: T_1 and T_2. The trees represent proposed most recent common ancestor relationships between 5 taxa, labeled a through e. These trees have branch lengths specified, but not all trees need have such information.

question the binary tree assumptions. Despite this, binary trees remain well established within the scientific community as a reasonable simplifying approximation in most cases.

1.2 Modeling Evolution

Generally, the first step in reconstructing a phylogenetic tree is to obtain a sample of representative individuals from each taxon to be included in the phylogeny, and then observe and quantify in them a number of morphological traits. The traits measured could range from simple physiological observations, e.g. skull volume, or they may detail the differences between the biological sequences, which comprise the chemistry of life.

Although it is possible to use a variety of data to form phylogenies, the majority of contemporary phylogenetic analysis is carried out on data obtained from either nucleic acid sequences (DNA or RNA) or amino acid sequences (*polypeptides* or *proteins*). In either case, the data consist of a string of characters from a fixed alphabet. In the case of nucleic acids, the alphabet consists of four characters, most commonly denoted A, C, G, and T. The encoded protein alphabet is larger, consisting of 20 amino acids. In subsequent discussion, we will discuss nucleic acid sequences exclusively, as the models are much more concise. However, the reader should keep in mind that a similar analysis can be carried out on polypeptide sequences as well, once the characteristics of the amino acid alphabet is taken into account.

When an organism reproduces, the DNA sequences found in the parents' germ-line cells are copied. These sequences may be preserved without change, or they may undergo a mutation. At any given position in the sequence, a number of things may occur during a mutation. A character may be *substituted* for another character, for example an A may change to become a C. A character may be *deleted*, shortening the sequence overall. Or, finally, one or more characters may be *inserted*, lengthening the sequence. A pair of characters in two sequences are *orthologous* if both are descended from the same ancestral character.

The possibility of the latter two types of mutation means that any time we wish to compare multiple sequences we may be first required to *align* them. Sequence alignments are an attempt to impute the unknown orthology relationships in a set of sequences. Aligning multiple sequences is known to be a very difficult problem; a NP-hard problem, to be precise [84]. However, despite the difficulty of the alignment problem, there are numerous heuristic methods which appear to produce reasonable results, and it is generally assumed that sequences used for phylogenetic reconstruction are aligned correctly. At the time of this writing, `MUSCLE` [44] and `MAFFT` [85] are among the most commonly used tools to obtain sequence alignments.

Once we have a sequence alignment in hand, a natural next question is how one might measure the (dis)similarity between the observed sequences. When observing data that take the form of real numbers, there is usually an obvious way to quantify the distance between two values, and most often the method is subtraction. However, sequences present a much more complicated problem, and there have been numerous ways proposed to quantify the differences between them.

The most common approach to modeling molecular evolution probabilistically is to treat the evolution of each character as an independent continuous-time Markov process. A Markov process is a stochastic process that has the 'memoryless' property: The future evolution of the process is conditional only on the current state of the system. In particular, it is independent of any behavior in the past, given the current state.

This model is motivated by the assumption that when a sequence is duplicated, the character in the new sequence is randomly selected from a distribution that depends only on the current state of the character. The character most likely remains unchanged, but substitutions are a possibility, and the probabilities of various substitutions depend only on the chemical dynamics of sequence replication. These two assumptions, at least, seem quite reasonable: the biochemistry underlying biological sequence replication is believed to be inherently stochastic in nature; and there is no known mechanism by which the state of a genetic sequence in the past could influence the duplication of genetic material in the present time.

Since evolution is assumed to take place slowly over a very large number of generations, and because it simplifies the calculations, it is customary to model time as a continuous variable, rather than by counting discrete generations. The dynamic behavior of such a system is determined by a *transition rate matrix*, typically denoted Q, which describes the rates at which the different types of substitution occur, and the initial character state. The rate matrix may change over time, or it may be constant, with the latter being a common simplifying assumption. Although we will, for the sake of brevity, make this assumption when discussing the following models, it can be relaxed at the expense of increased computational difficulty and model complexity.

One further important simplifying assumption is that the entire relevant period of evolution under study is understood to be stochastic fluctuation about the Markov Chain's equilibrium distribution (a concept we will introduce shortly). The biological implication of this assumption is that we are modeling only what is called *neutral evolution*, those mutations in the genome which become fixed in the population purely by chance, and do not confer any selective [dis]advantages. In particular, we are not modeling the directed processes collectively known as natural selection, that are often implied by the term 'evolution'.

A final limitation worth mentioning, which applies to all of the models we will discuss, is that in practice these models cannot be used to simultaneously estimate both the overall rate of base substitution and the amount of time that the Markov process has been evolving. Much of what we would like do with these models involves attempting to estimate the amount of time that a Markov process has evolved, given only observations of the beginning and ending states. However, the overall rate of substitution and the passage of time are intertwined in such a way that without imposing additional assumptions, it is only possible to estimate their product, which measures the mean number of substitution events expected to occur per site. Nevertheless, this quantity is often referenced simply as "time". Additional assumptions, such as a molecular clock, can allow one to estimate the passage of time directly, but it turns out that such measures are rarely necessary, and the assumptions needed are highly suspicious.

1.2.1 Introduction to Markov Processes

The behavior of a continuous-time Markov process on a state space with n elements, is governed by a $n \times n$ transition rate matrix, Q. The off-diagonal elements of Q represent the rates governing the exponentially distributed variables that are used to describe the amount of time that elapses before a particular type of base substitution occurs. The ij-th element of Q represents the rate at which characters in the i-th state are replaced with the j-th state. The rates are typically expressed with respect to a dimensionless "time" variable, usually denoted t. The diagonal elements of the rate matrix must be set such that every row in the matrix sums to zero.

The Q matrix can be used to compute a transition probability matrix, $P(t)$. This probability matrix, gives the probability that a character in the i-th state at the present time will be in state j after the passage of time t. If we use $X(t)$ to denote the state of a character time t in the future, then $\mathbb{P}[X(t) = j|X(0) = i] = P_{ij}(t)$. The transition probability matrix is related to the rate matrix by the matrix exponential,

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

An interesting property of these types of stochastic processes is that for certain classes of rate matrices, $P(t)$ converges to a fixed matrix as $t \to \infty$, and furthermore the rows of the limiting matrix may all be identical to a single vector, which we will denote π. When this occurs, it implies that behavior of the process at large distances is independent of the starting state of the system; for every possible starting state, in the far future the distribution governing the character state has the probability masses specified by the vector π. This limiting distribution is called the *stationary-state* (or *equilibrium*) distribution.

An interesting and useful property arises if a Markov process has a stationary distribution π satisfying the following relationship with the rate matrix Q,

$$\pi_i Q_{ij} = \pi_j Q_{ji}, \forall i, j. \tag{1.1}$$

This is known as the *detailed balance* equation, and when it holds the process is *reversible.* If a process is reversible, it means that once it has converged to the equilibrium distribution, the "arrow of time" disappears: there is no way to determine if a character's process $X(t)$ is indexed in the 'proper' direction, or if the time index has been reversed.

Reversibility turns out to be a desirable property if we want to study molecular evolution. Consider a most-recent common ancestor, with two daughter lineages. If the processes describing sequence evolution are reversible, then we do not need to consider the two lineages separately. The reversibility means that we can treat the daughters as endpoints of one long Markov chain that goes 'up' one lineage to the MRCA, and back 'down' the other lineage. If the model was not reversible, then this would not be a valid simplification. We would need to model each lineage discretely, in the 'correct' orientation from ancestor to descendant. All of the commonly used models are reversible, greatly reducing the complexity of many calculations.

1.2.2 The Jukes-Cantor Model

The simplest model of DNA evolution is the Jukes-Cantor (JC or JC69) model. In addition to the Markov process assumptions, it also assumes that there is a single transition rate that governs all types of substitution [83]. This assumption implies a transition matrix of the form,

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}.$$

We index the rows and columns of matrices for nucleotide models in the following order: adenine, guanine, cytosine, and thymine/uracil. In this case it does not matter, but in subsequent models this becomes important.

It turns out that if one takes the matrix exponential of tQ under the limit $t \to \infty$, one finds that a stationary distribution for the JC model exists, and is uniform on all possible character states, $\pi = (1/4, 1/4, 1/4, 1/4)$.

Recall that $P_{ij}(t) = \exp(tQ)_{ij}$ represents the probability of transitioning to state j when beginning in state i, if the time separating the sequences is t. Since the model is reversible, we can use this matrix to look up the likelihood of observing a particular pair of characters in a sequence alignment, assuming the sequences are separated by a time t. If the sites in the alignment are independent (as the JC model assumes), then the likelihood of the entire alignment is the product of the individual site likelihoods. That is, if the character pair i, j occurs in the alignment n_{ij} times, then the likelihood of the entire alignment is given by,

$$L(t) = \prod_{\forall i,j} P_{ij}(t)^{n_{ij}}.$$

For a variety of reasons, both theoretical and relating to numerical stability, it is more common to work with the log-likelihood,

$$l(t) = \log L(t) = \sum_{\forall i,j} n_{ij} \log P_{ij}(t).$$

We are now in a position to use a sequence alignment to estimate the evolutionary time separating the two sequences. We will do this by attempting to find the time t which maximizes the likelihood, given the observed alignment data. It turns out that it is fairly easy to obtain a closed-form solution for the JC problem. This solution is presented in many places (see, e.g., Pachter and Sturmfels [125]), and we will not discuss it at length here, except to note the expression used to estimate the JC distance between two sequences is

$$t = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p\right).$$

In this expression, the sufficient statistic p is the proportion of sites within the alignment which have different characters.

Although simple to understand and analyze, the Jukes-Cantor model, makes assumptions about the nature of DNA evolution that are considered to be unreasonable by many biologists. As a result, an increasingly complicated series of models has been developed which attempt to better accommodate the properties of empirical sequences.

1.2.3 The Kimura 2-parameter Model

Kimura's two-parameter model (K80) relaxes the JC assumptions by allowing for two different substitution rates. The motivation for this generalization is based in the fact that nucleic acid residues can be divided into two classes based on their chemical structure and properties, the purines (adenine and guanine) and the pyrimidines (cytosine and thymine/uracil). Empirical results suggest that it is significantly more likely for a replication error to result in a substitution with a nucleotide from the same class as the original nucleotide ($A \leftrightarrow G$ or $C \leftrightarrow T$). These types of substitutions are called transitions, whereas substitutions with a nucleotide from the other class (all other types of substitutions) are called transversions. For the K80 model, the matrix Q takes the form [184],

$$Q = \frac{1}{\kappa + 2} \begin{pmatrix} - & \kappa & 1 & 1 \\ \kappa & - & 1 & 1 \\ 1 & 1 & - & \kappa \\ 1 & 1 & \kappa & - \end{pmatrix},$$

where κ is a parameter controlling the ratio of the rate of the transition to that of transversion. It is fairly common to suppress the diagonal elements of transition matrices, because they are completely determined by the off-diagonal elements (so that the rows sum to zero), and are often complicated expressions which do not serve to convey any important information about the system to a human reader.

Like the JC model, the K80 model is reversible and has a stationary distribution that is uniform on all possible nucleotides. It also is relatively easy to obtain a closed-form solution to the maximum likelihood problem. If we let p be the proportion of sites showing a transition substitution and q be the proportion showing a transversion, then a closed form estimate of the K80 distance between two sequences is estimating using the following expression [183],

$$t = -\frac{1}{2}\ln(1 - 2p - q) - \frac{1}{4}\ln(1 - 2q).$$

1.2.4 The Hasegawa, Kishino, and Yano 1985 Model

The Hasegawa, Kishino, and Yano (HKY) model is a further generalization of the K80 model. It introduces additional parameters (the vector π) which allow the stationary distribution of character frequencies to depart from uniform. This is an important degree of flexibility, because base frequencies are known to vary significantly in nature, both between organisms as well as within a single genome. For example, in the complete 12 million character genome of common baker's yeast (*Saccharomyces cerevisiae*) the base frequencies vary from 19% each for cytosine and guanine, to 31% each for adenine and thymine [10]. Such a significant variation from a uniform distribution cannot reasonably be attributed to chance, and needs to be accommodated for by the model.

The rate matrix for the HKY model is

$$Q = \beta \begin{pmatrix} - & \kappa\pi_g & \pi_c & \pi_t \\ \kappa\pi_a & - & \pi_c & \pi_t \\ \pi_a & \pi_g & - & \kappa\pi_t \\ \pi_a & \pi_g & \kappa\pi_c & - \end{pmatrix}.$$

In this expression, κ is again a parameter describing the ratio of the rate of transitions to that of transversions, π_i are the equilibrium base frequencies for $i \in \{A, C, G, T\}$, and β is a constant which normalizes the rate to one overall. Although we could try to simultaneously optimize the log-likelihood over t, κ, and π, it is more common to estimate the base frequency distribution separately, and optimize only the parameters t and κ. The reason for this is that we often want to compute pairwise distances between all sequences in a multiple sequence alignment, and in this case the equilibrium frequencies must be shared by all sequences. Thus, it makes more sense to estimate the base frequencies once, using all of the available sequence data. This leads to more precise estimates of the base frequencies, as well as enforcing a common equilibrium distribution for the entire multiple alignment.

For models which allow base frequencies to depart from the uniform distribution, the rate normalizing constant for the rate matrix, β, is more complicated than in previous models. As before, it must be chosen so that the average rate of substitution is 1, but it must take into account the non-uniform base frequency distribution found in the alignment. For any reversible Q matrix, the normalizing constant is given by the following equation [98],

$$\beta = -1/\sum_i \pi_i Q_{ii}.$$

1.2.5 The Tamura-Nei 1993 Model

The Tamura-Nei 1993 (TN93) model expands slightly on the HKY model by adding a third substitution rate category. The three rate classes in the TN93 model are: $A \leftrightarrow G$ substitutions, $C \leftrightarrow T$ substitutions, and transversion substitutions. Thus, the TN93 rate matrix has the form,

$$Q = \beta \begin{pmatrix} - & \kappa_1\pi_g & \pi_c & \pi_t \\ \kappa_1\pi_a & - & \pi_c & \pi_t \\ \pi_a & \pi_g & - & \kappa_2\pi_t \\ \pi_a & \pi_g & \kappa_2\pi_c & - \end{pmatrix},$$

where κ_1 and κ_2 are parameters for two different types of transition and π_i is the base frequency of the state i.

1.2.6 The General Time Reversible Model

The General Time Reversible Model (GTR) model is the most flexible model of nucleotide substitution which preserves the time-reversibility of the Markov process [168]. It allows for all types of character substitution to occur at a distinct rate, as well as allowing for arbitrary equilibrium frequencies. See Tavaré [168] for the details.

1.2.7 Common Model Extensions

There are a few extensions that are sometimes added to any of the previously mentioned substitution models. These extensions are motivated by features commonly observed in empirical sequences that are not well fitted by any of the probabilistic models discussed thus far.

The first model extension allows for certain character sites to be classified as *invariant*. An invariant site is one where all substitutions are forbidden. This is motivated by the assumption that certain positions in a sequence are more important to the sequence function than others, and thus these sites experience strong purifying selection. A simple example of such a site is the region of a sequence which initiates or terminates protein translation. If these regions are disturbed by a mutation, there is little chance that the biological function of the sequence will be preserved in any meaningful way. Thus, mutations at these sites are assumed to be almost totally forbidden.

The second important possible model extension, Γ *rate categories*, is intended to account for the fact that different sites in a sequence might evolve at different rates overall. For example, DNA sequences are translated into amino acid sequences in in groups of three characters at a time. (e.g., the DNA sequence `ATG` translates into the amino acid methionine.) These sets of DNA base triplets are called *codons*. There are $4^3 = 64$ possible codons, but only 20 amino acids and a translation termination signal need to be encoded. The encoding is therefore redundant, with each amino acid encoded by an average of 3 different codons. (However, the redundancy varies from only a single encoding for the cases of methionine or tryptophan, to six encodings each for arginine, leucine, and serine.)

An interesting fact about the encoding is that the codons are not assigned to the amino acids in a random manner. When multiple codons encode a single amino acid, it is quite likely that the redundant encodings share common first and second characters, only varying in the third position. Conversely, changes at the second position are almost certain to result in a change in the translated sequence. For example, consider the codon `CTT`, which encodes the amino acid leucine. The third character can be freely substituted and the new codon will still translate to leucine. However, a substitution at the second position always changes the encoded amino acid.

Thus, substitutions at the second position in a codon should be subject to greater selective pressures than changes at the third position. These differences in selective pressure between the three codon positions should logically lead to differences in the overall substitution rate at each position. In the Γ model extension, a mixture of several scaled rate processes is used to model these disparities in substitution rate.

When calculating substitution probabilities, each category is allowed to scale the substitution rate matrix, Q, by a different constant, with the constraint that the combined total rate of substitution across all sites must remain equal to 1. The name of the extension is a reference to the fact that the scaling constants are obtained from quantiles of a mean-1 Gamma distribution. The user typically must specify both the number of categories as well as a constant (often named α, or the "shape parameter") which governs the variance of the mean-one Γ distribution used.

The use of these model extensions is typically indicated by the presence of "+I" or "+G" after a model code. For example, HKY+I+G means the HKY model was used with both the invariant sites and the Γ categories extensions. For more information, see Pachter and Sturmfels [125].

1.3 Reconstructing the Tree

Systematists often make inferences about the phylogenetic tree relating a set of organisms using a sequence alignment as input data. Many algorithms have been proposed for accomplishing this task, some are based explicitly in statistical methodology, whereas others are justified in other ways. In this section we briefly introduce several classes of methods for reconstructing a phylogenetic tree from a sequence alignment.

1.3.1 Distance-Based Methods

The *distance-based methods* of tree reconstruction work by first computing a pairwise distance matrix for the sequences in an alignment. The tree is then produced by a second algorithm, using only the distance matrix as input. This is in contrast to the other methods we shall discuss, which involve sequence alignment directly in the tree reconstruction algorithms.

A distance matrix is a non-negative, square, symmetric matrix with elements corresponding to estimates of some pairwise distance between the sequences in a set. The simplest definition distance uses the proportion of homologous sites in an alignment with differing characters, and is called the p-distance, or *Hamming* distance. Although the p-distance is simple to calculate and understand, it is more common to use one of the probabilistic definitions of evolutionary distance discussed in the previous section when producing the distance matrix.

Neighbor Joining.— The neighbor joining (NJ) method of tree reconstruction begins with a completely unresolved (star) tree and attempts at each step to further resolve the tree by adding a node which joins the most closely related nodes in the tree, as determined by a distance matrix. A new distance matrix is then computed where the rows and columns associated with the two newly joined taxa are replaced with new entries relating to to the new interior node, and the process is repeated until the tree is fully resolved.

The NJ method was developed by Saitou and Nei [149], and has been discussed extensively in other publications. We will not present details of the algorithm here, but rather refer interested readers to Haws et al. [64] for a fuller discussion. The related BIONJ algorithm of Gascuel [56] claims to offer improved performance when used on highly divergent alignments, as well as being capable of handling distance matrices with missing elements.

The NJ algorithm is among the fastest available methods of tree reconstruction. However, it does suffer from some drawbacks; particularly problematic is a lack of statistical consistency in certain situations. (A method is statistically consistent if it is almost certain to converge on the correct tree as the alignment length grows to infinity.) For these reasons, NJ is most often used to quickly obtain reasonable trees which can be used as starting locations for more computationally intensive tree reconstruction algorithms, such as the Maximum Likelihood or Bayesian methods.

Balanced Minimum Evolution.— Balanced minimum evolution is a tree reconstruction method which is roughly analogous to the least-squares method of fitting curves to observed data points [39]. As the name suggests, candidate trees produced by the BME computations are compared using the sum of their branch lengths (a measure of the total amount of evolution required to produce the tree), with smaller trees being considered superior. The principle of Occam's razor is typically cited as a rationale for this method of comparison.

The BME method describes a method of assigning lengths to the branches of an arbitrary tree topology in a way most compatible with a given (fixed) distance matrix, and which takes into account the fact that the variance of the pairwise distance estimates is smaller for closely related sequences than for highly divergent ones. Finding the optimal tree then involves finding the tree topology on which the total sum of the branch lengths is minimized. Fortunately, there is a simple and fast method of computing the total length of the branches on any given topology (actually computing all branch lengths is not required), known as Pauplin's Formula.

Unfortunately, unless the number of taxa in the tree is very small, a complete census of the possible topologies is infeasible. Not only is a complete search of the space of topologies computationally intractable, but there is no known way to reduce the computational complexity of the BME search to a reasonable level while also guaranteeing that the optimal solution is found [36]. Despite this, several fast heuristic algorithms have been developed that provide fairly good solutions to the BME problems, but without the guarantee that the globally optimal topology has been found. Desper and Gascuel [39] is an example of one such algorithm.

1.3.2 Maximum Parsimony

The maximum parsimony (MP) method, like BME, is a method that attempts to select a tree topology by minimizing the amount of evolution required to explain the inferred alignment. As such, it shares the drawbacks associated with the need to search the entire space of possible tree topologies. (Namely, there is no known method of easily obtaining a definitive solution.) However, unlike BME, it is not a distance-based method, but rather uses the entire sequence alignment in the calculation.

The principle underlying the MP tree is simple: For any given topology we assign sequences to the internal nodes of the tree in a way that minimizes the total number of base substitutions that are required to occur on the entire tree. This total number of base substitutions is used as the criterion by which a tree topology is selected. It should be noted that the MP method has the additional advantage of producing an estimate of the ancestral sequences as a byproduct of the computation, something that the distance-based methods do not do.

1.3.3 Methods explicitly based on probability models

The tree reconstruction methods discussed up to this point are not explicitly based on any probabilistic models of sequence evolution, although they may do so implicitly through the pairwise distance matrix. Conversely, the methods in this section are explicitly based on probability models, and the techniques used to obtain phylogenies are similarly grounded in statistical theory.

Maximum Likelihood.— The maximum likelihood (ML) methods of tree estimation share much of the theoretical machinery introduced in Section 1.2. However, instead of attempting to model a single Markov process connecting a pair of homologous characters, the ML methods posit a collection of Markov processes, one for each branch on a tree. Given a tree topology and an alignment, the ML methods attempt to assign branch lengths to the tree in such a way that the overall likelihood of the collection of Markov processes is maximized.

The calculation of the likelihood over an entire tree is significantly more complicated than in the case of a single pair of sequences, since the effect on the likelihood of branch lengths and internal character states are all highly interdependent. Fortunately, Felsenstein [50] developed an efficient *pruning algorithm* which greatly simplifies the calculation of the likelihoods.

Like the BME and MP algorithms, a ML tree is produced by searching for the combination of topology and branch lengths that maximize the likelihood value. Although no fast algorithm is known which is guaranteed to locate the global maximum value, several heuristic search methods are commonly used to explore the space of possible trees.

Bayesian Methods.— The primary alternative statistical approach to tree reconstruction is the Bayesian method. Like all Bayesian methods, the basic premise is to make inferences based on a *posterior* distribution of the relevant model parameters. In this case, the tree topology and branch lengths. The posterior distribution is computed from a model of substitution and a *prior* distribution π on the model parameters, using Bayes formula for reversing a conditional probability. Within the Bayesian framework, the prior distribution encodes the user's prior beliefs about what form the correct tree might take. However, in practice, little progress has been made in allowing biologists to reasonably specify priors on the space of trees. This lack of ability to specify priors has resulted in most papers utilizing one of a few basic prior distributions. The most common is probably a uniform distribution on the topology combined with exponential distributions for the edge lengths.

Although direct computation of the posterior distribution is usually impossible, due to the presence of a thoroughly intractable integral over the entire space of tree topologies, there is a fairly good method of obtaining a sample from the posterior known as the Metropolis-Hasting algorithm. The Metropolis-Hasting algorithm describes a method of implementing a discrete time Markov chain which generates (correlated) samples from arbitrary density functions, even if the normalizing constant for the density function is unknown [98]. This algorithm is one of the main workhorses in the class of methods known as *Markov Chain Monte Carlo* (MCMC).

The MCMC methods generate a sample of trees from the posterior distribution, and then use this sample as the basis of inferences about the true tree [78]. For example, a common Bayesian method of inferring the topology of a tree is to select the topology which occurs most commonly in the posterior sample, i.e., the *posterior mode.* The programs BEAST [16] and Mr. Bayes [78] are the most commonly used implementations of Bayesian tree reconstruction at the present time [43].

1.4 Model Selection

In Section 1.2, I presented a number of possible probability models of sequence evolution. The choice of probability model has the potential to change radically, not only the branch lengths, but also the topologies of any reconstructed trees. This naturally leads to the question: Which substitution model should one use when reconstructing a tree?

The question of model selection is a problem that has a rich history in statistical literature. Although one can always improve the fit to observed data by adding degrees of freedom to the model, such flexibility comes at a cost: a decrease in the precision of the model parameter estimates, and an increase in the amount of computational effort required to obtain them. Although a model which is too simple will typically result in biased inferences, a model which is too complex also often fails to be useful. In such "overfitted" models, the behavior of the parameter estimates is dominated by the statistical noise present in the data, making it difficult to observe any systematic patterns that may be present. In addition, models with many parameters are difficult for a human to meaningfully interpret, and usually perform quite poorly when used for predictive purposes, greatly limiting their utility.

Two popular methods of model selection, both of which have long histories in statistics are the Likelihood Ratio Test (LRT, also known as a χ^2 test, due to the asymptotic distribution of the test statistic) [119], and the Akaike Information Criterion (AIC) [3]. A third popular method, the

Bayesian Information Criterion (BIC), is sufficiently similar to AIC, both in calculation and use, that we will omit any further discussion of it in the interest of brevity [26].

The Hierarchical LRT is essentially a forward-selection method. Beginning with the simplest JC model, a sequence of increasingly complicated models is proposed, and a test is performed to indicate if the resultant improvement in the fit of the model, as determined by the log-likelihood, is sufficient to justify the increase in the number of parameters. One drawback of the LRT method is that the series of models to be tested must nest into each other in some way, with each model in the sequence being a generalization of the previous one.

The use of the AIC in model selection is simpler, because the requirement that the models nest within each other is not needed. Any model for which we can obtain a likelihood can be tested against any other, and the model with the smallest AIC value is deemed the best.

1.5 Distances between trees

As in the case of sequence analysis, there is no obvious answer to the question of how to quantify the differences between a pair of phylogenetic trees. A tree distance is a function, $d : \mathcal{T}_n \times \mathcal{T}_n \to \mathbb{R}_+$ that has, at a minimum, the properties $d(T, T') = d(T', T)$ and $d(T, T) = 0$, for all $T, T' \in \mathcal{T}_n$. There are numerous ways to define distances on the space of possible phylogenetic trees; some of these methods have convenient analytic or computational properties, while others have more natural biological interpretations.

Figure 1.2 depicts a pair of unrooted trees on five taxa. These trees T_1 and T_2 will be used as examples to demonstrate the calculations for the methods presented subsequently. If a distance measure does not require branch length information, then they may be ignored.

In the subsequent descriptions, we use n to denote the number of terminal taxa (or leaves) in the tree. The space of all possible trees on n taxa is called $\mathcal{T}_n$. This space may or may not incorporate branch length information, and the trees may or may not be rooted, depending on the context. We use $||\cdot||$ to represent the usual Euclidean length of a vector, and $|\cdot|$ to indicate the cardinality of a set. The symmetric difference between two sets is defined as $A \ominus B := (A\backslash B) \cup (B\backslash A)$. Many methods also require a vectorization function, $v : \mathcal{T}_n \to \mathbb{R}^p$, for some p, mapping phylogenetic trees into Euclidean space.

Splits.— A *split* is a bipartition of the set of leaves of a tree. A tree T is said to contain split s, if it is possible to remove an edge from T and form two sub-trees with leaves matching the bipartition specified by s. For example, $s = \{abc|de\}$ (sometimes shortened to simply *abc* when the context is clear), is one possible split of the leaves from the example trees. If we consider the example trees introduced in Figure 1.2, we find that split s is contained in T_1, since by removing the branch with length 2.2, two trees with the correct partitioning of leaves are formed. Conversely, s is not found in T_2; there is no way to remove a branch and obtain the desired bipartition.

Quartets.— Given any tree with more than 4 leaves, it is possible to form subtrees by pruning the tree down in various ways until it contains only 4 leaves. Such a subtree is called a *quartet* of the original tree. Each quartet contains a single non-trivial split, obtained by removing the single interior branch. Since there are three possible unrooted topologies for each quartet, the splits which they define carry information about the topology of the complete tree.

We say that a tree T contains the quartet $\{ab|cd\}$ if the quartet comprising the listed nodes contains the given split. For example, T_1 contains the quartets $\{ab|de\}$ and $\{ab|cd\}$, but not the quartet $\{cd|be\}$. T_2, on the other hand, contains all three of these quartets.

1.5.1 Squared Euclidean Distances

A tree distance $d(\cdot, \cdot)$ is *squared Euclidean* if there is a vectorization function v and a positive constant c, such that for all $T, T' \in \mathcal{T}_n$ the following relationship holds,

$$d(T, T') = c \cdot ||v(T) - v(T')||^2. \tag{1.2}$$

Several popular tree distances are squared Euclidean distances as will be demonstrated below.

Robinson–Foulds distance.— Let $S(T)$ denote the set of splits found in tree T. The normalized Robinson-Foulds (RF) distance is defined as half of the size of the symmetric difference between the sets of splits for the trees,

$$d_{RF}(T,T') := \frac{1}{2}|S(T) \ominus S(T')|.$$

The RF distance is a squared Euclidean distance, since we may define a vectorization function $v_{RF}\colon \mathcal{T}_n \to \mathbb{R}^{2^{n-1}-1}$ where the components of v_{RF} form an enumeration of the indicator functions on all possible tree splits. In other words, for each possible split of the leaves, $A|A^c$, there is an element of $v_{RF}(T)$ which is 1 if $A|A^c \in S(T)$, and zero otherwise. Thus, subtracting $v(T) - v(T')$ gives a vector where each element is non-zero if, and only if, the the corresponding split is contained in one tree, but not the other. It should be clear that the squared magnitude of this vector satisfies is equivalent to the number of such splits, and is thus equivalent to the Robinson–Foulds distance.

Revisiting the example trees, if the coordinates of v_{RF} are associated with the possible splits in the following way,

$$(a, b, c, d, e, ab, bc, ac, cd, bd, ad, de, cd, bd, ae),$$

then the trees T_1 and T_2 from Figure 1.2 are vectorized

$$\begin{aligned} v_{RF}(T_1) &= (1,1,1,1,1,1,0,0,0,0,0,0,0,0,1), \\ v_{RF}(T_2) &= (1,1,1,1,1,1,0,0,0,0,0,0,1,0,0). \end{aligned}$$

With these vectors, it is simple to calculate the normalized RF distance by applying Equation 1.2 and we find $d_{RF}(T_1, T_2) = 1$.

Note that it is somewhat common for authors or programs (e.g., PHYLIP) to define the RF distance as the size of the symmetric difference *without* the normalizing constant $\frac{1}{2}$. When comparing RF values from different sources it is important to determine if the conventions used are compatible.

Quartet distance.— Let $Q(T)$ be the set of quartets in a tree T. The quartet distance [48] is defined as a half of the size of the symmetric difference of quartets,

$$d_Q(T,T') := \frac{1}{2}|Q(T) \ominus Q(T')|.$$

As in the case of the RF distance, d_Q can be written as a squared euclidean distance such that using a vectorization function $v_Q : \mathcal{T}_n \to \mathbb{R}^{3\binom{n}{4}}$. This function maps tree T to the 0/1 vector $v_Q(T)$ whose entries are indicator functions of all possible quartet splits in T. For example, if the coordinates of v_Q are ordered in the following way,

$$(ab|cd,\ ac|bd,\ ad|bc,\ bc|de,\ bd|ce,\ be|cd,\ ab|ce,\ ac|be,\ ae|bc,\ ac|de,\ ad|ce,\ ae|cd,\ ab|de,\ ad|be,\ ae|bd),$$

then our example trees T_1 and T_2 from Figure 1.2 are vectorized as,

$$\begin{aligned} v_Q(T_1) &= (1,0,0,1,0,0,1,0,0,1,0,0,1,0,0), \\ v_Q(T_2) &= (1,0,0,0,0,1,1,0,0,0,0,1,1,0,0). \end{aligned}$$

The distance between T_1 and T_2 is easily computed to be

$$d_Q(T_1,T_2) = \frac{1}{2}||v_Q(T_1) - v_Q(T_2)||^2 = 2.$$

Dissimilarity map distance.— Given any tree T of n leaves with branch length information, one may produce a corresponding *distance matrix*, $D(T)$. The distance matrix is a $n \times n$ symmetric matrix of non-negative real numbers, with elements corresponding to the sum of the branch lengths between pairs of leaves in the tree. To calculate $D_{(ij)}(T)$, one simply determines which edges of the tree form the path from leaf i to leaf j, and then sums the lengths of these branches.

Since $D(T)$ is symmetric and has zeros on the diagonal, the upper-triangular portion of the matrix contains all of the unique information found in the matrix. We can vectorize T by enumerating this unique portion of the distance matrix,

$$v_D(T) := (D_{12}(T), D_{13}(T), \ldots, D_{23}(T), \ldots, D_{n-1,n}(T)).$$

The *squared dissimilarity map distance* is defined to be

$$d_D(T', T) := ||v_D(T) - v_D(T')||^2.$$

A discussion of the dissimilarity map distance can be found in Buneman [25].

If we order the columns and rows of the distance matrix alphabetically, then the example trees are vectorized as

$$\begin{aligned} v_D(T_1) &= (2.6, 5.0, 7.0, 6.8, 5.2, 7.2, 7.0, 5.0, 4.8, 2.4), \\ v_D(T_2) &= (3.6, 5.1, 7.4, 6.8, 5.7, 8.0, 7.4, 5.5, 4.9, 2.8). \end{aligned}$$

From this point, computing the distance between the trees is simple,

$$d_D(T_1, T_2) = ||v_D(T_1) - v_D(T_2)||^2 \approx 2.64.$$

Path difference.— The RF and Quartet distances are completely determined by the topologies of the trees, ignoring any edge lengths that may be present. Conversely, the dissimilarity map distance requires that the edge lengths be defined. The *path difference* distance d_P is a distance analogous to the dissimilarity map, but which does not require edge length information.

The calculation of the path difference is identical to the dissimilarity map, except that elements in the distance matrix $D(T)$ are determined by counting the number of edges between the leaves, rather than summing the edge lengths. (This is equivalent to the dissimilarity map distance with all edge lengths in the tree set equal to 1.) The path difference is studied and compared with the RF distances by Steel and Penny [161].

Using the same vector ordering as in the dissimilarity map example, we find that the path difference vectorizations of our example trees are

$$\begin{aligned} v_p(T_1) &= (2, 3, 4, 4, 3, 4, 4, 3, 3, 2), \\ v_p(T_2) &= (2, 4, 4, 3, 4, 4, 3, 2, 3, 3). \end{aligned}$$

The path difference is therefore, $d_p(T_1, T_2) = ||v_p(T_1) - v_p(T_2)||^2 = 6$.

1.5.2 Tree rearrangement distances

All of the tree distances discussed so far can be understood in terms of vector magnitudes in some Euclidean space. The distances discussed in this section are defined in a different way: Given a certain class of tree rearrangement operations, the distance between two trees is the minimum number of steps needed to transform one tree into another. These distances are all topological distances; they ignore any edge length information the trees may contain.

Nearest-Neighbor-Interchange distance.— For each internal branch in a tree, there are three possible configurations for the connected subtrees, as shown in Figure 1.3. A NNI operation operation (also known as a tree *rotation*) makes a small change to the topology of the tree by exchanging two adjacent subtrees, forming one of the alternative topologies [144]. The NNI distance between two trees is the minimum number of such moves required to transform one tree into the other. Although conceptually simple, computing the NNI distance is a NP-hard problem [35].

For instance, each of the example trees can each be rotated about the length 2.2 branch, exchanging the leaves c and e, to form the other tree. Thus, the trees are separated by a NNI distance of 1.

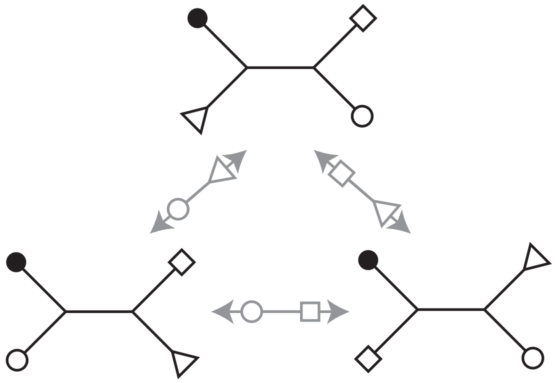


Figure 1.3: A nearest-neighbor interchange (NNI) move begins by selecting an internal branch from the tree. The selected internal branch defines four subtrees, which are represented in simplified form by different shapes. These four subtrees may be arranged in one of three possible topologies. The NNI move is completed by exchanging two of the subtrees, forming one of the alternative topologies.

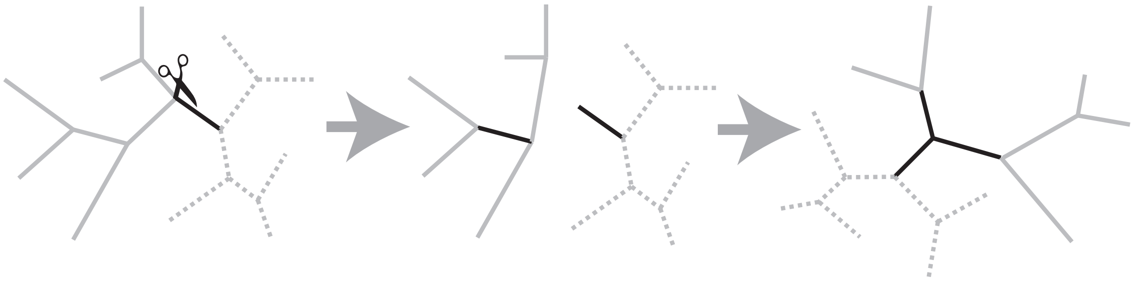


Figure 1.4: A subtree-prune-and-regraft (SPR) move: [Left] A subtree is selected and pruned from the main tree. [Middle] A branch is chosen from the main tree to receive the subtree. [Right] The subtree is regrafted onto the main tree.

Subtree-Prune-and-Regraft distance.— Like the NNI distance, the *Subtree-Prune-and-Regraft (SPR)* distance is defined by a minimum number of operations required to transform one tree into another. The steps of a SPR move are depicted in Figure 1.4. Succinctly, a subtree is pruned from the main tree, and then reattached to the middle of an edge elsewhere in the tree. An NNI move is a special case of an SPR move, where the detached subtree can only be moved to one of two possible locations.

Unfortunately, computing the SPR distance also is a NP-hard problem [71]. However, since a NNI move is also a SPR move, we know that that the example trees are also separated by SPR distance 1.

Tree-Bisection-and-Regrafting distance.— A further generalization of SPR, *tree bisection and regrafting* (TBR) operations on trees can also be used to define tree distances [155]. In a TBR operation the pruned subtree can be reattached to the main tree in a more general fashion than in a SPR move. An example of a TBR move is depicted in Figure 1.5. Like the other distances in this section, the TBR distance is defined as the minimum number of such moves required to transform one tree into another. Computation of the TBR distance is also NP-hard [4]. However, in the case of our example trees, the distance is also 1, since TBR is a generalization of both SPR and NNI.

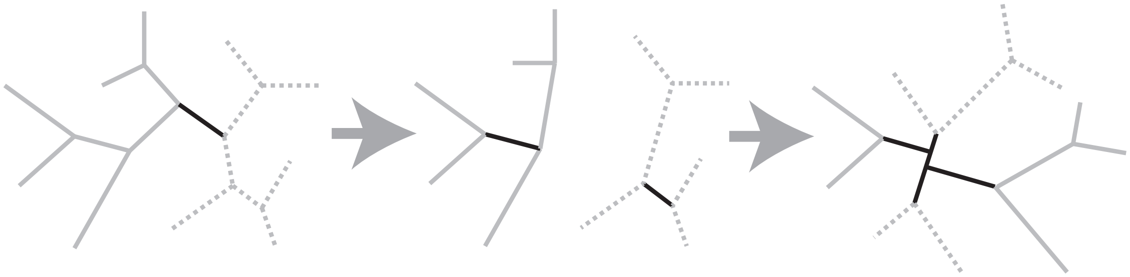


Figure 1.5: A tree-bisection-and-regraft (TBR) move: [Left] The tree is bisected by removing an internal branch. [Middle] Two branches are chosen from the resulting subtrees. [Right] The two branches are regrafted together in one of the three possible topologies.

Disagree distance.— Steel and Penny [161] also describe a quantity they name the *disagree distance*. This distance is defined by the minimum number of taxa that must be removed from the phylogeny before the trees become congruent. For example, for our example trees are separated by a disagree distance of 1, because removing any one of the taxa c, d, or e, results in the trees being topologically congruent. Puigbó, Garcia-Vallvé, and McInerney [138] present an algorithm for computing the disagree distance.

1.5.3 Other tree distances

This section contains a few other notable tree distances which do not fall into any of the previous categories.

Matching splits distance.— A recently introduced distance is the *matching splits distance*, developed by Bogdanowicz and Giaro [14]. Roughly speaking, the matching splits distance refines the RF distance by allowing splits to partially match each other when a portion of the split is shared by both trees.

Maximum Parsimony distance.— Fischer and Kelk [52] introduced a notion of the *Maximum Parsimony (MP) distance* between phylogenetic trees. The MP distance between trees is simply the difference between the MP scores of the given trees.

1.6 Billera-Holmes-Vogtmann Treespace

Billera et al. [13] introduced a continuous metric space which can be used to model the set of phylogenetic trees with edge lengths on a fixed set of leaves. The Billera-Holmes-Vogtmann (BHV) tree space is not Euclidean, but it is non-positively curved. Such a space is known as a CAT(0), or *Hadamard*, space, and such spaces have the property that any two points are connected by a unique shortest path through the space, called a *geodesic*. The distance between two trees is defined as the length of the geodesic connecting them.

Consider a rooted tree with n leaves. Such a tree has at most $2n-2$ edges; there are n terminal edges, which connect are connected to leaves, and as many as $n-2$ internal edges. The maximum number of edges is achieved when the tree is binary, but the number of edges can be lower if the tree contains any polytomies (vertices with degree greater than 3). With each distinct tree topology, we associate a Euclidean *orthant*, of dimension equal to the number of edges that the topology possesses. (Here, we may regard an orthant to be the subset of Euclidean space with all coordinates non-negative.) For each topology, the orthant coordinates correspond to edge lengths in the tree.

Before continuing, we should note that all tree topologies have the same set of n terminal leaves, and each of these leaves is associated with a single terminal edge. A consequence of these facts is that the portion of the BHV space which corresponds with the leaf edges is equivalent to the Euclidean orthant $\mathbb{R}_{+}^{n}$, and furthermore the entire BHV space is the cartesian product of that Euclidean orthant with a non-Euclidean space representing the internal structure of the tree.

Since the information contained in the terminal edge lengths is, in a sense, orthogonal to the part of the space encoding the topological structure of the trees, we will often simplify our discussion by ignoring the terminal edge lengths, and concern ourselves primarily with the portion of each orthant which describes the internal edges. (Recall that this space has at most $n-2$ dimensions.) Even though we will often make this simplification, we should keep in mind that we could reincorporate the leaf edges without any serious complication.

With this simplification, each of the coordinates in our simplified orthant corresponds to the length of one of the internal edges in the tree. The orthant boundaries (where at least one coordinate is zero) thus represent trees with collapsed internal edges, or in other words, trees with polytomies. These points can be thought of as as corresponding to trees with slightly different (but closely related) topologies. The BHV space is constructed by noting that the boundary trees from two different orthants may describe the same topology. With this insight, we may set about constructing the space by grafting orthant boundaries together when the topologies of the trees they represent coincide.

Figure 1.6 depicts a portion of the BHV space on rooted trees with 4 leaves, which we denote $\mathcal{T}_4$. The depicted portion of the space includes five orthants (topologies) and the structure of the connections between them. Since rooted binary trees on four leaves have two internal nodes, the space consists of two dimensional orthants. Each point within an orthant corresponds to the tree with its associated topology and given internal edge lengths. The origin of each orthant corresponds to the tree with no internal edges (the *star tree*), and the boundary rays correspond to trees where one edge is collapsed, forming a single internal node with three children.

While it is not possible to depict the entire space $\mathcal{T}_4$ in a two dimensional space in the manner of Figure 1.6, it is possible to represent the complete structure of the grafting, as is done in Figure 1.7. In this graph, each edge represents a continuous path through a single orthant, from one boundary ray to the other. Conceptually, this path is formed by exchanging length between the internal edges of the tree. If two edges are joined by a node, then there are trees along the boundaries of the orthants which share a common (polytomic) topology. These boundaries are grafted together, making it possible to form a continuous path between the two orthants.

If two trees are within the same orthant, then we define the distance between them using the Euclidean distance between the corresponding points. However, if the trees have different topologies, then things become more difficult. In Figure 1.8 we have plotted the example trees (by arbitrarily designating node e as the root) onto the portion of the space from Figure 1.6. One possible continuous path between any two trees can be formed by shrinking the internal nodes of one tree down to zero (forming the star tree), and then expanding the tree again in the correct topology. This path is called the *cone path* and is depicted by the dotted line.

However, there is a shorter path connecting the trees, in which only one internal edge is collapsed and the resulting polytomy can be resolved directly into the topology of the other tree. This path is depicted by the solid line. Considering only these two options, it is clear that cone path is longer. However, we have yet to establish that we have, in fact, found the shortest among all possible paths.

Since each orthant is locally a Euclidean space, the shortest path between two points within a a single orthant is a straight line. The difficulty comes in establishing which sequence of orthants joining the two topologies will contain the geodesic. In the case of four leaves, we could do this through a brute-force search, but we cannot hope to do so with larger trees. Owen and Provan [124] present a quartic-time algorithm (in the number of leaves) for finding the geodesic path between any two points in the space. Once the geodesic is known, computing its length—and thus the distance between the trees—is a simple matter.

1.7 Statistical methods for testing congruency between trees

Another fundamental problem within systematics is how to characterize differences between phylogenetic trees. For example, conflicting phylogenies arise when different phylogenetic reconstruction

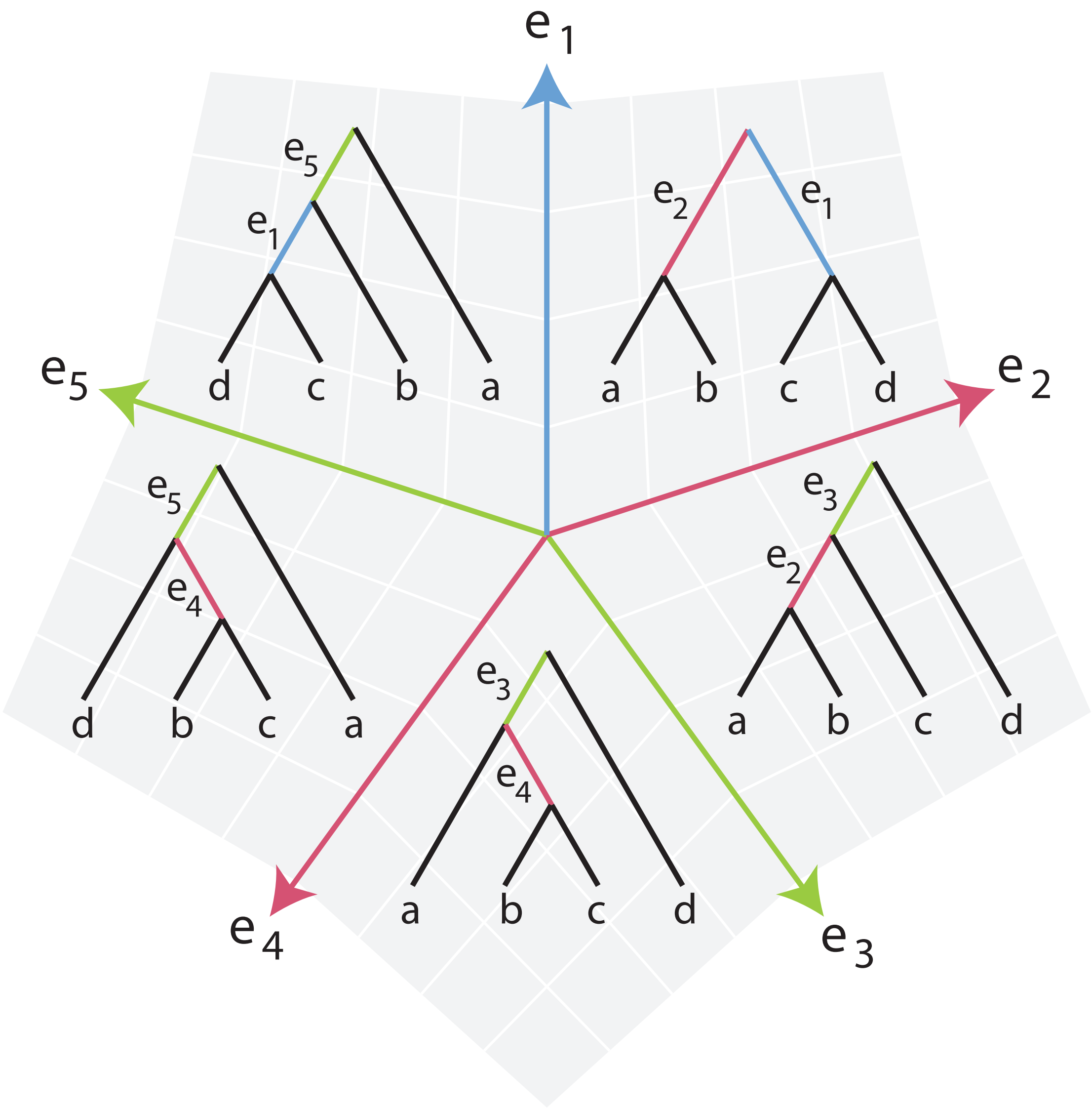


Figure 1.6: A portion of the space of rooted trees with 4 leaves. The space is formed by grafting together orthants, each corresponding to a particular topology. The full space contains several additional orthants, and is depicted schematically in Figure 1.7.

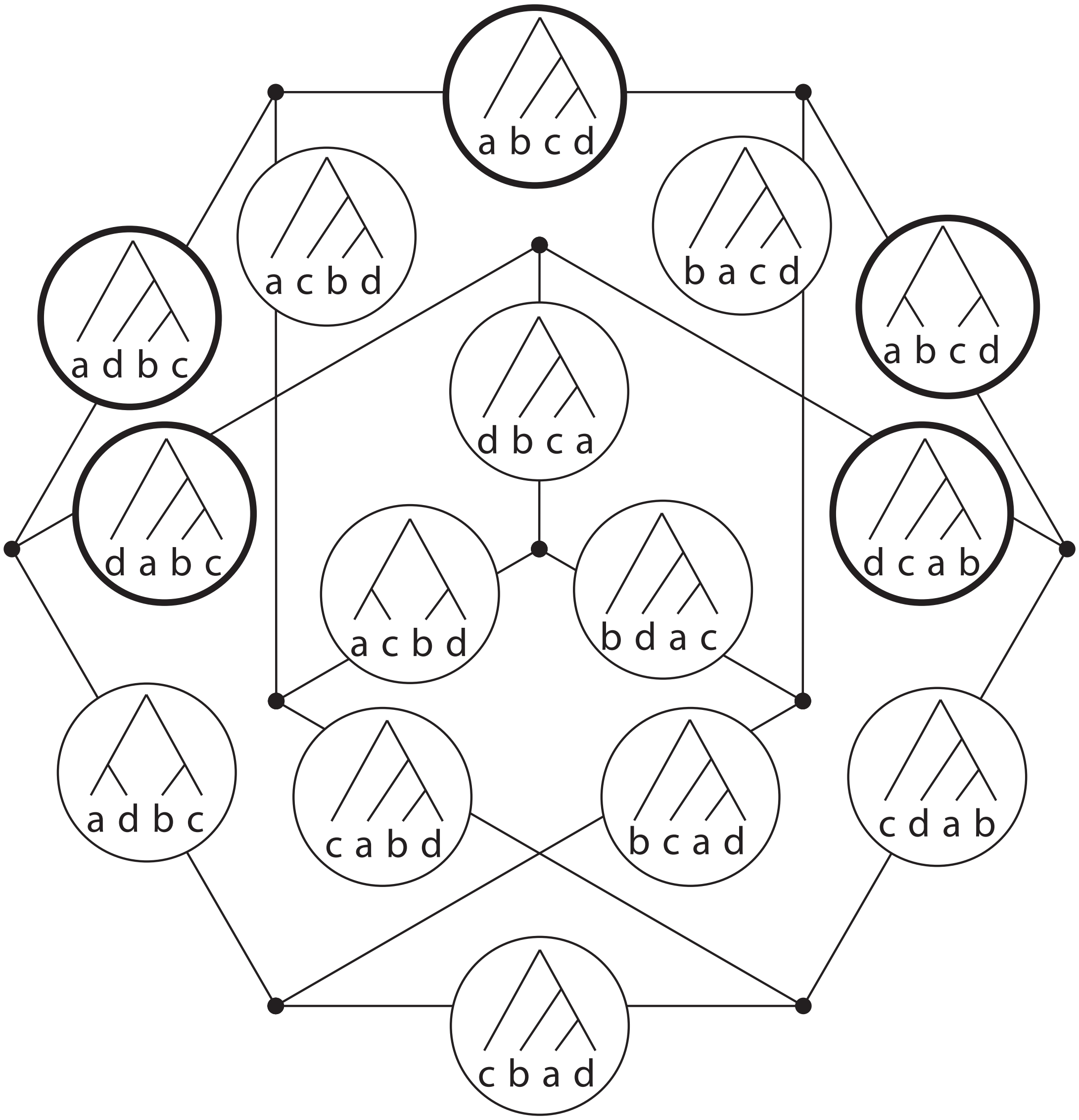


Figure 1.7: A schematic representation of the full space of trees on four leaves, forming a Petersen graph. Each edge represents a rooted binary tree topology, and each node represents the grafting together of orthant boundaries from the connected topologies. The portion of the graph depicted in Figure 1.6 is bolded.

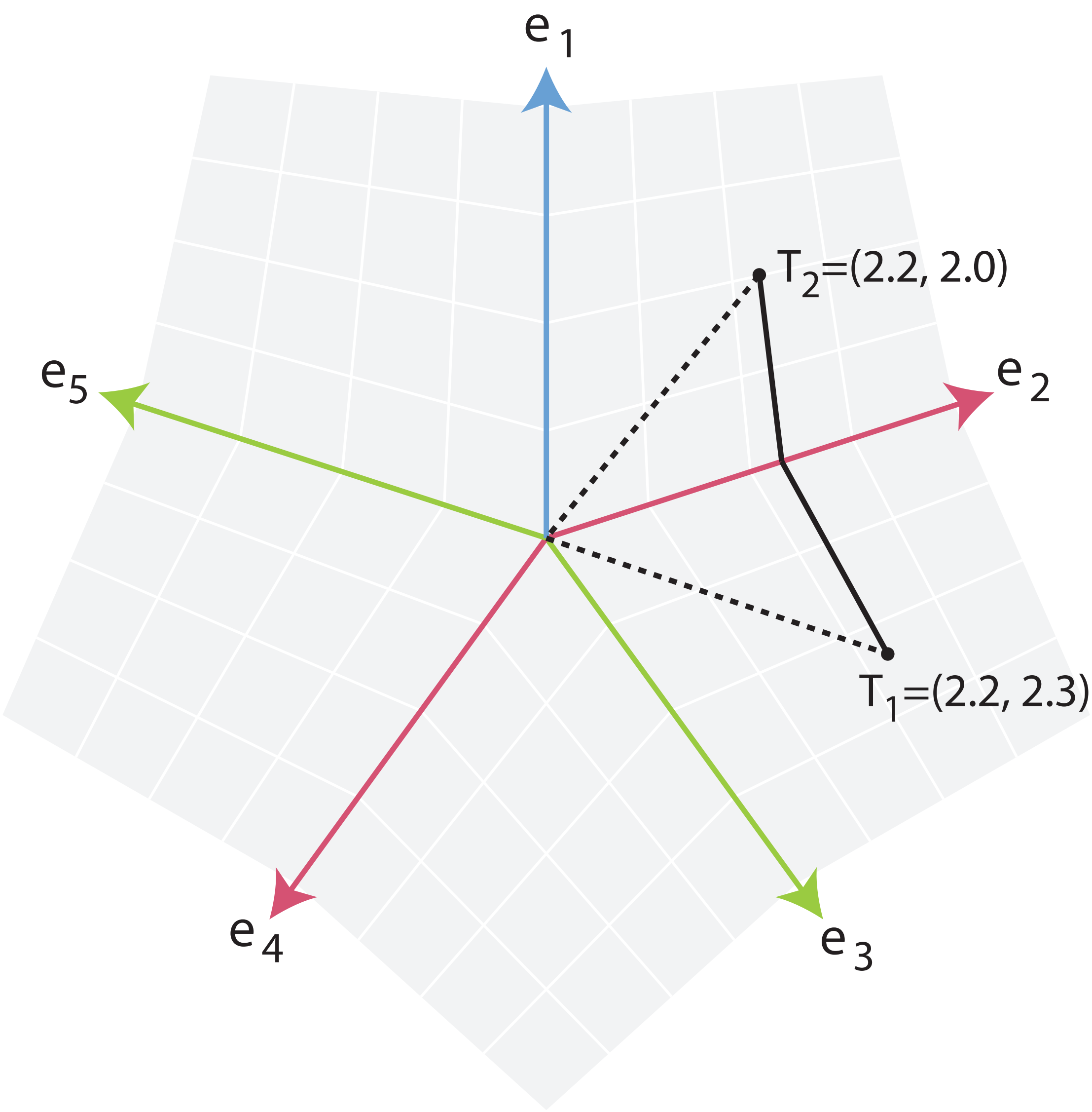


Figure 1.8: If we relabel leaf e as the root, then we can plot our example trees in the reduced tree-space. (The orthants shown correspond to the same topologies as in Figure 1.6.) The cone path between the trees is shown as the dotted line. The shortest path connecting the trees is called a geodesic, and is plotted as the solid black path.

methods are applied to the same data set, or even with one reconstruction method applied to multiple different genes. Gene trees may be considered to be codivergent by virtue of exact congruence or simply because they are insignificantly incongruent. Conversely, one may conclude that a set of trees displays a significant incongruence [110]. All of these outcomes are fundamentally interesting. Congruence of gene trees (or subtrees) is often considered the most desirable outcome of phylogenetic analysis, because such a result indicates that all sequences in the clade are orthologs (homologs derived from the same ancestral sequence without a history of gene duplication or lateral transfer), and that discrete monophyletic clades can be unambiguously identified. In contrast, gene trees that are incongruent are often considered problematic because the precise resolution of speciation events is apparently unclear. This suggests that the ability to identify significant incongruence within sets of gene trees would be useful and interesting, since these events represent non-canonical evolutionary processes [111, 46, 107, 106].

The most common pattern of gene trees in genome evolution is one of codivergence, the parallel divergence of ecologically associated lineages [128, 130, 162, 32, 93]. However, deviations from codivergence are known to exist and can include gene duplications; lateral interspecific gene transfers; retention of ancestral sequence polymorphisms through speciation events through the action of balancing selection; loss of a particular gene within some populations; or accelerated evolution by neofunctionalization, i.e., the gain of novel gene function through sequence divergence by a duplicate copy of a progenitor gene. These six commonly recognized types of evolutionary events all represent different biologically interesting phenomena which may be interesting to study in a variety of contexts [126].

In addition to these six scenarios, there is another phenomenon known to systematists which tends to cause incongruence within a set of trees known as long-branch attraction (LBA). LBA is an erroneous grouping of two or more long branches as sister groups due to methodological artifacts [11]. It was shown in Page and Holmes [131] that the problem of LBA may be severe in the case of trees for four sequences but it may not be as problematic in the case of trees with larger numbers of taxa.

However, recently Fares et al. [49] showed that positive selection for increased expression and consequent rapid cell growth after a whole-genome duplication event, subsequent rearrangements, and later gene loss might be the cause of LBA artifacts in phylogenetic trees discussed in the yeast literature. These artifacts could be the cause of conflicting topologies among neighbor-joining (NJ) trees reconstructed from alignments at different loci. In addition, Kück et al. [95] showed that the LBA affects the maximum likelihood estimation of a phylogenetic tree, even in the event the correct model is used. Differences in selection pressure across the genome has also been suggested as a cause for gene tree incongruence [145, 30].

Deviations from strict phylogenetic codivergence of genes within a genome generally elicit considerable interest in the scientific community and even in the public. The notion of lateral (or horizontal) gene transfer (LGT) is an excellent example. Claims of LGT between very distantly related organisms (e.g., taxonomic domains or kingdoms) regularly appear in high profile publications, and the possibility also underlies many public concerns about genetically modified organisms [86], or the potential emergence of new pathogenic "superbugs" that defy many types of antibiotics [120]. Given these important concerns for medicine, agriculture and the environment, additional methodology which assesses possible cases of evolutionary processes such as LGT are sorely needed.

Much of the evidence for LGT so far has been controversial, especially for eukaryotes [150, 160]. Such claims generally involve first an assessment to check whether the sequence in question is grouped within a clade dominated by homologs from another kingdom. Subsequent statistical tests typically compare alternative tree topologies [141, 89], but the underlying assumption remains untested; that is, it remains unclear whether the tree is actually indicative of LGT rather than some other evolutionary process that would also cause deviation from the species tree. For example, paralogy and gene loss are not tested statistically even though such processes are consistently evident in gene trees touted as evidence of LGT. In prokaryotes, LGT of plasmids and mobile genomic islands (e.g., pathogenicity islands) evident from phylogenetics have been experimentally substantiated [105], but even in these cases a variety of other evolutionary processes undoubtedly operate on these elements. Thus, it would be highly beneficial to move beyond the identification of disparities between phylogenies of genes and genomes or hosts and parasites, and to elucidate the most likely causes of those disparities by means of explicit statistical tests for different processes underlying gene/genome coevolution and host/parasite coevolution.

While there has been a well-established understanding of the discordant phylogenetic relationships that can exist among independent gene trees drawn from a common species tree [133, 166, 110, 15], phylogenetic studies have only recently begun to shift away from single gene or concatenated gene estimates of phylogeny towards these multi-locus approaches [28, 181, 12, 68, 169]. These newer approaches focus on the effect of genetic drift in producing patterns of incomplete lineage sorting and gene tree/species tree discordance, largely using coalescent theory [146, 147, 38, 108, 91, 180, 170]. These theoretical developments have been used to reconstruct species trees from distributions of estimated gene trees [111, 29, 46, 115, 148, 92, 179, 2, 75].

In statistics, one possible relationship between gene and species trees is well-understood in terms of the coalescent processes [87, 67]. However coalescent models usually assume that genes cannot be transferred between members of different species. Just as host switching can cause parasite trees to disagree with host trees, LGT can cause gene trees to disagree with species trees. Combinatorially, these mechanisms correspond to *subtree prune and regraft* (SPR) operations [155]. Many techniques have been developed to compare gene trees [106, 46, 5, 57, 172, 88, 171, 100], and host and parasite trees [42, 153, 76, 61].

The increased use of multi-locus data sets for phylogenetic reconstruction has increased the need to determine whether a set of gene trees significantly deviates from the phylogenetic patterns of other genes. Motivated by this problem, there has been significant work devoted to the development of statistical methods for testing hypotheses of discordance between the trees in a collection. For example, the Bayesian estimation methods [106, 46, 5], the Templeton test implemented in `paup*` [165, 57], the partition-homogeneity test (PHT) also implemented with `paup*` [172], Kishino-Hasegawa test [88], and the likelihood ratio test (LRT) [171] are statistical methods to see if there is a "significant" level of incongruence between the trees (these methods are also called partition likelihood support (PLS) [100]).

On the other hand, the methods which are used in the host-parasite analysis aim to test whether there is a "significant" level of congruence between the trees. Since Henning [69] (see also a nice summary of works in Dowling et al. [42] and the references within), there have been many studies analyzing host-parasite cospeciation. For example, the LRT of Huelsenbeck et al. [77], applying the Markov chain Monte Carlo (MCMC) techniques for estimating lateral transfers as in Huelsenbeck et al. [76], methods that compare trees' pairwise distance matrices, (e.g., by the Mantel test [61], ParaFit [101], and [152]), Brooks parsimony analysis (PSA) [20, 21, 22, 24, 23], and PSA [41], implemented in the software `TreeMap` [127, 129], are statistical methods to test if there is codivergence between trees.

1.8 Organization of this dissertation

The remainder of this dissertation is a series of manuscripts which attempt to address some of these problems by developing novel methods for the analysis of sets of phylogenetic trees. Chapter 2 develops a technique applicable to the BHV tree space which allows us to quickly screen a set of phylogenetic trees for potential outliers. This method is modeled on the kernel density techniques of nonparametric statistics. Chapter 3 presents a refinement of this method, where we address an issue arising in BHV space which does not occur in the Euclidean spaces where kernel methods are typically carried out: the volume of a kernel in BHV space depends not only on the bandwidth of the kernel, but also on the location of the kernel within the space. Chapter 4 presents a wholly different approach to the problem of analyzing sets of trees which is modeled on the ideas of principal component analysis. Finally, Chapter 5 presents a few possibilities for further projects based on the techniques already presented. A glossary of symbols used in this manuscript may be found in an appendix.

Chapter 2

kdetrees: Nonparametric Density Estimation for Phylogenetic Trees

This chapter consists of a manuscript that has been published as Weyenberg et al. [174].

Abstract

While the majority of gene histories found in a clade of organisms are expected to be generated by a common process (e.g. the coalescent process), it is well-known that numerous other coexisting processes (e.g. horizontal gene transfers, gene duplication and subsequent neofunctionalization) will cause some genes to exhibit a history quite distinct from those of the majority of genes. Such "outlying" gene trees are considered to be biologically interesting and identifying these genes has become an important problem in phylogenetics.

We propose and implement KDETREES, a nonparametric method of estimating distributions of phylogenetic trees, with the goal of identifying trees which are significantly different from the rest of the trees in the sample. Our method compares favorably with a similar recently-published method, featuring an improvement of one polynomial order of computational complexity (to quadratic in the number of trees analyzed), with simulation studies suggesting only a small penalty to classification accuracy. Application of KDETREES to a set of Apicomplexa genes identified several unreliable sequence alignments which had escaped previous detection, as well as a gene independently reported as a possible case of horizontal gene transfer. We also analyze a set of *Epichloë* genes, fungi symbiotic with grasses, successfully identifying a contrived instance of paralogy.

Our method for estimating tree distributions and identifying outlying trees is implemented as the R package KDETREES, and is available for download from CRAN.

2.1 Introduction

A central problem in systematic biology is the reconstruction of the evolutionary history of populations and species from numerous gene trees with varying levels of discordance [19, 45]. Although there is a well-established understanding that discordant phylogenetic relationships will exist among independent gene trees drawn from a common species tree [133, 166, 110], phylogenetic studies have only recently begun to shift away from single-gene and concatenated-gene estimates of phylogeny in favor of multi-locus methods [28]. These newer approaches focus on the role of genetic drift in producing patterns of incomplete lineage sorting and gene tree/species tree discordance, largely using coalescent theory [146, 147, 38]. These theoretical developments have been used to reconstruct species trees from samples of estimated gene trees [111, 29, 46, 116, 148].

Detecting concordance among gene trees is also a topic of interest. For example, Ané et al. [6] developed a Bayesian method to estimate concordance among gene trees using molecular sequence data from multiple loci. The method can produce estimated gene trees as well as an estimate of the proportion of the genome that support a particular clade. However, *a priori* assumptions must be made about the degree and structure of concordance present in the gene trees.

Although there is a tremendous amount of ongoing effort to develop better parametric models for gene tree distributions, the parametric framework has inherent limitations. While a parametric method typically makes the most efficient use of a given data set when the model is specified correctly, they achieve this efficiency by assuming that the true distribution of gene trees is one of a relatively small class of distributions. This can lead to erroneous inferences when the the true distribution does not resemble any of the models in the proposed class. Given that many questions remain about the proper way to incorporate a number of important processes into a parametric model (e.g. geographic barriers to migration, or a population bottleneck), the problem of model mis-specification is very real. Nonparametric methods avoid the majority of these modeling issues, enabling unbiased estimation for a much larger class of true tree distributions at a cost of statistical efficiency.

Numerous processes can reduce the correlation among gene trees. Negative or balancing selection on a particular locus is expected to increase the probability that ancestral gene copies are maintained through speciation events [167]. Horizontal transfer introduces divergent gene copies into a different species through a vector, or shuffles gene copies among species via hybridization [110]. The correlation may also be reduced by naive sampling of loci for analysis. For example, paralogous gene copies will result in a gene tree that conflates gene duplication with speciation. Similarly, sampled sequence data that span one or more recombination events will yield "gene trees" that are hybrids of two or more genealogical histories [137]. These non-coalescent processes can strongly influence phylogenetic inference [137, 112, 45]. In addition, Rivera et al. [142] showed that an analysis of complete genomes indicated a massive prokaryotic gene transfer (or transfers) preceding the formation of the eukaryotic cell, arguing that there is significant genomic evidence for more than one distinct class of genes. These examples suggest that the distribution of eukaryotic gene trees may be more accurately modeled as a mixture of a number of more fundamental distributions.

In this paper, we focus on the problem of identifying significant *discordance* among gene trees, as well as estimating the distribution of gene trees as a whole. This set of gene trees is assumed to consist mostly of "typical" (or "non-outlier") gene trees, which are assumed to be independently sampled from some distribution f. For example, gene trees have evolved neutrally under a coalescent process. In addition, there are a smaller number of "outlier" gene trees which are sampled from a very different distribution f'. These genes are assumed to arise from less common evolutionary processes; for example, paralogy, neofunctionalization, horizontal gene transfer, or periods of rapid molecular evolution. In addition, more mundane errors—such as incorrect sequencing, alignment, tree reconstruction, or annotation—can also produce outlier trees in a data set [74]. Our method produces a *nonparametric* estimate of the distribution f and also attempts to identify potential outlier gene trees which are probably not generated by f. Trees identified as outliers can then be inspected more closely for biologically interesting properties. In particular, identifying and removing outliers that violate model assumptions can improve the accuracy of inferences made from a collection of gene trees (e.g. Disotell and Raaum [40], Martin and Burg [112], Edwards [45], Posada and Crandall [137]).

2.1.1 Related Work

The method presented in this paper is not, at its present state of development, a statistical method for hypothesis testing, but rather for discovering possible outliers present in a given collection of orthologous genes. However, there has been significant work devoted to the development of statistical methods for testing hypotheses of discordance between the trees in a collection. The reviewed methods in Poptsova [136] are the following: (i) likelihood-based tests of tree topologies, such as the Kishino-Hasegawa [88], Shimodaira-Hasegawa [158] test, and Approximately Unbiased [157] tests; (ii) tree distance methods, such as Robinson and Foulds [143] and subtree pruning and regrafting distances [58]; and (iii) genome spectral approaches, such as bipartition [109] and quartet decomposition analyses [135].

The likelihood-based tests of tree topologies and tree distance methods are statistical hypothesis tests that detect significant incongruence between trees, i.e., they are testing the following hypotheses:

H_0: Given trees are topologically congruent.
H_1: Given trees are topologically incongruent.

The distinction between likelihood and distance based methods is in how they calculate the p-value of these hypotheses. The likelihood-based tests compare each gene tree with a species/reference tree using a likelihood value, to see if the incongruence is "statistically significant." These methods are also known as partition likelihood support (PLS) [100]. Tree distance methods estimate the p-value of the hypotheses above by computing a distance between a reference tree and each gene tree. Holmes [73] describes a framework for statistical hypothesis testing on trees based on tree distances using distributions of phylogenetic trees (e.g. a posterior distribution or bootstrap resampling). Holmes also presents a statistical method to compare two sets of bootstrap sampling distributions, using the mean and variance of each distribution [73, Section 4.4.1]. A nonparametric method for detecting significant discordance between two sets of trees via supporting vector machines (SVMs) was introduced by Haws et al. [63]. This is a nonparametric method for statistical testing of the hypotheses:

$$\begin{aligned} H_0 &: \text{ Two sets of trees are drawn from the same distribution.} \\ H_1 &: \text{ Two sets of trees are not drawn from the same distribution.} \end{aligned}$$

While likelihood-based tests assume that the species tree is known, genome spectral approaches do not use such a reference tree. Genome spectral methods summarize a set of gene trees with phylogenetic spectra (frequencies), such as splits or quartets. These frequencies can be used to approximate the distribution of gene trees, instead of producing a summarizing tree. Outlier trees can be identified by looking for trees whose highly supported features disagree with prevalent features in the spectra [118].

A non-statistical approach for summarizing collections of gene trees is presented by Nye [121]. Treating each gene tree as a leaf node, a "meta-tree" is constructed where nodes correspond to phylogenetic trees; distances between nodes of the meta-tree correspond to distances between phylogenetic trees, and internal nodes correspond to gene trees with various branches collapsed. When using the Robinson–Foulds distance, the nonparametric method proposed in this paper can be viewed as a numerical summarization of the meta-tree in [121].

Recently, de Vienne et al. [37] developed a statistical nonparametric method to detect outlier trees from the set of gene trees. They first convert gene trees into vectors in a multi-dimensional Euclidean space and then apply Multiple Co-Inertia Analysis—an extension of Principal Coordinate Analysis (PCO)—directly to these vectorized gene trees. Their method, Phylo-MCOA, also detects outlier species, those whose position varies widely from tree to tree. Included in our results are simulation studies comparing our nonparametric method with Phylo-MCOA.

2.2 Methods

2.2.1 Algorithm

Let $\mathcal{T}_n$ denote the set of all tree topologies (including multifurcating trees) on n taxa (which we call *tree space*). We consider trees to be unrooted, but rooted trees can be treated similarly. Our main object of study is a sample, $\{T_i\}_{i=1}^N$, of N trees (gene trees) mostly drawn from a distribution f on $\mathcal{T}_n$. If n is large enough that $|\mathcal{T}_n| \gg N$ then many tree topologies in the sample may have low empirical frequency. In this case, f cannot be estimated well by assigning $\hat{f}(T)$ to be the empirical frequency of T in the sample. On the other hand, if f corresponds to a model such as the coalescent, it is reasonable to expect that topologies "close" to many observed trees will have a higher likelihood than topologies "far away" from the observed trees.

Kernel density estimation is a nonparametric technique to estimate a distribution that generated a sample, by leveraging the fact that points close to sample points tend to have higher likelihood than distant outlier points (under adequate assumptions on the distribution, namely, the distribution is square-integrable [114]). Kernel density estimation can be viewed as a refined version of histogram-based estimation of a density. As the term *density* suggests, kernel density estimation is typically formulated for continuous variables over $\mathbb{R}^d$. However, similar methods can also be devised to estimate distributions over a finite set such as tree space.

A key ingredient is the ability to measure similarity between trees. Fortunately, research in phylogenetics has produced several classical distances on tree space, such as the dissimilarity map distance [25], the topological dissimilarity distance measure [161], the Robinson–Foulds distance [143], and the quartet distance [48]. More recently Billera et al. [13] introduced the notion of geodesic distances. [124] showed that there is an efficient algorithm for computing this distance in $O(n^3)$, where n is the number of taxa.

Our method uses existing tree distances to estimate a tree distribution by mimicking kernel density estimation. Our main goal is to identify regions of $\mathcal{T}_n$ which have high probability, as well as observed trees with markedly low estimated probability. These low-probability trees are potentially outlier trees; i.e., trees having evolutionary histories unlikely to have arisen from the same model that generated the non-outlier trees. Our approach is nonparametric, which makes it quite general, and avoids problematic issues such as model design and selection that one encounters when using a parametric model (such as the coalescent). Unfortunately, using a small sample to learn an arbitrary distribution on tree space is inherently difficult, especially as the dimension of $\mathcal{T}_n$ grows, and we do not expect to learn the tree distribution with high accuracy for every tree topology. However, estimates of the density in regions where the probability is high can be quite good.

Our method identifies potential outliers in a set of trees by comparing the values of the non-normalized density estimates (which we call "tree scores") of the trees. An unusually low score indicates that a tree is relatively distant from the other trees in the sample. We implement a simple classification scheme which is based on the interquartile range (IQR) of the density estimates, as is commonly done when creating box-and-whisker plots.

Given an independent and identically distributed sample of trees $T_1, \ldots, T_N$, we propose a non-parametric estimator of the distribution that generated the sample with the form

$$\hat{f}(T) \propto \frac{1}{N} \sum_{i=1}^{N} k(T, T_i).$$

Here $k(\cdot)$, the kernel function, is a non-negative function defined on pairs of trees which measures how "similar" two trees are. For our approach, we do not require $k(\cdot)$ to be a kernel in a strict statistical sense.

In KDETREES we have implemented a kernel of the form

$$k(T, T_i) \propto \frac{1}{h_i} \exp\left(-\left(\frac{d(T, T_i)}{h_i} \right)^{\delta} \right).$$

A distance function on the space of trees, $d(T, T')$, is used to define a univariate projection $\mathcal{T}_n \to \mathbb{R}_+$ in the natural way for each fixed $T \in \mathcal{T}_n$, mapping $T' \mapsto d(T, T')$. The "shape" parameter $\delta > 0$, and the "bandwidth" parameters $h_i > 0$ control how tightly each contribution $k(T, T_i)$ will be centered on T_i. Allowing the bandwidth to vary with the sample points, T_i, is called an *adaptive bandwidth* method. Alternatively the bandwidth can be set to a constant value for all T_i.

In general, we can remove the symmetry and triangle inequality requirements for d, and it is possible that the sum over tree space, $\sum_{T \in \mathcal{T}} k(T, T')$, will vary with T'. Ideally, we would remedy this issue by normalizing $k(\cdot, T')$ so that $\sum_{T \in \mathcal{T}} k(T, T') = 1$. (This is the case most analogous to kernel density estimation.) However, for the d implemented by KDETREES, Monte Carlo estimates of this sum do not appear to vary significantly across T', and so the current version of the software assumes that it is constant. (Additional information about these estimates is presented in Figure B.1.)

Since the ultimate goal is to detect outlier trees, T_j, which are not actually drawn from the true distribution f, we are most concerned with estimating the density at the observed sample points. In this context, it makes sense to use a "leave-one-out" estimator which excludes the contribution of the point in question from the tree score,

$$\hat{g}(T_j) = \frac{1}{N-1} \sum_{i \neq j} k(T_j, T_i).$$

This estimator simply transforms probability estimates via $A(x) = N(x - c)/(N - 1)$ for some c. Assuming the sample is drawn i.i.d. from a distribution f, for fixed d and δ, both $\hat{g}(T)$ and $\hat{f}(T)$ (once normalized) will converge to f as $N \to \infty$, so long as the $h_i(N) \to 0$. This result follows immediately from the finiteness of tree space.

Once we have computed the scores, $\{\hat{g}(T_i)\}$, we classify tree T_j as an outlier if $\hat{g}(T_j)$ is less than $Q_1 - \kappa \cdot IQR$. Where Q_1 and IQR are the first quartile and the interquartile range of the set of tree scores, respectively; and κ is a classification tuning parameter. The choice of κ affects the sensitivity and specificity of the classifier, and is set to 1.5 by default, although the user may supply their own value.

This idea can be extended to also exclude the contribution of a number of trees which are determined to be outliers. Since the magnitude of $\hat{g}(T_j)$ can be used as a measure of evidence for T_j being an outlier, KDETREES can iteratively remove from the calculation the contribution of the tree which minimizes $\hat{g}(T_j)$, and recompute the estimator $\hat{g}$ with the reduced sample. This process can be iterated to remove multiple putative outliers.

Choice of tree distance.— In our approach, trees can be incorporated into a statistical framework by converting them into a numerical vector format based on a distance matrix or map, and several such vectorization schemes were introduced in Chapter 1. These vectorized trees can then be analyzed as points in a multi-dimensional space where the distance between trees increases as they become more dissimilar [72, 156, 60].

For the choice of d, we propose distances derived from three different distances on trees: *dissimilarity map* d_d, *topological dissimilarity map* d_t, and *geodesic distance* d_{geo}. The dissimilarity map distance measure between two trees is the Euclidean distance,

$$d_d(T', T) = ||v_d(T) - v_d(T')||_2,$$

where $v_d(T)$ is a vectorization of trees, $\mathcal{T}_n \to \mathbb{R}^{\binom{n}{2}}$, based on an enumeration of the pairwise distances between the tips [25]. The topological dissimilarity map distance measure between two trees is defined similarly,

$$d_t(T', T) = ||v_t(T) - v_t(T')||_2,$$

but uses a vectorization $v_t(T)$ that counts the number of edges between the tips [161]. An example calculation of both v_d and v_t is shown in Figure B.2.

Billera et al. [13] showed that the space of rooted trees with a fixed number of taxa is the union of positive cones in $\mathbb{R}^{\binom{n}{2}}$. Thus, the space of trees is the set of all metrics derived from valid trees, and is a subspace of the space of all distance matrices. The geodesic distance d_g is the shortest distance between two valid trees when the connecting path is constrained within this tree space (note that this subspace of valid trees is not itself Euclidean). Owen and Provan [124] developed an $O(n^4)$ algorithm to compute the geodesic distance $d_g(T, T')$ between any two valid trees.

Missing taxa.— It is desirable for phylogenetic analyses to be able to deal with situations with incomplete data. In this case, the most relevant type of missing data is when some gene trees are missing a tip which is present in other trees in the data set. Our method is capable of handling such a situation if the dissimilarity or topological distance maps are used. In this situation we impute missing tip-to-tip distances in the tree vectors with the median value found in trees containing the missing tip. Unfortunately, the geodesic distance algorithm we employed does not currently allow us to perform such an imputation, and so KDETREES cannot handle missing tips if the geodesic distance map is selected.

If the trees have node labels which correspond to support for the given split (obtained, for example, by a bootstrap analysis), then the software can accommodate this information by collapsing nodes with support less than a given value. This behavior is disabled by default.

Kernel bandwidth.— The estimator $\hat{g}$ depends crucially on the choice of the bandwidth parameter h. We employ a nearest-neighbor approach to estimate an adaptive bandwidth for each sample point. To estimate the bandwidth for a point T_j, we use the distance to the m-th closest sample point. This approach has the effect of causing the kernels to be concentrated in areas where there is a lot of data, and diffuse in the tails of the distribution. In the current version of KDETREES m is defaulted to be 20% of the sample size, a heuristic value chosen based on simulation results.

Alternatively, the bandwidth can be set to a constant value for all T_i. In order to do this we must find a way to choose an optimal value for the bandwidth h. We experimented with a constant bandwidth chosen by estimating the partition function $Z_h = \sum_T \hat{g}_h(T)$ using a random sample of trees. However, it seems that we tend to under-estimate the bandwidth h and the results are not as robust as in the case of the adaptive bandwidth.

```
$S$ is a set of sets of gene trees.
$g$ is the number of non-outlier trees in each simulation.
$r$ is the number of outlier trees.
$\kappa$ is the classifier tuning parameter.
for all iterations in simulation do
    Generate non-outlier trees (Sample $g/|S|$ coalescent trees from each $s \in S$.)
    Generate $r$ random outlier gene trees. (Each within a new random species tree.)
    Analyze data with both kdetrees and Phylo-MCOA
    Tally true and false outlier identifications for each method
end for
```

Figure 2.1: Summary of the simulation comparing kdetrees and Phylo-MCOA. (See Figure B.3 for a plot of the species tree used.) For the "single" simulations, S contains a single tree (top left of Figure B.3), while for the "mixed" simulations it contained 5 trees (remainder of Figure B.3). For our simulations, $r = 1$ and $g = 100$.

Tuning parameters.— The outlier classifier's sensitivity depends on the choice of a tuning parameter, κ. The default value, 1.5, is chosen for historical reasons. In our simulations smaller values of κ, around 0.75 to 1, often resulted in false positive rates close to 5%. Creating plots of the tree scores may be helpful in choosing an appropriate value for a given data set.

Computational complexity.— The running time of kdetrees is dominated by the step where pairwise tree distances are calculated. For N trees, each with n taxa, this step takes $O(n^2N^2)$ operations when using the dissimilarity or topological distances, or $O(n^4N^2)$ if using the geodesic distance.

2.2.2 Simulations

We conducted a series of simulations comparing the performance of kdetrees and Phylo-MCOA. (Code and documentation for the simulations is included in a package vignette with kdetrees.) (Phylo-MCOA is a `R` package and one of the functions in the software is to identify putative outlying genes in a data set.) The simulated data consisted of coalescent trees generated by the Python library DendroPy [164]. Six species trees (see Figure B.3) were used to contain coalescent gene trees. A data set consisted of a small number of "outlier" gene trees, together with a larger number of "non-outlier" gene trees. In the "single" coalescent simulations, the non-outlier trees are all contained within the top left tree in Figure B.3. In the "mixed" coalescent simulations, an equal number of non-outlier genes were sampled from each of the other 5 trees. Pseudocode in Algorithm 2.1 summarizes the simulation processes.

Our first simulation investigated the classification characteristics of the methods, producing receiver operating characteristic (ROC) curves comparing kdetrees and Phylo-MCOA, by varying the classification tuning parameter of each method. (A ROC curve is a graphical plot of the fraction of true positive rate vs. the fraction of false positive rate at various threshold settings [62].) In this simulation we set the effective population size of the coalescent process generating the trees to 2000, a value which produced a moderate amount of variance in the generated coalescent trees.

A second simulation compares the true positive rates of the methods as the variance of the coalescent trees increases. (Variance of the random trees is controlled by the coalescent population parameter.) This simulation was carried out both with the default classification tuning values, as well as values chosen based on the ROC simulation results to limit the false positive rate (FPR) to around 5%.

A third simulation compared the distribution of outlier tree scores to the distribution of non-outlier tree scores. The simulation process is summarized in the pseudocode in Algorithm 2.2.

```
Generate $g$ coalescent trees within a fixed species tree.
Use kdetrees to obtain scores for non-outlier trees.
for all iterations in simulation do
    Generate a single outlier tree within a new species tree.
    Append outlier tree to set of non-outlier trees.
    Obtain and record outlier tree score.
end for
Plot kernel density estimates of both score distributions.
```

Figure 2.2: Summary of the simulation design for the simulation comparing the tree score distributions for outlier trees and non-outlier trees. For our simulations both g and R are set to 500, and the coalescent parameter is 2000.

2.2.3 Biological datasets

Apicomplexa.— The Apicomplexa data set presented by Kuo et al. [96] consists of trees reconstructed from 268 single-copy genes from the following species: *Babesia bovis* (Bb) [17], *Cryptosporidium parvum* (Cp) [1] from CryptoDB.org [66], *Eimeria tenella* (Et) from GeneDB.org [70], *Plasmodium falciparum* (Pf) [55] and *Plasmodium vivax* (Pv) from PlasmoDB.org [8], *Theileria annulata* (Ta) [132] from GeneDB.org [70], and *Toxoplasma gondii* (Tg) from Toxo-DB.org [54]. A free-living ciliate, *Tetrahymena thermophila* (Tt) [47], was used as the outgroup. To this set of sequences, we appended the Set8 gene, which has been identified by Kishore et al. [90] as a probable case of horizontal gene transfer from a higher eukaryote to an ancestor of the Apicomplexa.

Epichloë.— Another set of biological sequences to use as a test case was generated from housekeeping genes and a known pair of paralogs in *Epichloë* species and related plant symbionts and parasites in the fungal family Clavicipitaceae. We previously reported sequencing, annotation, and the identification of orthologs in genome of *Epichloë amarillans* strain E57, *E. brachyelytri* E4804, *E. festucae* strains E2368 and Fl1, *E. glyceriae* E277, *E. poae* E5819, *E. typhina* E8, *Aciculosporium take* MAFF-241224, *Claviceps fusiformis* PRL 1980, *C. paspali* RRC-1481, *C. purpurea* 20.1, *Neotyphodium gansuense* e7080, and *Periglandula ipomoeae* IasaF13 [154]. We compiled the inferred protein sequences for ten housekeeping proteins, namely, γ-actin (ActG), DNA lyase (ApnB), a calmodulin-dependent protein kinase (CpkA), the largest and second largest subunits of RNA polymerase II (rpbA and rpbB), translation elongation factor 1-α (TefA), α-tubulin (paralogs TubB and TubC), and β-tubulin (paralogs TubB and TubP). As the expected phylogenetic outlier, we compiled two known paralogous proteins, namely, LolC (which catalyzes synthesis of a loline alkaloid intermediate), and the very closely related *O*-acetylhomoserine(thiol)-lyase (CysD, which scavenges H_2S for synthesis of a methionine intermediate) [159]. Of the 13 fungal strains, three had *lolC* genes but not *cysD*, nine had *cysD* but not *lolC*, and one (*E. glyceriae* E277) had both genes. Both LolC/CysD datasets had one sequence from each strain, but they differed in containing either LolC or CysD from *E. glyceriae.*

2.3 Results

We present the software package kdetrees for nonparametric estimation of tree distributions and detection of outlier trees. The software takes as input a sample of trees in Newick format, and estimates for each tree a “score” based on a nonparametric estimator of the tree density. It can then use these scores to identify putative outlying trees in the sample. The trees scores and summary plots are produced as output.

The kdetrees package is written in R [139], and depends on packages distory [31], ggplot2 [178], and ape [134]. The software is available for download from CRAN and is compatible with all systems supported by R.

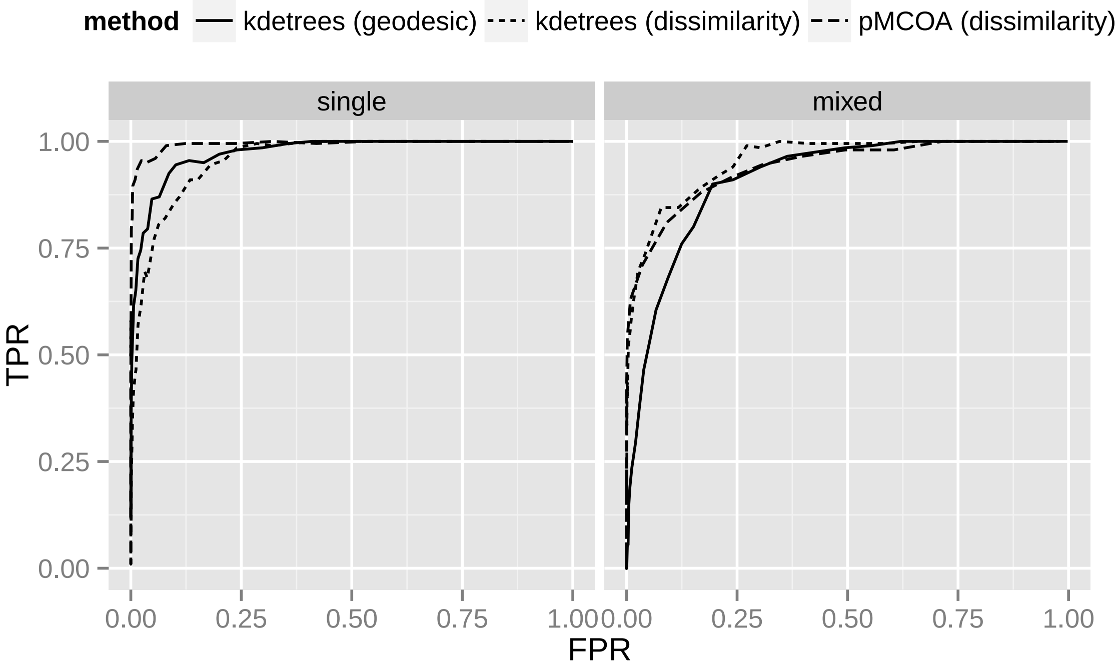


Figure 2.3: ROC curves comparing KDETREES and PHYLO-MCOA as the classification tuning parameter is varied. (In general higher is better, a very effective classifier will pass close to the upper left corner.) The effective population size is 2000 for the coalescent trees. At left are the "single" contained coalescent simulations, with the non-outlier trees all contained within a single species tree. At right are results from a "mixed" simulation, with the non-outlier trees generated from a mixture of 5 species trees.

2.3.1 Simulation Results

Our first simulation, presented in Figure 2.3, produced ROC curves comparing the various methods of outlier identification. We find that the performance of KDETREES and PHYLO-MCOA is similar, with PHYLO-MCOA having a slightly better curve in the single simulations, and KDETREES in the mixed scenarios. Interestingly, the geodesic distance worked better for the "single" data than the dissimilarity map, while the relationship is reversed for the "mixed" simulation. These results were almost completely unaffected by changes in the proportion of outliers in the sample (proportions between 1 to 10% were tested).

The variability of the coalescent trees is determined by the effective population size, the parameter studied in our second simulation. The proportion of the simulated data sets where each method correctly identified an added outlier tree is illustrated in Figure 2.4. This simulation was run both with default tuning parameters and ones chosen based on the ROC curve simulation results. If optimal tuning parameters are selected, PHYLO-MCOA can outperform KDETREES, however, selecting these correctly can be difficult.

We ran a third simulation studying the difference between the score distributions of outlier trees and non-outlier trees, as the ability of our method to reliably detect outlying trees depends on a tendency by outlier trees to produce scores significantly lower than the scores of non-outlier trees. The results are presented in Figure 2.5. We found that while there is some overlap between the score distributions, the distribution of scores for outlier trees lies significantly below that of non-outlier trees.

Finally, Figure 2.6 summarizes the running times of the algorithms as the number of trees in the

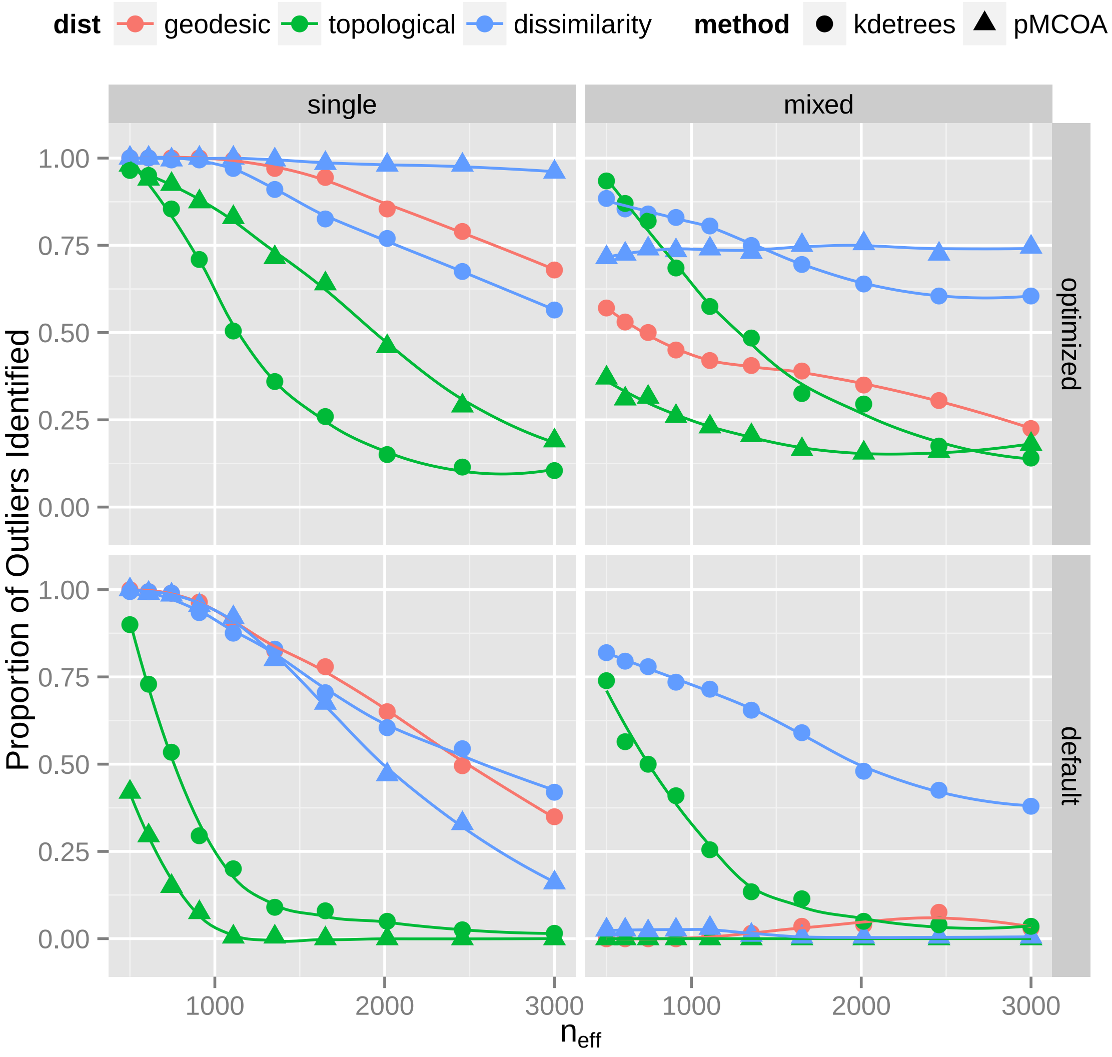


Figure 2.4: Summary of simulation results comparing performance of KDETREES and PHYLO-MCOA for various values of the effective population size. Shown is the proportion of simulated data sets in which the methods identified the outlier tree. The top two plots use use tuning parameters chosen based on results of the ROC simulation, while the bottom plots use default values. For KDETREES the optimal tuning parameter was $\kappa = 0.7$, while for PHYLO-MCOA it was $\kappa = 0.25$. The default values are both $\kappa = 1.5$.

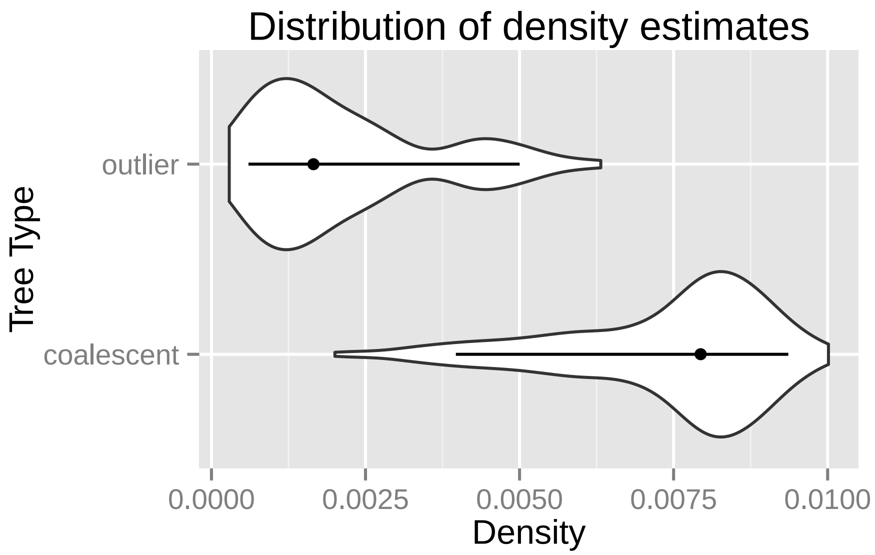


Figure 2.5: Kernel density estimates of the observed distribution of tree scores. The "coalescent" scores are for contained coalescent trees generated within a fixed species tree (bottom). A single random outlier tree is added to this data set and its score computed. This process is replicated to generate the sample of "outlier" tree scores (top). Lines and dots represent the 5%–95% quantiles and the median, respectively. An effective population size of 2000 was used to produce these estimates.

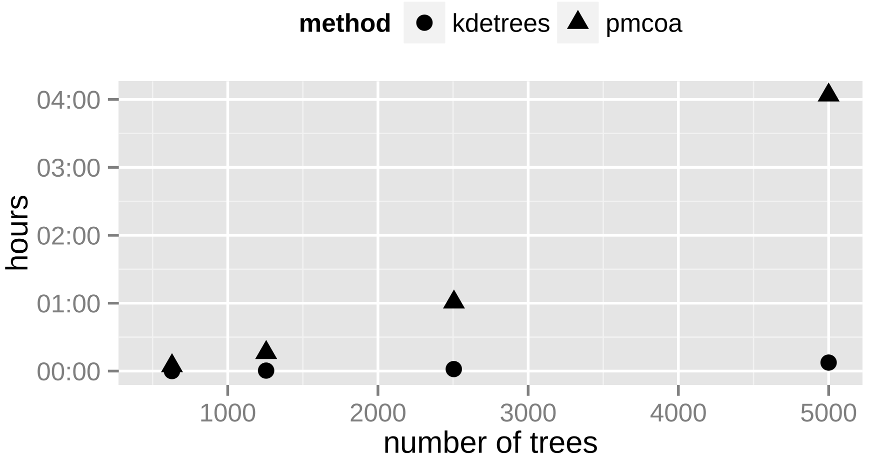


Figure 2.6: The running time (in hours) of KDETREES and PHYLO-MCOA as the number of trees in the data set increases. The trees used here have 50 tips each.

data set is increased. Here KDETREES vastly outperforms PHYLO-MCOA. For a data set consisting of 5000 trees, each with 50 tips, KDETREES completed in about 7.5 minutes, while PHYLO-MCOA required slightly over 4 hours. For smaller data sets, of a few hundred trees, KDETREES runs in less than a second, while PHYLO-MCOA requires a few minutes.

Table 2.1: Apicomplexa gene sets identified as outliers by KDETREES. All annotations except 728 are putative.

No.[a]	GeneID[b]	Functional Annotation
488	PF08_0086	RNA-binding protein
497	PF13_0228	40S ribosomal subunit protein S6
515	PFA0390w	DNA repair exonuclease
546	PFF0285c	DNA repair protein RAD50
547	PFL1345c	Radical SAM protein
641	PFE0750c	hypothetical protein, conserved
660	PF10_0043	ribosomal protein L13
662	PF11_0463	coat protein, gamma subunit
728	MAL13P1.22	DNA ligase 1
747	PFB0550w	Peptide chain release factor subunit 1
773	PFF0120w	geranylgeranyltransferase
780	PFD0420c	flap exonuclease

[a] Based on geneset designations in Kuo et al. [96].
[b] Geneset represented by GeneID for *Plasmodium falciparum*.

2.3.2 Biological data results

Apicomplexa.— The list of putative outlier genes identified by KDETREES in the Apicomplexa data is presented in Table 2.1, with additional discussion in Table B.1. When employing either the dissimilarity maps or geodesic distance, our method identified the same set of putative outlier trees. (The first four trees identified as putative outliers are also plotted in Figures B.4-B.7, and the entire set of estimated scores are summarized in Figure B.8.) These trees all contain a branch with a length that is far too long in proportion to the other branches, leading to their identification as outliers. Closer inspection of these trees suggested that they correspond to questionable sequence alignments which likely non-homologues included due to poor annotation, many involving *Eimeria tenella* (Et) sequences.

Since there appeared to be pervasive problems with the Et sequence data, we removed this species from the data set and recreated the phylogenetic analysis as in Kuo et al. [96]. With the reduced set of gene trees, KDETREES identified a different set of outlier trees, and in this case the Set8 gene was selected as the furthest outlying tree.

Epichloë.— The fungal datasets included alignments with known paralogs, LolC and CysD. Because *E. glyceriae* E277 had both *lolC* and *cysD*, we ran the analysis on alternative data sets with either LolC or CysD eliminated for that strain. In both analyses, the LolC/CysD tree was identified as one of two outlier genes, the other being the DNA lyase protein ApnB. Topologically, the LolC/CysD gene tree differed markedly from the others, which is as expected because CysD sequences grouped together in a clade apart from LolC. However, the topology of the ApnB tree was similar to that of other housekeeping genes, suggesting that it had significantly different relative branch lengths.

2.3.3 Running Time

A significant advantage of KDETREES over PHYLO-MCOA is a significant improvement in computational speed, especially with larger data sets. Actual KDETREES running times are well fitted by a $O(N^2)$ curve, as suggested by the complexity of the algorithm discussed previously, while the PHYLO-MCOA times are $O(N^3)$.

2.4 Discussion

2.4.1 Simulations

The simulation results were generally positive for KDETREES. Although PHYLO-MCOA was often able to slightly outperform KDETREES in classification accuracy, the difference was often relatively small. However, in terms of computational time, KDETREES vastly outperforms PHYLO-MCOA, especially as the number of trees in the data set increases.

In all cases studied, methods incorporating branch length information outperformed the topology only methods. The performance of the geodesic distance was better in the "single" simulations than the "mixed" simulations, although the reason for this is unclear. All of the methods were able to correctly identify the outlier tree when the effective population size (and thus tree variance) was low, provided that a suitable tuning parameter was chosen. As the variance of the coalescent trees increased, the performance of PHYLO-MCOA tended to degrade at a slightly slower rate than KDETREES.

It should be noted that choosing a suitable tuning parameter can be quite difficult, as the optimal value depends on not only the details of the data set, but also one's subjective opinions on the relative merits of the sensitivity and specificity of the classifier. As such, we also studied the behavior of the algorithms when using their default tuning parameters. This information is relevant, since many users will not change the parameters from their default values. With these values we found that KDETREES is slightly superior to PHYLO-MCOA in the single-distribution simulations. In the mixed-distribution simulations the default values for PHYLO-MCOA resulted in very poor performance, while KDETREES's rate of outlier identification was much higher.

The third simulation set compared the distribution of scores for outlier trees to the scores of non-outlier trees. Although the distributions are not completely distinct, it is clear that the outlier trees tend to have scores smaller than the majority of non-outlier trees. Since the outlier trees were generated as completely random coalescent trees, there will inevitably be trees generated which have structure similar to the non-outlier trees, simply by chance, and this accounts for some of the overlap between the distributions. With real data, such trees would correspond to genes which have some exotic history, but nonetheless appear to have a phylogeny substantially similar to the rest of the genes in the genome. In this case, it is ambiguous whether or not such a gene should be legitimately classified as an outlier.

The main advantage of KDETREES over PHYLO-MCOA lies in the vast improvement in running time on data sets with larger numbers of gene trees. For small data sets the difference is not material, however for data sets with several thousand trees, PHYLO-MCOA requires many hours to complete, while KDETREES will finish within a few minutes on contemporary commodity hardware.

2.4.2 Biological datasets

Apicomplexa.— The phylum Apicomplexa contains many important protozoan pathogens [102], including the mosquito-transmitted *Plasmodium* spp., the causative agents of malaria; *T. gondii*, which is one of the most prevalent zoonotic pathogens worldwide; and the water-born pathogen *Cryptosporidium* spp. Several members of the Apicomplexa also cause significant morbidity and mortality in both wildlife and domestic animals. Due to their medical and veterinary importance, whole genome sequencing projects have been completed for multiple prominent members of the Apicomplexa.

The data set presented in Kuo et al. [96] consists of 268 orthologous genes from seven species of Apicomplexa and one outgroup ciliate, *Tetrahymena thermophelia.* To this set of genes we appended sequences from the Set8 gene, which has been identified by Kishore et al. [90] as a probable case of horizontal gene transfer from a higher eukaryote to an ancestor of the Apicomplexa.

Of the trees identified as outliers by our method with the dissimilarity map, it appears that most suffer from incorrect annotation or the inclusion of non-orthologous genes. (The most common culprits were sequences from *Eimeria tenella* (Et). See Table 2.1 and Table B.1 for more details.) The interpretation of results from the topological dissimilarity map was less decisive. In most of these cases there were no clearly identifiable problems with the outlying trees or sequences. This result is similar to that found in the simulation studies, suggesting that the incorporation of the branch length information by the dissimilarity map provides superior results.

While the Set8 gene was not identified initially by KDETREES as an outlier gene, its score was very close to the classification threshold, and is the next gene to be classified as an outlier if the tuning parameter is lowered slightly, from 1.5 to 1.3. Since many of the outliers in the analysis seem to be caused by questionable annotation in the Et sequences, we removed this species from the data set and generated new gene trees. In the new analysis, the Set8 gene was identified as the furthest outlier tree.

These results demonstrate the potential applicability of the KDETREES method to the curation of genetic data sets by providing a simple tool for highlighting sequences or alignments that may be of further interest. The successful identification of the Set8 outlier indicates that our method is able to highlight interesting cases which warrant further attention from investigators.

Epichloë.— The application of KDETREES to the set of fungal protein alignments successfully identified the contrived paralogous alignment of LolC and CysD as an outlier. This is a scenario that could easily arise in phylogenomic analysis, where OrthoMCL [103] identified the genes as orthologs, though the group was subsequently broken into separate ortholog sets by application of COCO-CL [82] to the OrthoMCL output. It seems likely that the lolC gene was evolutionarily derived from cysD [159]. Inspection of synteny relationships, and identification by BLAST of remnants of *cysD* that were not identified as genes by `FGeneSH`, indicated that LolC and CysD were indeed paralogous. This result was expected and is indicative of the utility of this program to identify outliers arising from paralogy.

2.5 Conclusion

The ongoing development of ever-cheaper sequencing methods is producing a plethora of data suitable for phylogenomic analysis. One of the great promises of modern genomics is that phylogenetics applied at the genomic scale (phylogenomics) should be especially powerful for elucidating gene and genome evolution, relationships among species and populations, and processes of speciation and molecular evolution. However, genomic data that can now be generated relatively cheaply and quickly, but for which computationally efficient analytical tools are lacking. There is a major need to explore new approaches to undertake comparative genomic and phylogenomic studies much more rapidly and robustly than existing tools allow.

In simulations and applications to biological data, we address particular challenges posed by bioinformatic artifacts, as well as interesting biological phenomena such as gene duplications and horizontal gene transfer. As we observed in the Apicomplexa and fungal data sets, our approach also serves as a means of identifying “interesting” gene trees which may arise from horizontal gene transfer, paralogy, or experimental artifacts such as misannotations or misalignments.

A further advantage of our method is that it may be applied in a straightforward way to phylogenetic reconstruction methods which produce a a sample of many trees as output, rather than a single “best fit” tree. Indeed, methods that produce only a point estimate does not represent the full set of possible phylogenies compatible with the gene sequences. We can circumvent this issue by building a kernel for each gene based on a collection or sample of reconstructed topologies (via the estimated posterior distribution of each gene, for example), rather than using only a point estimate of each gene tree.

In future work we intend to extend our method to clustering trees based on similarity, in addition to identifying outliers. The identification and exclusion of outlier points is an important preliminary step in many clustering methods. The removal of outlier points facilitates better inference at the clustering stage [27, 80, 79].

A long-term goal for this project is to develop a phylogenomic pipeline that is convenient and accessible, as well as robust. To accomplish this aim, important problems that need attention are (1) refinement of gene calls based on comparison among orthologs from multiple genomes and (2) comparing thousands of gene phylogenies across whole genomes. Therefore, our approach is focused on the efficiency of the algorithm in terms of computational complexity and memory requirements, with less emphasis on achieving the highest classification accuracy possible. Such a tradeoff makes our approach more attractive candidate for inclusion in a pipeline for genome-wide phylogenetics as an annotation supplement or as a discovery aid for instances where evolutionary processes deviate significantly from normal.

Chapter 3

Normalizing kernels in BHV treespace

At the time of the publication of this dissertation, material in this chapter currently is in review for publication, with preprints available on the arXiv [175]. This chapter represents further progress in refining the methods pioneered in the previous chapter.

Abstract

As costs of genome sequencing have dropped precipitously, development of efficient bioinformatic methods to analyze genome structure and evolution have become ever more urgent. For example, most published phylogenomic studies involve either massive concatenation of sequences, or informal comparisons of phylogenies inferred on a small subset of orthologous genes, neither of which provides a comprehensive overview of evolution or systematic identification of genes with unusual and interesting evolution (e.g. horizontal gene transfers, gene duplication and subsequent neofunctionalization). We are interested in identifying such "outlying" gene trees from the set of gene trees and estimating the distribution of the tree over the "space of phylogenetic trees."

This paper describes an improvement to the KDETREES algorithm, an adaptation of classical kernel density estimation to the metric space of phylogenetic trees (Billera-Holmes-Vogtmann treespace), whereby the kernel normalizing constants, are estimated through the use of the novel holonomic gradient methods. As the original KDETREES paper, we have applied KDETREES to a set of Apicomplexa genes and it identified several unreliable sequence alignments which had escaped previous detection, as well as a gene independently reported as a possible case of horizontal gene transfer.

The updated version of the KDETREES software package is available both from CRAN, as well as from the development repository on Github [176].

3.1 Introduction

One of the great opportunities offered by modern genomics is that phylogenetics applied on a genomic scale (phylogenomics) should be especially powerful for elucidating gene and genome evolution, relationships among species and populations, and processes of speciation and molecular evolution. However, a well-recognized hurdle is the sheer volume of genomic data that can now be generated relatively cheaply and quickly, but for which analytical tools are lagging. There is a major need to explore new approaches to undertake comparative genomic and phylogenomic studies much more rapidly and robustly than existing tools allow. Here, we focus on the problem of *efficiently* identifying significant *discordance* among a set of gene trees, as well as estimating the distribution of gene trees from the given set of trees.

The KDETREES algorithm introduced in Weyenberg et al. [174] (Chapter 2) is an adaptation of classical kernel density estimation to the metric space of phylogenetic trees defined by Billera et al. [13]. It is a computationally efficient method of estimating the density of the trees over the Billera-Holmes-Vogtmann (BHV) treespace, and relies on a fast implementation of the BHV geodesic distance function provided by Owen and Provan [124]. The method then uses the density estimates to identify putative outlier observations. This paper describes an improvement to KDETREES, whereby the kernel normalizing constants, are estimated through the use of the novel holonomic gradient methods [94, 113].

In our original paper describing the KDETREES method, we propose a nonparametric estimator of the form,

$$\hat{f}(T) \propto \frac{1}{N}\sum_{i=1}^{N} k(T, T_i, h_i).$$

In the KDETREES software, the kernel function implemented is a spherically symmetric gaussian kernel, i.e.,

$$k(T, T', h) \propto \exp\left(-\left(\frac{d(T, T')}{h}\right)^2\right). \tag{3.1}$$

Since we are, for the moment, interested primarily in using the estimator $\hat{f}$ for outlier detection, knowledge of the overall proportionality constant for $\hat{f}$ is not of significant importance. However, it is important to know how the normalizing constant associated with $k(T, T', h)$ varies with the selected bandwidth and with the location of the kernel's center. In our original paper we argued that, in practice, estimates of the normalizing constant do not appear to have significant systematic variation, and that assuming a constant value was a reasonable first approximation. This paper presents basic results and techniques for obtaining better approximate values for these normalizing constants.

In the case case of Euclidean k-space with the usual metric, the kernel (3.1) corresponds to a spherically symmetric multivariate normal distribution centered on the point T', and the kernel normalization constant is given by

$$c(T', h_i) = (2\pi h_i)^{-k/2}. \tag{3.2}$$

Note that not only is there a simple closed form solution for the constant, but the constant is invariant under changes to the central point T'. However, when applied to the BHV treespace with the geodesic metric, not only is such a closed form solution apparently unavailable, but it is also clear that the constant will depend on the location of the central point.

For example, consider the case where $T' = 0$, i.e., the star tree, located at the origin of BHV space. In this case, the kernel integral,

$$c(T', h) = \int_{\mathcal{T}} k(T, T', h)\, dT, \tag{3.3}$$

is symmetric over each orthant comprising BHV treespace. Within each orthant, the integral is equivalent to the normalizing constant of a zero-mean multivariate normal truncated to $\mathbb{R}_n^+$. Thus, expression (3.3) is equivalent to the number of orthants in the space, n_O, multiplied by the corresponding truncated normal constant.

On the other extreme, consider a central tree T' such that every edge is large compared to the bandwidth h, i.e., the tree is relatively far away from any orthant boundary. In this case, the kernel integral will be very close to the value given in expression (3.2). If the central point is placed arbitrarily far away from any orthant boundary, then the integral over any orthant other than the one containing T' can be made arbitrarily small. Thus, the integral over the orthant containing T' itself will be an increasingly good estimate of the entire normalizing constant as the central point is moved further away from orthant boundaries.

The updated version of KDETREES presented in this paper improves on the first generation algorithm by estimating the kernel normalizing constants $c(T', h)$. This is accomplished by finding bounding functions in each orthant which can be more easily integrated than the true kernel function. While some analytic simplification is possible, certain expressions cannot be evaluated other than through numerical methods.

3.1.1 Holonomic Gradient Method

The *holonomic gradient method* (HGM) is a non-stochastic numerical method for calculating certain types of integrals. The HGM is a variation on the gradient descent method of function optimization, and is suitable for application to holonomic functions [117, 94]. Roughly, a holonomic function is a solution to a homogenous ordinary differential equation with polynomial coefficients [182]. Several integrals of interest to statisticians turn out to be expressible as solutions to an optimization problem within a holonomic system.

For our present problem two cases are of particular use. Marumo et al. [113] demonstrates the use of HGM to calculate the normalizing constant for a multivariate normal distribution truncated to the positive orthant, i.e., $\mathbb{R}^n_+$. In addition, Hayakawa and Takemura [65] provides the constants for the so-called exponential-polynomial family of probability densities,

$$f(x|\theta_1, \ldots, \theta_k) \propto \exp\left(x\theta_1 + \ldots + x^k\theta_k\right).$$

As was briefly discussed in the introductory section, BHV treespace is a simplical complex of positive Euclidean orthants, and the normalizing constant for a truncated multivariate normal distribution is an ingredient for a scheme to approximate the kernel constants in BHV treespace. In section 3.3 we show that it is possible to use the normalizing constants for the truncated multivariate normal and the exponential-polynomial family, computed either by HGM or some other method, to construct approximations to the kernel normalizing constants for BHV space.

3.2 Methods

3.2.1 Normalizing Constants

In this paper we use k to denote the *unnormalized* kernel function, i.e., with unit constant of proportionality in (3.1). If we are given a fixed tree T_0 and bandwidth h, our objective is to compute bounds for the integral $\int k(T, T_0, h)\ dT$ over the entire BHV treespace, so that we may normalize the kernel function.

One suitable lower bound function is based on the use of the triangle inequality.

Lemma 3.1. For any pair of trees, $k(T, T', h) \geq \underline{k}(T, T', h)$, where,

$$\underline{k}(T, T', h) := \exp\left(-\frac{(d(T, 0) + d(0, T'))^2}{h^2}\right).$$

Proof. This is an immediate consequence of the fact that the geodesic path between any two trees is the shortest path connecting the trees. In particular, it is shorter than the cone path, $d(T, T') \leq d(T, 0) + d(0, T')$. □

However, $\underline{k}$ does better than simply providing a global lower bound for k. In fact, the bound is sharp, as $\underline{k}$ is equivalent to k whenever the geodesic between T and T' passes through the origin. This turns out to be a quite common occurrence, as geodesics between trees which are not separated by a small number of NNI interchanges are likely to pass through the origin. As a result for much of the space, integrating over $\underline{k}$ will be equivalent to integrating over k itself. Happily, integrating $\underline{k}$ over a single orthant affords an opportunity for analytical simplification.

Theorem 3.1. Let O be an arbitrary fixed orthant in BHV treespace, and let p denote its dimension. Then, the integral of $\underline{k}(T, T', h)$ over that orthant is given by the expression

$$\underline{C}_O(T', h) := \int_O \underline{k}(T, T', h)\ dT = \frac{\pi^{p/2} e^{-d(0,T')^2/h^2}}{2^{p-1}\Gamma(p/2)} \underline{A}(T', h), \tag{3.4}$$

where, if we let $\theta_1 = -2d(0, T')/h^2$ and $\theta_2 = -h^{-2}$, then

$$\underline{A}(T', h) = \int_0^\infty \rho^{p-1} \exp\left(\theta_1\rho + \theta_2\rho^2\right)\ d\rho. \tag{3.5}$$

Proof. The distance $d(T, 0)$ is the usual l_2-norm of the vector of edge lengths for the tree T, and O is the positive orthant $\mathbb{R}^p_+$. Thus, if we express the integral over O in an angular coordinate system,

$$\underline{C}_O(T', h) = e^{-\frac{d(0,T')^2}{h^2}} \int_O e^{-\frac{(d(T,0)^2 + 2d(0,T')d(T,0))}{h^2}}\ dT = e^{-\frac{-d(0,T')^2}{h^2}} \int_0^\infty \int_\Theta e^{\theta_1\rho + \theta_2\rho^2} dV(\rho, \Theta).$$

Now the volume element in $\mathbb{R}^p$ in an angular coordinate system is

$$dV(\rho, \Theta) = \rho^{p-1} d\rho \prod_{k=1}^{p-1} \sin^{p-k-1}(\theta_k)\, d\theta_k,$$

and it so happens that one of the definitions of the Beta function yields

$$\int_0^{\pi/2} \sin^{a-b-1}(\theta)\, d\theta = \frac{1}{2} B((a-b)/2, 1/2).$$

Integrating in all the radial coordinates yields a product of beta functions which telescopes down to the constant appearing in (3.4). This reduces the problem to a single integral in the radial coordinate, which is equivalent to $\underline{A}(T', h)$. □

Unfortunately, the function $\underline{A}(T', h)$ has no general closed form solution. However, there are several methods that we can use to obtain a numerical estimate of this value. The HGM method developed in Hayakawa and Takemura [65] is one such method for obtaining this value. It is also reasonable to calculate this particular integral using classical quadrature methods.

Lemma 3.2. Following the notation of Hayakawa and Takemura [65],

$$\underline{A}(T', h) = \partial_1^{p-1} A_2(\theta_1, \theta_2).$$

Here, $A_2(\theta_1, \theta_2)$ is the normalizing constant for the exponential-polynomial distribution of order 2, the θ are defined as in Theorem 3.1, and ∂_1^m means the m-th partial derivative with respect to θ_1. Furthermore, Hayakawa gives the following equivalence for the first partial derivative

$$\partial_1 A_2(\theta_1, \theta_2) = -\frac{1}{2\theta_2} \left\{1 + \theta_1 A_2(\theta_1, \theta_2)\right\},$$

and for the the higher derivatives, $m \geq 2$, the partials can be expressed recursively in terms of lower order derivatives,

$$\partial_1^m A_2(\theta_1, \theta_2) = -\frac{1}{2\theta_2} \{(m-1)\partial_1^{m-2} A_2(\theta_1, \theta_2) + \theta_1 \partial_1^{m-1} A_2(\theta_1, \theta_2)\}.$$

Proof. See Hayakawa and Takemura [65], Section 2, equations (4) and (7). The latter expression can be easily obtained by differentiation of the expression for the first partial. □

These results are sufficient to use the *hgm* package described in Koyama et al. [94] to implement the lower bound for the orthant integral, $\underline{C}_O(T', h)$. The desired normalization constant for function (3.1) can be decomposed as the sum of integrals over each orthant in BHV space. Thus, if n_O is the number of orthants in the BHV space, then $n_O \cdot \underline{C}_O(T', h)$ is a crude lower bound for the overall normalizing constant. Although this is a poor bound, it can be improved by obtaining better bounds for the contribution from various orthants and adjusting accordingly.

The most obvious orthant to begin with is the orthant containing the "central" tree T', which we shall call $O_{T'}$. This is the orthant where the difference between k and $\underline{k}$ will be the greatest, and thus the largest improvement to the bounding constant is to be found here. Note that in this case, the integral over $O_{T'}$ is given by,

$$C_{O_{T'}}(T', h) = \int_{O_{T'}} \exp\left(d(T', T)^2/h^2\right)\, dT = \int_{\mathbb{R}_+^p} \exp\left(-||x - x_{T'}||/h^2\right)\, dx.$$

This is simply the integral of a radially-symmetric multivariate gaussian kernel centered at the point T over the positive orthant. Such a normalizing constant can also be calculated using HGM, and an implementation is included in the *hgm* R package [94].

Further improvements to the integral for orthants which adjoin directly to $O_{T'}$ can be made by noting that a relationship similar to that of Lemma 3.1 will hold, but with the third point in the triangle inequality being somewhere on the orthant boundary, instead of the origin. However, in practice the improvements to the bounds obtained in this way are quite small, given the typical values of the bandwidths which occur in practice and the small number of orthants to which the calculations apply. For this reason, and in the interests of controlling the overall numerical complexity of the KDETREES algorithm, these “second-order” improvements are not implemented at at this time, but may appear in future updates.

3.2.2 Outlier Test

In Chapter 2, we chose to implement a outlier test of the form, $\hat{f}(T_i) < c^*$, where the critical value c^* is selected using Tukey’s quartile method,

$$c^* = Q_1 - k(Q_3 - Q_1).$$

Here, $\hat{f}(T)$ denotes our density estimate for tree T, and the quartiles, Q_1, Q_3,are calculated using all observed tree density estimates.

Further experimentation with the method has suggested that better performance is obtained if the tree density scores are transformed to the log-scale before the classification step takes place, $\log f(T_i) < c^*$. This transformation was chosen because the raw scores, $f(T_i)$, are themselves bounded below by zero, and the log transformation removes this bound. The quantiles used to compute the critical value are also obtained using the log-transformed scores. Due to the better performance characteristics of this method, the default classifier algorithm for KDETREES has been changed to operate on the log-density scale.

3.2.3 Leaf Edges

The dimension of the orthants comprising tree space is determined by the number of taxa in the trees, with each edge in the fully-resolved tree contributing a dimension to each of the orthants comprising the space. However, the n leaf edges are represented in the space in such a way that the space can be decomposed into a Cartesian product $S \times \mathbb{R}^n_+$. The portion of the space S is associated with the internal edges of the trees, while the positive Euclidean orthant $\mathbb{R}^n_+$ is associated with the leaf edges [13]. Because of this decomposition, there is not a large amount of topological information contained in the portion of the space corresponding to the leaf edge lengths.

If we remove the leaf edges from the calculation, the dimension of treespace can be reduced by approximately half, while retaining the important topological information. This has the benefit of simplifying the overall density estimation problem, as well as the computation of the normalizing constant estimates by the HGM methods. While the original KDETREES algorithm included the leaf edges in the geodesic calculations, the updated version omits them from consideration.

3.3 Results

The updated version of the KDETREES software package is available both from CRAN (the official R package system), as well as from the official development repository on Github [176].

3.3.1 Simulations

A set of simulated datasets were constructed and analyzed, using a similar design as the first simulation described in Chapter 2. The simulations measure the true and false positive rates for identification of known outlier trees within a set of trees drawn from a common distribution. Results of the simulation are summarized as ROC curves, and are presented in Figure 3.1.

3.3.2 Apicomplexa

The Apicomplexa data set presented by [96] consists of trees reconstructed from 268 single-copy genes from the following species: *Babesia bovis* (Bb) [17], *Cryptosporidium parvum* (Cp) [1] from CryptoDB.org [66], *Eimeria tenella* (Et) from GeneDB.org [70], *Plasmodium falciparum* (Pf) [55] and

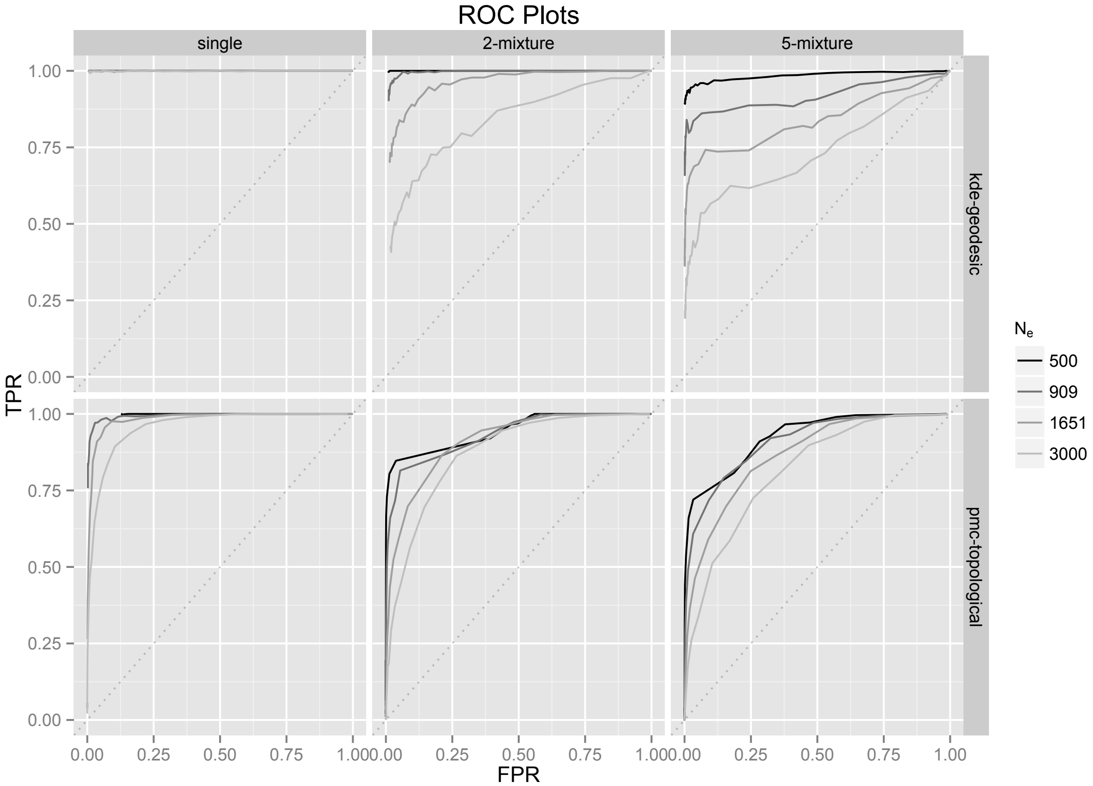


Figure 3.1: Receiver operating characteristic (ROC) plots comparing the updated KDETREES algorithm with Phylo-MCOA. ROC plots summarize the true and false positive rates (TPR and FPR, respectively) for a binary classifier as the tuning parameters are changed. A perfect classifier would be represented as a single point at the upper left corner of the plot, while a completely random scheme would follow the dotted diagonal lines. Curves are shown from simulated coalescent data, using a variety of effective population sizes (N_e). Larger values for N_e correspond to more variability in the generated trees, and thus a more difficult classification problem.

Plasmodium vivax (Pv) from PlasmoDB.org [8], *Theileria annulata* (Ta) [132] from GeneDB.org [70], and *Toxoplasma gondii* (Tg) from Toxo-DB.org [54]. A free-living ciliate, *Tetrahymena thermophila* (Tt) [47], was used as the outgroup. To this set of sequences, we appended the Set8 gene, which has been identified by [90] as a probable case of horizontal gene transfer from a higher eukaryote to an ancestor of the Apicomplexa. This is the same data set analyzed as part of the original KDETREES paper, which was analyzed again with the updated algorithm. The new set of outlier trees is presented in Table 3.1. The newly identified set of outlier trees are presented in a series of supplementary figures which appear in Appendix C. The figures are in ascending order by the kdetrees tree score, i.e., the first tree depicted is the furthest outlying tree.

Table 3.1: Apicomplexa gene sets identified as outliers by the updated KDETREES. Genes which were not identified as outliers by the original algorithm are marked with a ∗. [a]Based on geneset designations in [96]. [b]Geneset represented by GeneID for *Plasmodium falciparum*. (Pf = *Plasmodium falciparum*, Pv = *Plasmodium vivax*, Bb = *Babesia bovis*, Ta = *Theileria annulata*, Et = *Eimeria tenella*, Tg = *Toxoplasma gondii*, Cp = *Cryptosporidium parvum*, and Tt = *Tetrahymena thermophila* (outgroup).)

No.[a]	GeneID[b]	Functional Annotation	Analysis
∗ 472	PF14_0059	hypothetical protein	Tree topology inconsistent with phylogeny. Bb and Cp on same branch, with Ta distant from sister species Bb. Sequence alignment looks good in some regions, but with numerous gaps and other regions with poor alignment. Multiple homopolymer stretches in Pv and Pf.
∗ 478	PF14_0326	hypothetical protein	Tree topology not consistent with phylogeny of the species. Bb branches with the outgroup Tt instead of itâĂŹs closely-related sister species Ta. Poor alignment with numerous gaps, numerous homopolymer stretches, particularly in Et.
488	PF08_0086	RNA-binding protein, putative	Significant sequence length disparity (164 a.a. for Ta vs 1075a.a. for Pf). Generally good sequence alignment in one region of 100 residues; otherwise, alignment is poor.
∗ 505	PF14_0143	protein kinase, putative	Ta/Bb and Pf/Pv not monophyletic; split by outgroup Tt. Good sequence alignment in multiple blocks, but significant sequence length differences. Pf/Pv have multiple insertions and Et and Cp sequences are truncated.
515	PFA0390w	DNA repair exonuclease, putative	Short sequences for Et and Cp. Several homopolymer stretches in Et. Modest to good alignment in multiple blocks, Et being an exception in several regions. Possible incorrect annotation of Et sequence.
∗ 553	PFC0730w	conserved protein, putative	Tree topology inconsistent with phylogeny. Bb and Ta are distant not monophyletic with Pv/Pf. Short regions exhibiting good sequence alignment. Et sequence is truncated.

Continued on next page…

Table 3.1 – continued from previous page

No.^{a}	GeneID^{b}	Functional Annotation	Analysis
∗ 578	PF14_0042	U3 small nucleolar ribonucleoprotein, U3 snoRNP putative	Tree topology very inconsistent with phylogeny; Tg branch with outgroup Tt, Et branch with Bb and Ta. Poor alignment. Significant sequence length differences; Tg sequence is 4126 residues in length, Cp and Tt 2000 residues, Pf and Pv are 450 residues.
∗ 585	PF10_0054	hypothetical protein	Cp exhibits anomalous placement in tree. Significant sequence length differences; Pf, Pv, Tg about 1100 residues, Et only 349 residues, so numerous gaps in alignment. Some regions show good alignment.
∗ 588	PFI1020c	Inosine-5'-monophosphate dehydrogenase	Sequence alignment looks reasonably good. Tree shows Cp branching with outgroup Tt and distant from other Api species. Bb split from Ta.
∗ 630	PFL2120w	hypothetical protein, conserved	Tree topology inconsistent with phylogeny. Cp branching with Pv/Pf. Ta/Bb not monophyletic with Pv/Pf. Several blocks of sequence showing good alignment, but numerous gaps, due mostly to Ta insertions.
641	PFE0750c	hypothetical protein, conserved	Et on a very long branch with other species tightly clustered. Large difference in sequence lengths; 269 residues for Et vs. 848 for Pf. Central region with modest to good alignment; Et exhibited poor sequence identity.
∗ 645	PF14_0635	RNA binding protein, putative	Tree topology looks proper, although Pv and Pf are on a somewhat long branch. Modest to good alignment.
∗ 662	PF11_0463	coat protein, gamma subunit, putative	Multiple homopolymer stretches in Et sequence. Generally good alignment for all but Et; sequence might not be homologous.
∗ 725	PF14_0428	histidine – tRNA ligase	Tree topology appears proper, but Pf/Pv on long branch. Good alignment in two large blocks, but significant gaps and poor alignment in other regions. Et sequence truncated (339 vs. 1000 residues for others). Ta sequence also truncated (583 residues).
∗ 745	PF11_0049	hypothetical protein, conserved	Ta and Bb branch is distant from other Api species, which cluster tightly. Regions of good sequence alignment by with several large gaps. Sequence length differences; Pf and Pv = 3300 residues, Tg and Cp = 2600, Et only 347.
∗ 750	PFE1050w	adenosylhomo-cysteinase (S-adenosyl-L-homocystein e hydrolase)	Ta and Bb somewhat distant from other Api species. Relatively good sequence alignment, although Et sequence truncated (291 vs. 480 for others).

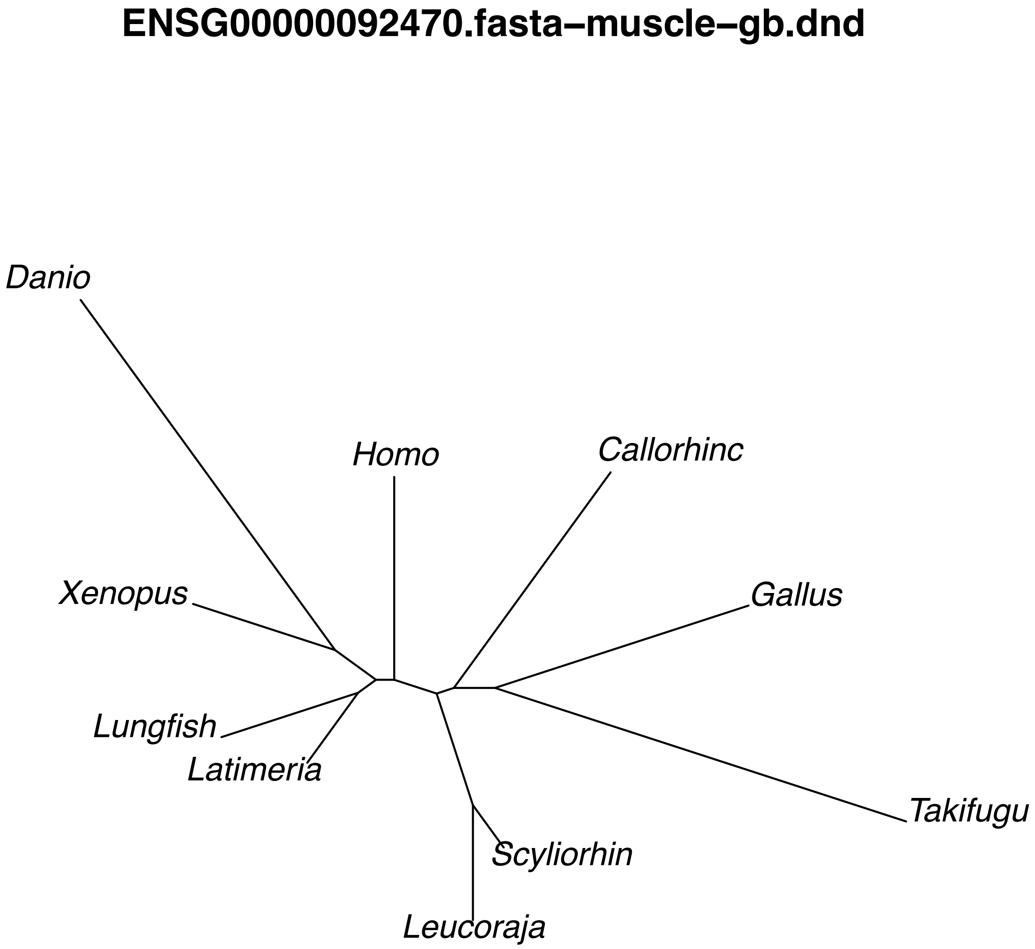


Figure 3.2: An outlier tree from the Lungfish dataset, featuring a unusual topology.

3.3.3 Lungfish

We also analyzed a dataset of 1290 sequences presented by Liang et al. [104]. The sequences were aligned using MUSCLE [44], and trees were reconstructed using the neighbor-joining method [149]. This dataset features 10 taxa, representing tetrapods (*Homo, Gallus, Xenopus*), cartilaginous fish (*Leucoraja, Scyliorhin, Callorhinchidae*), bony fish (*Takifugu, Danio*), and transitional fishes (*Latimeria*, Lungfish). The paper presenting the dataset suggests that the reconstructed trees generally fall into one of the three unrooted topologies which one can defined using the four overarching clades described above, with a few instances of trees where these clades are not reconstructed as expected.

Table 3.2 contains a description of the outlier trees identified in the Lungfish dataset. Figure 3.2 depicts an example of one of the outlying trees which features a clearly unusual topology.

3.4 Discussion

3.4.1 Apicomplexa

The outliers identified by KDETREES in the Apicomplexa dataset are substantially different than those reported in our original paper. In the original paper, many of the outliers were trees containing

Table 3.2: Outliers identified by KDETREES in the Lungfish dataset.

score	tree	notes
-21.08	ENSG00000109762	Disproportionate internal branch.
-19.61	ENSG00000124279	Disproportionate internal branch.
-17.97	ENSG00000115970	Disproportionate internal branch.
-16.69	ENSG00000000457	Disproportionate internal branch, Leucoraja placed with bony fish.
-11.40	ENSG00000142798	Disproportionate internal branch, Callorhinchidae placed with bony fish.
-6.86	ENSG00000180694	Disproportionate internal branch.
-6.53	ENSG00000159733	Lungfish and Callorhinchidae placed with bony fish.
-3.76	ENSG00000185917	Callorhinchidae placed with bony fish.
-2.33	ENSG00000015479	Disproportionate internal branch.
-1.99	ENSG00000165124	Cartilaginous fish do not form clade.
-1.01	ENSG00000189079	Disproportionate internal branch.
2.14	ENSG00000112200	Disproportionate internal branch.
2.74	ENSG00000163512	Disproportionate internal branch.
3.98	ENSG00000164252	Highly disproportionate external edge length.
4.21	ENSG00000134759	Scyliorhinidae placed with bony fish.
4.43	ENSG00000159267	Highly unusual tree topology.
4.71	ENSG00000119431	
4.72	ENSG00000092470	Highly unusual tree topology.
5.15	ENSG00000140263	Xenopus not placed with tetrapods.
5.37	ENSG00000109775	
5.38	ENSG00000145901	
5.97	ENSG00000132952	
6.03	ENSG00000152223	
6.11	ENSG00000128708	
6.11	ENSG00000214367	Xenopus not placed with tetrapods.
6.12	ENSG00000103932	
6.37	ENSG00000099991	Scyliorhinidae placed with bony fish.
6.62	ENSG00000134900	
6.63	ENSG00000165309	Xenopus not placed with tetrapods.
6.87	ENSG00000189306	Disproportionate internal branch.
6.92	ENSG00000151023	Xenopus not placed with tetrapods.
6.98	ENSG00000116863	Disproportionate internal branch.
7.02	ENSG00000196290	
7.43	ENSG00000066136	Disproportionate internal branch.
7.50	ENSG00000106144	
7.54	ENSG00000126777	Disproportionate internal branch.

a single edge much longer than any other edge. This was largely attributable to one or more poorly aligned sequences within the larger multiple sequence alignment. Thus, the disproportionate edges were often leaf edges, and as a result of the changes in the algorithm, the leaf edges are no longer taken into consideration by the default settings. As a result, many of the trees identified as outliers in the original paper are no longer identified as outliers by the default parameter settings of the improved version. However, the software retains the “dissimilarity mode” from the original implementation, which always uses the terminal branch length information.

The cumulative result of the changes in the algorithm is an increased focus on differences in topology in the dataset. The new set of 16 outlying trees differ from the non-outliers primarily in the placement of the Bb, Cp, Ta, and Tt genes within the trees. In the non-outlier trees, Cp generally forms a clade with Tt, while Ta forms a clade with Bb. In the outlier trees, however, these taxa are placed in widely varying locations within the trees, as demonstrated by the drawings of the outlier trees appearing in the series of figures in Appendix C.

Together, these results demonstrate that the updated KDETREES algorithm is more sensitive to topological differences in the trees than the previous version, at the expense of the loss of information from the terminal edge lengths. However, the original functionality using the terminal edge information is still available for use, by setting the appropriate flags.

3.4.2 Simulations

The performance of the classifier with the simulated datasets is substantially better than the original version of the algorithm. Although there is a modest performance penalty associated with the estimation of the normalizing constants, KDETREES remains significantly faster than the competitor PHYLO-MCOA algorithm [37].

When the non-outlier trees are drawn from a single coalescent distribution, the performance of the classifier is nearly perfect, identifying the correct outlier in every simulation iteration, even when the variance of the coalescent distributions (controlled by the effective population size parameter N_e) was quite large. (Of course, due to the nature of the classifier, false positives are inevitable if the tuning parameter is chosen poorly.) In the more difficult cases where the non-outliers were drawn from a mixture of coalescent distributions, the updated algorithm remained superior to the Phylo-MCOA algorithm, showing greater area under the ROC curve for all cases except for the case of the most highly variable 5-part mixture distribution for the non-outlier trees.

3.5 Conclusion

Our proposed method is motivated by the fact that existing methods of phylogenetic analysis and tree comparisons are not adequate for genomic scale phylogenetic analysis, particularly in cases of certain non-canonical evolutionary phenomena. Furthermore, the scenario in our mixed coalescent distribution simulation—where the non-outlier trees are sampled from an unknown mixture of distributions—cannot be handled by parametric methods, with the possible exception of the genome spectral methods. However, even the genome spectral methods ignore possible statistical dependencies between different feature spectra. In contrast, we propose analyzing a collection of gene trees without reducing gene trees to summarizing information. Our KDETREES approach also possesses a considerable advantage in speed over other methods, which is of paramount importance for a tool used in whole-genome phylogenetic analysis.

In addition, one of the applications of our method is an inference on the species tree or a tree that reflects the evolution of most genes in the genome. We can use our method to identify genes which produce discordant trees (outlying trees) and then we can remove them from phylogenetic analysis. By doing this we can use the genes that share the same evolutionary history and we can build a tree that reflects the evolution of the species or that of most of the genome.

We have been interested in developing a phylogenomic pipeline that is convenient and accessible, as well as robust. To accomplish this aim, an important problem that needs attention is to comparing thousands of gene phylogenies across whole genomes. Thus, our approach is focused on the efficiency of the algorithm in terms of computational complexity and memory requirements, with less emphasis on achieving the highest classification accuracy possible. Thus, it is very important for us to improve the accuracy of the method while we maintain the speed of the algorithm. Compared with the original KDETREES and the Phylo-MCOA algorithm, in this paper, we have demonstrated the significant improvement in the accuracy of the method without a corresponding impact on computational speed, as shown by the simulation results in Figure 3.1.

In future work we intend to apply KDETREES to clustering trees based on similarity. Unsupervised clustering is an important method to learn the structure of unlabeled data. The aim of clustering methods is to group patterns on the basis of a similarity (or dissimilarity) criteria where groups (or clusters) are set of similar patterns.

Many traditional clustering algorithms (e.g., K-Means, Fuzzy c-Means, SOM and Neural Gas) take the form of kernel-based algorithms [79, 80, 27]. The use of kernels allows us to implicitly map data into an high dimensional space, called feature space; computing a linear partitioning in this feature space results in a nonlinear separation between clusters in the input space. We intend to further expand on these results by using the kernels to develop similar clustering methods for trees in BHV space.

Acknowledgement

We would like to extend our thanks to Nobuki Takayama and Tomonari Sei for their help in adapting the HGM software for use in our software package.

Chapter 4

Principal Components in BHV treespace

This chapter represents the progress so far in an ongoing project with Dr. Tom Nye at Newcastle University. The software algorithms described here are currently under development, and when completed this will be developed into a more complete methods paper, analyzing sets of trees similar those studied in the previous chapters.

4.1 Introduction

Principal Components Analysis (PCA) is a classical dimension reduction technique which can be applied to multivariate data which takes values in some Euclidean space $\mathbb{R}^s$. The goal is to attempt to find a set of $r < s$ orthogonal vectors which, if the data points are projected onto the subspace spanned by the vectors, minimizes the differences between the original data and the projected points. The coordinate values in the lower-dimensional projected space are intended to be a good approximation of the position in the high-dimensional space. The component vectors are understood as somehow being a more natural representation of the major kinds of variability present in the dataset, which may not correspond well with the units in which the data was originally collected [81].

In this chapter we develop a technique analogous to PCA, but which may be applied to the space of phylogenetic trees described by Billera, Holmes, and Vogtmann [13] (BHV treespeace). The space is constructed as complex of Euclidean orthants, but it is itself not Euclidean, and so certain properties which are assumed for the PCA algorithms to work do not hold in treespace. Fortunately, although not Euclidean, the BHV space is well behaved enough in the sense that it is a CAT(0) space (it is non-positively curved everywhere). This property ensures the existence of geodesic paths, which allows us to formulate definitions for objects in the BHV space which we believe are most analogous to the principal component vectors of classical PCA.

There have been many previous generalizations of the PCA techniques. Some of these techniques retain the assumption that the observations are points in a Euclidean space, while allowing the the space being projected onto to deviate from a linear subspace of the higher dimensional space. Sammon [151] (Sammon Mapping) and Gorban and Zinovyev [59] (Elastic Maps) provide two important examples of these types of generalizations. More recently development has begun on techniques which relax the Euclidean space assumption. Fletcher et al. [53], for example, aims to conduct a PCA-like analysis of the shapes of physical objects (specifically oddly-shaped glands found in the Human anatomy) using so-called "medial axis parameters", which are not elements of any Euclidean space. This work is most similar in spirit to that of Fletcher, but rather than using their set of "principal directions", we develop a distinct object which we believe is better applicable to the BHV treespace.

4.1.1 GeoPhytter

GeoPhytter is a software package and algorithm for constructing "principal geodesics" in the BHV treespace. A geodesic between two trees, T, T', is the shortest continuous path connecting the trees within the metric space, and we use the notation $\Gamma_{TT'}$ to describe the entire geodesic, and $\Gamma_{TT'}(t)$ to refer to a particular tree on the geodesic, at proportion $t \in [0, 1]$ of the way along the geodesic from T to T'. A principal geodesic is defined by a pair of endpoint trees which are chosen so that the total square-distance from the observed trees to the principal geodesic is minimized. The principal geodesic "summarizes the most variable features of a sample of trees ... and can be visualized as an animation of smoothly changing trees." [123]

The algorithm works by performing a stochastic search of the space of possible principal geodesics. After being initialized with some initial pair of points (possibly random), GeoPhytter attempts to improve the fit of the principal geodesic by randomly perturbing the endpoints in a variety of ways. It should be noted that the algorithm is not guaranteed to find the optimal solution to the principal geodesic problem, as it may become trapped within a local optimum.

The algorithm we describe in this chapter is essentially a generalization of GeoPhytter. Using a collection of "vertex" points, we define a principal object. Then, the vertices are perturbed, utilizing the same methods as GeoPhytter, in an attempt to minimize the total square-distance from the observed data to the principal object so constructed. While GeoPhytter limits the number of vertices to two, this algorithm is limited only by the dimension of the underlying treespace and the computational feasibility of the optimization problem.

4.1.2 Convex Hulls

The principal geodesic appears to be a fairly natural object in BHV treespace, since it can be interpreted in similar ways to the first principal component in classical PCA [123]. As a principal geodesic is defined by a pair of endpoints, a set in treespace defined by a collection of three (or more) points seems to be a natural candidate for a generalization to third (or higher) order principal objects.

An immediately attractive object defined by a collection of points is the *convex hull.* The convex hull of a collection of points is the smallest geodesically closed set which contains the given points. The hull is an attractive object for several reasons. First, as a convex object, the projection from an arbitrary point in treespace onto the hull is well-defined and unique. Secondly, the hull can be defined by any number of vertex points. Furthermore, the definition degenerates to the principal geodesic in the case of two vertices , and to the Fréchet mean in the case of a single vertex.

Unfortunately, the convex hull has a critical drawback which significantly weakens its claim to be the natural higher-order generalization of the principal geodesic. In Euclidean space the convex hull formed from r points has dimension at most $r-1$. In other words, increasing the number of points used to specify the hull has a very well-understood effect on the dimensions of the object constructed. However, in BHV space, it turns out that the dimension of a convex hull can grow faster than the number of points used to define it. The following example demonstrates that the convex hull of 3 points in the BHV space of unrooted trees on six leaves can fill a 3-dimensional region of space, rather than the two dimensional surface one might hope to obtain.

Example 4.1. Consider three unrooted trees in $\mathcal{T}_6$ with the topologies shown in Figure 4.1, and the internal branches labeled as shown. Tree 1 is separated from both Trees 2 and 3 by a single NNI move, while Trees 2 and 3 are separated from each other by two NNI moves. Suppose we assign to Tree 1 the internal lengths $(x, y, z) = (1, 1, 2)$, to Tree 2 the lengths $(w, y, z) = (2, 1, 1)$, and to Tree 3 the lengths $(v, x, z) = (2, 1, 1)$. Figure 4.2 depicts the geodesics connecting the three trees within the BHV treespace. Shown is the portion of the space corresponding to the internal branches of the trees which is orthogonal to the length of the z-edge.

Consider the surface formed by the point representing Tree 1 and the points where the geodesics Γ_{12} and Γ_{13} intersect the orthant boundary, with coordinates $(x, y, z) = (0, 1, 1\frac{2}{3})$ and $(x, y, z) = (1, 0, 1\frac{2}{3})$, respectively. If we restrict our attention to only this particular orthant, then the hull of the three points defines a two dimensional surface.

Now, note that Γ_{23} also intersects the orthant containing Tree 1. Although the intersection is only at a single point, with coordinates $(x, y, z) = (0, 0, 1)$, the convex hull must also contain all geodesics from this new point to the surface we just defined. However, the new point is not co-planar with the three points defining the surface, and since within the orthant we are simply dealing with Euclidean space, the four non-planar points form a hull which is 3-dimensional. This 3-dimensional object must, by definition, be contained within the convex hull of the three trees, as defined in the full BHV space.

Thus, we have shown that it is possible for the convex hull to contain a region of higher dimension than expected, given the number of vertices used in the definition. In fact, the trees used in this example are not particularly special. The dimension escalation problem described is possible whenever a geodesic connecting two vertices passes through an orthant containing another vertex, a situation which is by no means uncommon for arbitrary sets of vertex trees.

One consequence of the possible explosion of the convex hull dimension is that it is quite difficult to explicitly decide if a particular point in BHV space is within a given convex hull or not. The difficulty in actually constructing and defining the convex hull gives rise to difficulty in the development

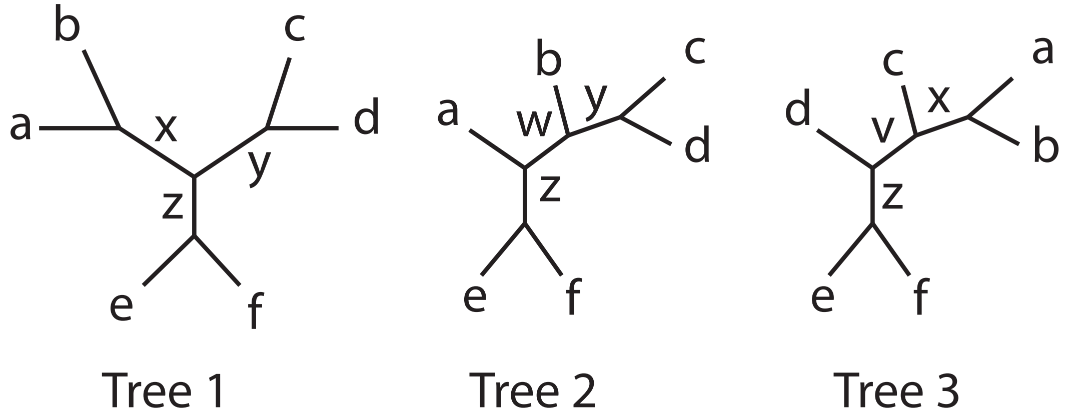


Figure 4.1: Three trees used in the space-filling convex hull example.

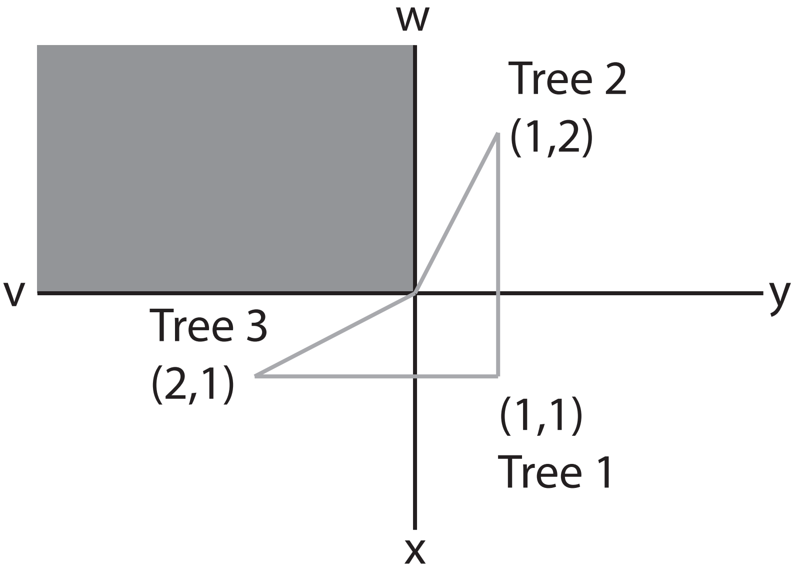


Figure 4.2: Plot showing the geodesics connecting the three trees. The x-dimension of the orthants has been suppressed in this illustration, although it may be understood to be extending orthogonally out of the page.

```
function WeightedMean($\vec{v}, \vec{w}, \epsilon$)
    $y \leftarrow v_1$                                  ▷ Initialize at first vertex.
    $n \leftarrow 1$                                    ▷ Iteration counter.
    repeat
        $y' \leftarrow y$                               ▷ Store old value.
        $n \leftarrow n + 1$
        $\lambda \leftarrow 1/n$
        for $i$ in $1$: length($\vec{v}$) do           ▷ For each vertex,
            $t \leftarrow \frac{2\lambda\, w_i}{1+2\lambda\, w_i}$
            $y \leftarrow \Gamma_{yv_i}(t)$             ▷ Move proportion $t$ along geodesic to vertex.
        end for
    until $d(y, y') < \epsilon$                         ▷ Convergence check.
    return $y$
end function
```

Figure 4.3: A deterministic algorithm for approximating a weighted Fréchet mean.

of algorithms for projecting onto it, and thus with obtaining a "principal hull" via the stochastic optimization process described in Nye [123]. Furthermore, even if a oracle existed which was able to construct a principal hull, any interpretation of the hull would remain extremely challenging. Without having a good way to construct the hull, any effort to visualize the hull itself, or what sort of variation it represents, would seem virtually impossible.

4.1.3 Fréchet Means

The weighted Fréchet mean (also known as a *barycenter* of a distribution F) is a generalization of the notion of the usual arithmetic mean to a metric space, defined by,

$$m(F) := \arg\min_{x \in \mathcal{T}} \int_{\mathcal{T}} d(x,t)^2 \; dF(t).$$

It is clear that this definition coincides with the familiar arithmetic mean in the Euclidean case. However, this definition can be used with any metric space, and in particular it is compatible with the BHV treespace $\mathcal{T}_n$.

For the purposes of this chapter we shall concern ourselves only with the special case of a weighted mean of a finite set of r vertices, $\vec{v} := (v_1, \ldots, v_r)$, in a fixed treespace $\mathcal{T}_n$,

$$m(\vec{v}, \vec{w}) := \arg\min_{x \in \mathcal{T}_n} \sum_{i=1}^{r} w_i \, d(x, v_i)^2. \tag{4.1}$$

The weights vector $\vec{w}$ encodes the probability mass assigned to each vertex by the probability distribution. If the space is CAT(0), as is the case with BHV space, then such a minimizer exists and is unique [9, Theorem 2.4].

The algorithm described in Figure 4.3 approximates the weighted Fréchet mean defined in (4.1). The sequence of intermediate values of the algorithm is shown to converge to the minimizer $m(\vec{v}, \vec{w})$ by Bačák [9, Theorem 3.4].

4.2 Methods

Consider the *locus of Fréchet means* (LFM) for a fixed vertex set of size r,

$$M(\vec{v}) := \{m(\vec{v}, \vec{w}) : \vec{w} \text{ is a set of probability weights}\}.$$

We claim that the LFM is a more natural way to generalize a principal object in the BHV space from a set of specified vertices, given that the convex hull exhibits the dimension explosion problem discussed in the introduction. The LFM has the basic properties one might assume that a principal object should have: The vertices themselves are contained in the LFM, $\vec{v} \subset M(\vec{v})$, since placing all of the probability mass on a single vertex implies that the mean lies on that vertex. Furthermore, the geodesics connecting the vertices are contained in the LFM, $\Gamma_{v_i v_j} \subset M(\vec{v})$, since the weighted Fréchet mean of two points lies on the geodesic connecting them. However, the most desirable property of the set $M(\vec{v})$, which sets it apart from the convex hull, is that the dimension is limited by the number of vertices used to define it, minus one.

Unfortunately, the locus of Fréchet means is not a perfectly ideal object to serve as a generalization of a principal component. The locus of Fréchet means is not, in general, convex. If such were the case, then $\text{hull}(\vec{v}) \subset M$, and we have already shown cannot be the case, as the dimension of the convex hull in the example exceeds $r - 1$. However, not all is lost, as the locus of means is closed and bounded, and this guarantees that a projection onto $M(\vec{v})$ does exist, in the sense that there exists a (possibly non-unique) point in $M(\vec{v})$ which minimizes the distance to the projected point.

Theorem 4.1. The weighted Fréchet mean $m(\vec{v}, \vec{w})$ is a jointly continuous function of the weights and the vertices.

Proof. This proof uses the fact that the Bačák [9] algorithm for obtaining the mean or median is continuous.

Consider the map used by the to update the estimate: $y \mapsto \Gamma_{yv}(t)$. The new point is obtained by moving some distance along the geodesic from the current point to one of the vertices. This updating step is jointly continuous in the starting point, the vertex, and the proportion t.

First, note that the complement of the movement proportion is

$$1 - t_i = (1 + 2\lambda w_i)^{-1}.$$

This is clearly continuous for $w_i \in [0, 1]$, and thus so are t_i and $\Gamma_{yv_i}(t_i)$. Next, consider moving both the starting location and the target vertex up to distance ϵ from their original positions. An application of Sturm [163, Corollary 2.5] (see Figure 4.4) yields

$$d(\Gamma_{yv}(t), \Gamma_{y'v'}(t)) \leq (1 - t)d(y, y') + td(v, v') \leq \epsilon.$$

Next, we note that by the triangle inequality,

$$d(\Gamma_{yv}(t), \Gamma_{y'v'}(t')) \leq d(\Gamma_{yv}(t), \Gamma_{y'v'}(t)) + d(\Gamma_{y'v'}(t), \Gamma_{y'v'}(t')).$$

Since we have shown that both of the distances in the right hand term are continuous, we may conclude that $y \mapsto \Gamma_{yv}(t)$ is jointly continuous in all of the arguments.

The composition of continuous maps are continuous, so this also shows that the ending point after a finite number of iterations of the Bačák algorithm is jointly continuous in the weights and the vertices.

Now using the algorithm we construct a pair of sequences, $\{y_i\} \to m(\vec{v}, \vec{w})$ and $\{y'_i\} \to m(\vec{v}', \vec{w}')$. Since both sequences converge to their respective Fréchet means, there exists a N_ϵ such that $d(y_i, m(\vec{v}, \vec{w})) < \epsilon$ and $d(y'_i, m(\vec{v}', \vec{w}')) < \epsilon$ for any $\epsilon > 0$ and $i \geq N_\epsilon$. Finally, we note that

$$\begin{array}{rcl} d(m(\vec{v}, \vec{w}), m(\vec{v}', \vec{w}')) & \leq & d(m(\vec{v}, \vec{w}), y_{N_\epsilon}) + d(y_{N_\epsilon}, y'_{N_\epsilon}) + d(y'_{N_\epsilon}, m(\vec{v}', \vec{w}')) \\ & \leq & 2\epsilon + d(y_{N_\epsilon}, y'_{N_\epsilon}). \end{array}$$

Since we have proven that the map consisting of finite number of iterations of the algorithm is continuous, we may conclude that $m(\vec{v}, \vec{w})$ is continuous. □

4.2.1 Projecting onto the locus of weighted means

The continuity result above gives us some hope of using numerical methods to attack the problem of projection onto the LFM. Given a fixed vertex set of size r, and an algorithm for computing a weighted Fréchet mean, the problem becomes a fairly straightforward, although non-convex, constrained optimization problem.

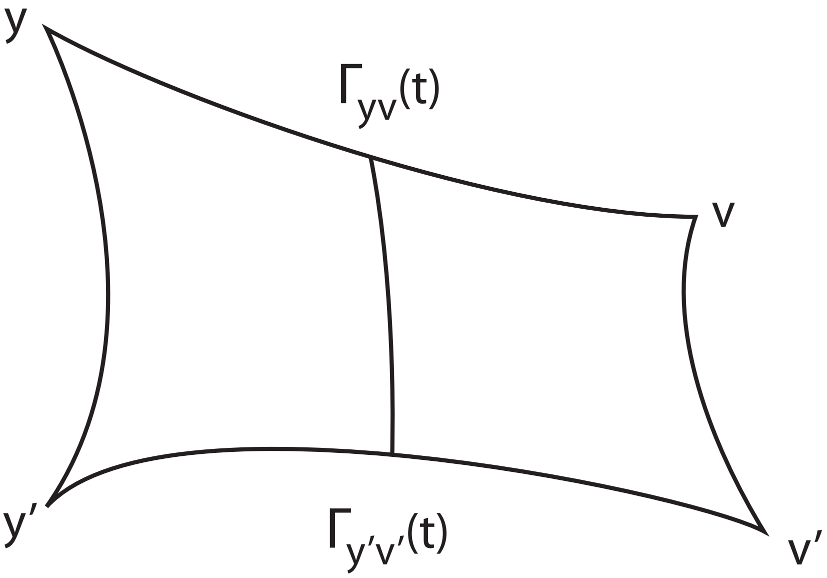


Figure 4.4: Illustration of Sturm [163, Corollary 2.5].

The objective function which must be optimized in the projection of a point x onto a LFM may be expressed as a function of the weight vector, $\Omega_x(\vec{w}) = d(x, m(\vec{v}, \vec{w}))$. The constraints placed on $\vec{w}$ for it to be a valid weight vector are fairly straightforward linear constraints, and so the problem can be attacked using existing techniques. For example, the R package provides the function CONSTROPTIM, which is designed to handle this type of problem [97].

We should note at this point that the non-convexity of the LFM allows for the possibility of a non-unique global optimum for Ω_x, and thus for the value of the projection itself. However, this does not actually interfere with our goal of finding a principal LFM, as the criterion used to definition of that object depends only on the optimal value of Ω_x. If there are multiple distinct points in the LFM which all optimize Ω_x, then it is immaterial to which of these points is actually found by the projection operation.

The potential for local optima thus presents the major difficulty in the the projection problem. However, in the course of practical use of our technique we do not expect the problem to be particularly severe. Although it is possible in theory to define a principal LFM using an arbitrary number of vertex points, it seems unlikely that there will be much practical use in trying to fit a LFM with more than perhaps four or five vertices. One reason for this is that the vertex optimization technique described in the subsequent section becomes very difficult as the number of vertices increases. A second important consideration which limits the number of vertices is the difficulty with both visualization and interpretation of LFMs defined using large vertex sets. Given these considerations, we will assume that practical applications of the technique deal exclusively with very modestly numbers of vertices.

One heuristic which is commonly used with non-convex optimization problems is to start the optimizer at a variety of points in the space. Hopefully, the optimizer will converge to the same solution in every case, but if not, the best solution found may still be used. It should be noted that if the number of vertices to be used in the LFM definition is kept small, then the dimension of the space of possible weight vectors over which Ω_x must be optimized is similarly limited. For vertex sets on the scale described above, the dimension of the space is not large enough to make a relatively dense exploration of starting positions computationally infeasible. By employing this heuristic, we may may be reasonably confident the projection algorithm will obtain a good estimate of the true distance from x to $M(\vec{v})$.

4.2.2 Searching for optimal vertices

The search for a set of optimal vertices for the principal LFM may be carried out in the same manner as in GeoPhytter, once a projection algorithm is available. Four perturbation methods are used to propose changes to the set of LFM vertices:

- Gaussian Random Walk
- Nearest Neighbor Interchange
- Random Jump
- End-point move

See Nye [123], for a detailed description of these perturbation operations.

The algorithm operates by iteratively proposing a new vertex set, projecting the observed data onto the new LFM, and then checking to see if the new vertex set improves or degrades the fit. If the fit is improved, the new vertex set is retained, otherwise it is discarded. The process may be repeated until the user is satisfied that the algorithm has converged.

It should be noted that this method of perturbing the vertex set may allow us to make a significant optimization of the projection algorithm discussed in the previous section. If a single vertex within $\vec{v}$ is perturbed slightly, then the continuity of the Fréchet mean implies that for a fixed $\vec{w}$, $m(\vec{v}, \vec{w})$ will shift in a continuous manner, and that the entire LFM as a whole will not move discontinuously. While this does not guarantee that the projection of a point x is continuous in $\vec{v}$, it does suggest that the vector of weights associated with the projection of x onto $M(\vec{v})$ is a good starting location heuristic for the projection algorithm when calculating the new projection of x onto the proposed LFM.

4.2.3 Visualization of the LFM

Although perhaps not as baffling as the convex hull, the LFM in treespace is still a somewhat exotic construction. For the principle geodesic Nye [123] suggests visualizing the principle geodesic in the form of a smooth animation displaying all of the points on the geodesic between the two endpoint trees. It is argued that this allows the viewer to visually interpret the main type of variability present within the data.

This technique suggests a similar approach for a higher order principal LFM. Since the points on the LFM are parameterized by the weight vector, $\vec{w}$, we may allow the user to explore the space by continuously varying the weight vector, while viewing an animation of the corresponding trees on the LFM. If the number of vertices used to define the LFM is kept modest, then a manual exploration of this space is reasonable for the same reasons that it is feasible to use the multiple starting location heuristic in the projection algorithm. While the user interface to required to achieve this cannot be as straightforward as in case of the principal geodesic, this method of interpretation should allow the viewer to get a sense for what types of variability are represented by the LFM.

4.2.4 Datasets

When the coding is complete, the method will tested by application to three empirical datasets. The first empirical dataset was originally presented by Archibald and Roger [7], and consists of chaperonin genes sequenced in archea. There are 12 sequences observed in six species: *Pyrodictium occultum, Aeropyrum pernix* and *Pyrobaculum aerophilum*, together with 3 closely related *Sulfolobus* species. This data has been analyzed by several similar techniques, including the previous versions of GeoPhytter [122, 123, 51]. This example is chosen primarily to facilitate comparisons with previous methods.

The second dataset to be analyzed was first presented by Liang et al. [104], and concerns the relationship between tetrapods (*Homo, Gallus, Xenopus*), bony fishes (*Takifugu, Danio*), cartilaginous fishes (*Leucoraja, Scyliorhin, Callorhinchidae*), and two transitional fishes (*Latimeria* and the Lungfish). The dataset contains 1290 trees and the original study suggests that the trees generally cluster into three distinct topologies. This dataset has also been analyzed by the methods of Chapter 3, which has identified 37 putative outliers observations in the data. Analyses are to be performed both with and without these putative outliers in the dataset, to investigate the sensitivity of the method to outlier observations.

The third dataset is presented in Nye [122] and reanalyzed in [123]. It consists of a parametric bootstrap sample obtained from a source tree with 41 taxa representing important eukaryotes, along with an archean outgroup. The tree contains two long branches, and is analyzed in order to better understand the effect of the long branches on the new algorithm. The source tree was originally presented by Brinkmann et al. [18] and the set of new trees was simulated by first generating amino acid alignments within the base tree using seq-gen [140]. Each generated alignment contains 300 base pairs and was simulated using the WAG+4Γ model of evolution. The results were then fed into the phyML software to obtain an ML estimate tree for each sequence.

4.3 Results

We have named the software implementing the principal LFM GeoPhytter+, and at the time of the publication of this dissertation it is in the early stage of code development, and is not yet suitable for distribution.

Chapter 5

Future Directions

This chapter briefly describes possible programs for future improvement on the techniques introduced in this dissertation.

5.1 Improved LFM projection algorithms

While the LFM is continuous, it is not, in general, built out of convex segments within each orthant, it is instead "curved", as demonstrated by an argument outlined in a personal communication from TM Nye (June 2015). However, it appears that it is possible to construct a local tangent approximation to the LFM at points within orthant interiors.

First, we note that the geodesic connecting a tree x with the vertex v_i is characterized by a set of splits $A^{(1)}_{xv_i}, \cdots, A^{(r_i)}_{xv_i}$ in the tree x and $B^{(1)}_{xv_i}, \cdots, B^{(r_i)}_{xv_i}$ in the tree v_i. Furthermore, suppose we restrict our attention to an open set R_i, such that all $x \in R_i$ form geodesics Γ_{xv_i} which traverse the same common sequence of orthants, and thus possess a common set of splits. (Recall that Γ_{xy} is the geodesic connecting trees x and y.) Within such a set, the distance $d(x, v_i)$ may be written in the following form, which is a generalization of the expression given by Owen and Provan [124], obtained by dropping the assumption that the geodesic endpoints have no splits in common:

$$d(x, v_i)^2 = ||\mathcal{A}_{xv_i} + \mathcal{B}_{xv_i}||^2 + ||\mathcal{C}_{xv_i} - \mathcal{D}_{xv_i}||^2.$$

Here, $\mathcal{A}_{xv_i}$ and $\mathcal{B}_{xv_i}$ are r_i-dimensional vectors defined as, e.g., $\mathcal{A}_{xv_i} = \Big(||A^{(1)}_{xv_i}||, \cdots, ||A^{(r_i)}_{xv_i}||\Big)$, with

$$||A^{(l)}_{xv_i}|| = \Big(\sum_{e \in A^{(l)}_{xv_i}} e^2\Big)^{\frac{1}{2}},$$

which is the norm of the vector of edge lengths found in $A^{(l)}_{xv_i}$. The vectors $\mathcal{C}_{xv_i}$ and $\mathcal{D}_{xv_i}$ are defined similarly, using the edge lengths in the splits shared by both x and v_i, respectively.

If we let $||x||$ denote the norm of all (non-leaf) edges in the tree x, then we can rewrite the previous distance expression as,

$$d(x, v_i)^2 = ||x||^2 + ||v_i||^2 + 2\langle \mathcal{A}_{xv_i} | \mathcal{B}_{xv_i} \rangle - 2\langle \mathcal{C}_{xv_i} | \mathcal{D}_{xv_i} \rangle.$$

This leads to the following form for the objective function which is minimized in the definition of the Fréchet mean,

$$\Omega(x) = \sum_{i=1}^{r} w_i\, d(x, v_i)^2 = ||x||^2 + \sum_{i=1}^{r} w_i\, (||v_i||^2 + 2\langle \mathcal{A}_{xv_i} | \mathcal{B}_{xv_i} \rangle - 2\langle \mathcal{C}_{xv_i} | \mathcal{D}_{xv_i} \rangle). \tag{5.1}$$

Now suppose we wish to project the point x onto $M(\vec{v})$ and that we are able to construct a starting point $x \in \bigcap_{i=1}^{r} R_i$, which is near to the true projected point, $\underset{m \in M(\vec{v})}{\arg\min}\, d(x, m)$.

Consider the gradient of the objective function $\nabla \Omega(x)$. It would appear that it may be possible to use (5.1) to construct a linear approximation to the gradient, which in turn defines a local tangent plane, which we can use to approximate the LFM near x. If so, points on this tangent plane can be described by a weighting vector $\vec{w}$ and a set of vectors $\vec{\alpha}$, and the pair can be chosen such that the weight vector corresponds locally to the weights in the Fréchet mean. This approximation suggests the algorithm described in Figure 5.1, which exploits the local linearity to iteratively improve an initial guess for the weights associated with the projection of x onto $M(\vec{v})$.

```
function ProjectLFM($x, \vec{v}$)                      ▷ Project a point $x$ onto $M(\vec{v})$.
    Initialize $\vec{w}$ somehow.                      ▷ Possibly randomly, or using a previous guess.
    repeat
        $z \leftarrow m(\vec{v}, \vec{w})$             ▷ Find the point on the LFM associated with $\vec{w}$.
        Use analytical form of $\Omega$ to obtain tangent vectors to LFM at $z$.
        Update weights $\vec{w}$ by optimizing distance to $x$ on the tangent plane.
        $z^* \leftarrow z$                             ▷ Store old value.
        $z \leftarrow m(\vec{v}, \vec{w})$             ▷ Use new weights to find new point on LFM.
    until $d(z, z^*) < \epsilon$                       ▷ Converge if the projected point does not move.
    return $z, \vec{w}$                                ▷ The weights may be useful in subsequent steps.
end function
```

Figure 5.1: An algorithm which attempts to locate the projection of a tree x onto the LFM of the vertices $\vec{v}$ by constructing an approximating tangent plane.

5.2 kdetrees for big data

A significant portion of the computational effort in the KDETREES method is expended in the calculation of the pairwise distance matrix. The size of this distance matrix grows with the square of the number of trees in the dataset, and the difficulty of calculating the individual elements grows with the cube of the number of tips in the trees.

While we have demonstrated that the technique is computationally feasible for datasets on the scale of 20,000 trees each containing 13 tips, beyond this size the computational time begins to become prohibitive when using contemporary commodity hardware. Two avenues for addressing this problem are immediately apparent. The first is simply to parallelize the computation of the pairwise distance matrix. As currently implemented, R, and therefore KDETREES utilizes a single processor thread for all computations. However, the elements of the distance matrix are essentially independent of each other (for the purposes of direct computation), and this naturally suggests an opportunity for significant improvement in speed through parallel evaluation of the matrix elements.

A second, and perhaps more fruitful, avenue of approach may be found in the framework developed by Lawson and Adams [99]. This framework provides a systematic approach to the question of how to extract information from a partially computed distance/similarity matrix, by focusing computational effort on obtaining exact values for those elements which carry the most non-redundant information, and providing an emulator which is able to quickly compute estimates for the missing entries.

The KDETREES method with large numbers of trees in the input appears to be a prime candidate for the application of the framework. If the trees in the dataset are clustered, then we may expect that the entries in the distance matrix to be dependent, as all trees in a cluster are close to each other and far from trees in other clusters. Simulations conducted by [99] suggest that the framework is particularly effective for efficient reconstruction of the distance matrix in this type of situation.

The extension of KDETREES to incorporate the Lawson Framework will require exploration of the space of options available for both the *choice function*, which selects the next unobserved element to be evaluated, and for the *emulator* which quickly imputes missing values. While Lawson makes several possible suggestions for each component, it is not clear which of these, if any, will be well adapted for use with KDETREES.

Appendix A

Symbol Glossary

Symbols are grouped by the chapter in which they first appear. A small number of symbols may be redefined in later chapters when it is deemed unlikely that confusion will occur.

Chapter 1

$||\cdot||$ Vector norm

$|\cdot|$ Set cardinality

$\ominus$ Set symmetric difference

$\{abc\ldots|xyz\ldots\}$ a split (or quartet) of tip taxa in a phylogenetic tree

β Rate matrix normalizing constant

$d(\cdot,\cdot)$ Distance function

$D(T)$ Tip-to-tip pairwise distance matrix for tree T.

κ transition/transversion rate ratio

$L(\cdots)$ Likelihood function.

$l(\cdots)$ Log-likelihood function

n The number of states available to a Markov process, or the number of tips in a phylogenetic tree.

n_{ij} Number of loci in a sequence alignment at where the the first sequence expresses character i and the second sequence expresses character j.

$P(t)$ Markov process transition probability matrix

π Character stationary distribution

Q Markov process transition rate matrix

$Q(T)$ The set of quartets found in tree T

$\mathbb{R}^d$ The Euclidean space with dimension d

$\mathbb{R}^d_+$ The orthant in $\mathbb{R}^d$ with all *non-negative* coordinates

$S(T)$ The set of splits found in tree T

t evolutionary "time" separating two sequences, actually the expected number of character substitutions per site in the alignment

T a tree in $\mathcal{T}_n$

$\mathcal{T}_n$ The space of trees on n leaves

$v(T)$ A tree vectorization function

Chapter 2

δ kernel shape parameter

$f(T)$ probability density at tree T

$\hat{f}, \hat{g}$ estimates of the probability density, including all observations, and the leave-one-out estimator, respectively

h kernel bandwidth

$k(T, T')$ kernel function centered at T', evaluated at T

κ classification tuning parameter

N number of trees in sample

n number of tips in trees

n_{eff} effective population size for the coalescent process

Q_1, Q_3 quartile values

Z a partition function

Chapter 3

0 the star tree

A_k Normalizing constant for degree k exponential-polynomial distribution

$B(\cdot, \cdot)$ the Beta function

$c(T, h)$ volume of unnormalized kernel with bandwidth h centered on tree T.

F a distribution

$\underline{C}_O(T', h)$ lower bound for the integral of $\underline{k}(T, T', h)$ over the orthant O.

$\underline{k}$ A lower bound function for the kernel k.

ρ radial coordinate

S portion of the BHV space corresponding to the internal edges

θ exponential-polynomial family parameters

Θ angular coordinate vector

dV volume element of a polar coordinate system

Chapters 4–5

$\mathcal{A}_{xy}, \mathcal{B}_{xy}, \mathcal{C}_{xy}, \mathcal{D}_{xy}$ vectors containing the sequence of splits corresponding to the geodesic connecting x to y.

$\Gamma_{TT'}$ the entire geodesic connecting T to T'

$\Gamma_{TT'}(t)$ a single point on the geodesic, parameterized by $t \in [0, 1]$

$\mathbf{hull}(\vec{v})$ the convex hull of a vertex set

$m(\vec{v}, \vec{w})$ a weighted Fréchet mean

$M(\vec{v})$ the locus of Fréchet means for all valid weightings

r total number of vertices in vertex set

R_i an open set for where every tree $x \in R_i$ defines a geodesic Γ_{xv_i} which passes through the same sequence of orthants

$\vec{v}$ a set of vertices

$\vec{w}$ a probability weight vector

Appendix B

Supplementary Material for Chapter 2

Table B.1: Analysis of Apicomplexa gene-sets identified as outliers. Pf = *Plasmodium falciparum*, Pv = *Plasmodium vivax*, Bb = *Babesia bovis*, Ta = *Theileria annulata*, Et = *Eimeria tenella*, Tg = *Toxoplasma gondii*, Cp = *Cryptosporidium parvum*, and Tt = *Tetrahymena thermophila* (outgroup).

Gene ID	Functional Annotation	Analysis
PF08_0086	RNA-binding protein, putative	Significant sequence length disparity (164 a.a. for Ta vs 1075a.a. for Pf). Generally good sequence alignment in one region of 100 residues; otherwise, alignment is poor.
PF13_0228	40S ribosomal subunit protein S6, putative	Tt sequence much longer than all others; long N-terminal and C-terminal extensions. Very good alignment in blocks, but with lengthy insertions for outgroup Tt. Possible incorrect annotation of Tg sequence.
PFA0390w	DNA repair exonuclease, putative	Short sequences for Et and Cp. Several homopolymer stretches in Et. Modest to good alignment in multiple blocks, Et being an exception in several regions. Possible incorrect annotation of Et sequence.
PFF0285c	DNA repair protein RAD50, putative	Poor alignment in general. Three modest blocks (50-100 aa) of reasonable sequence alignment. Et sequence contains long homopolymeric stretches. Pf and Pv have long insertions that might be translated introns.
PFL1345c	Radical SAM protein, putative	Relatively short sequence for Et. Homopolymeric stretch at N-terminus of Tg. Modest to good alignment in blocks.
PFE0750c	hypothetical protein, conserved	Large difference in sequence lengths; 269 residues for Et vs. 848 for Pf. Central region with modest to good alignment; Et exhibited poor sequence identity suggestion it might not be a homologue.
PF10_0043	ribosomal protein L13, putative	80 residue N-terminal extension in Tg. Good sequence alignment, with Tt (outgroup) being an exception. Tt sequence might not be a homologue.
PF11_0463	coat protein, gamma subunit, putative	Multiple homopolymer stretches in Et sequence. Generally good alignment for all but Et; sequence might not be homologous.
MAL13P1.22	DNA ligase 1	Homopolymer stretches in Et sequence with poor alignment to other sequences. Et sequence might be incorrectly annotated and/or might not be homologous.

Continued on next page…

Table B.1 – continued from previous page

Gene ID	Functional Annotation	Analysis
PFB0550w	Peptide chain release factor subunit 1, putative	Short sequence for Et (132 residues), with long homopolymer stretch. Other sequences are approximately 425 a.a. in length. Generally good alignment, even for Et over a short region (50 residues). Possible incorrect annotation of Et sequence.
PFF0120w	putative geranylgeranyl-transferase	Two homopolymer stretches (serine) in Et sequence. Moderately good alignment. Possible incorrect annotation of Et sequence.
PFD0420c	flap exonuclease, putative	Very discrepant sequence lengths; 179 a.a. for Et vs. 2213 a.a. for Tt. All other sequences 500 – 600 residues in length. Good alignment over several regions, although sequence for Et is absent in portions of these regions. Very long N-terminal extensions and insertions in Tt sequence. Possible incorrect annotations for Et and Tt.

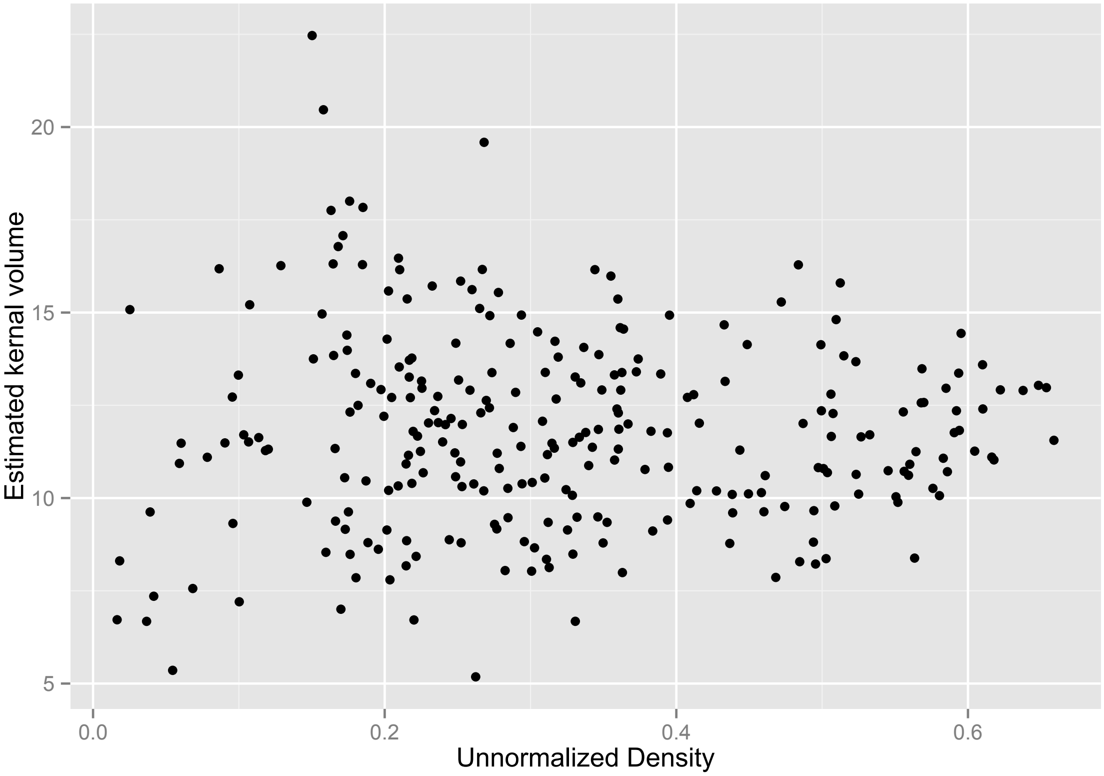
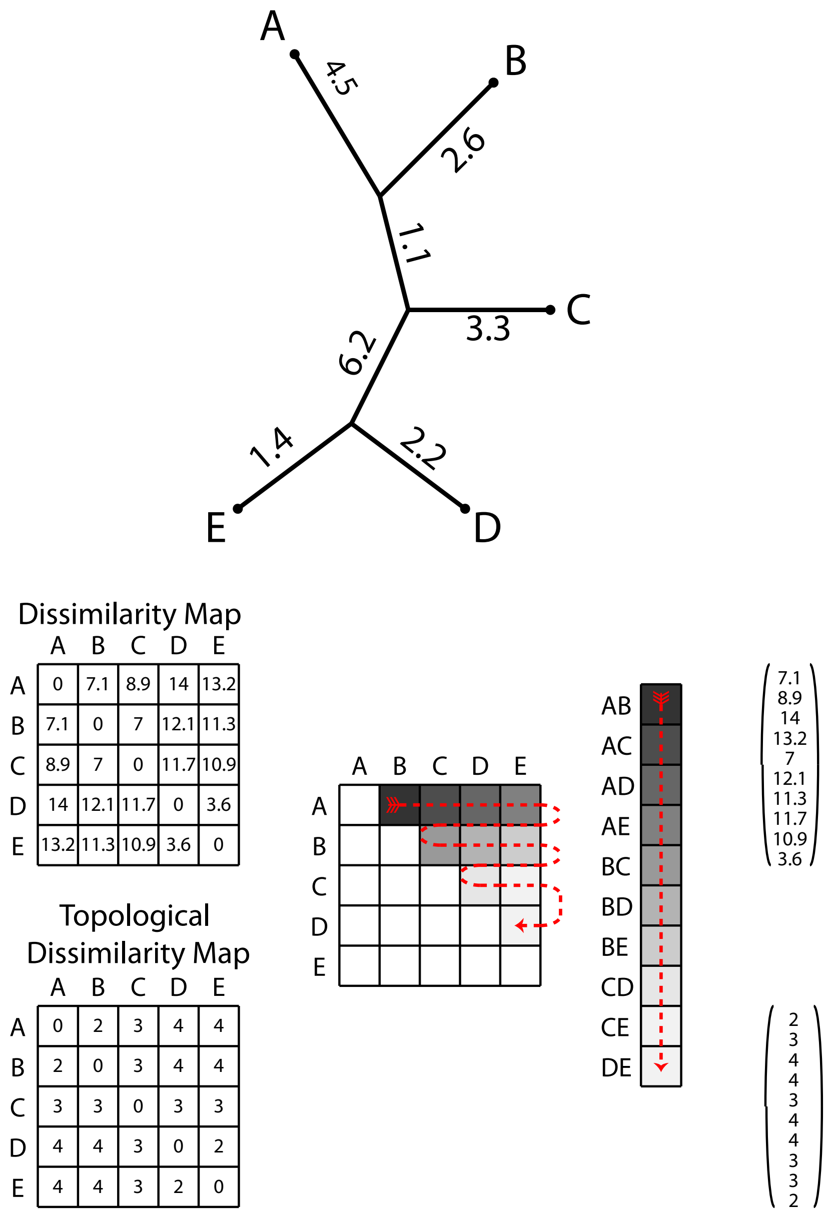


Figure B.1: Monte Carlo estimates of $\sum_{T \in \mathcal{T}} k(T, T')$ are plotted against the unnormalized tree score for each tree T' in the Apicomplexa data. There is no significant evidence that the sum is related to the tree score ($p = 0.72$).



Dissimilarity Map

	A	B	C	D	E
A	0	7.1	8.9	14	13.2
B	7.1	0	7	12.1	11.3
C	8.9	7	0	11.7	10.9
D	14	12.1	11.7	0	3.6
E	13.2	11.3	10.9	3.6	0

Topological Dissimilarity Map

	A	B	C	D	E
A	0	2	3	4	4
B	2	0	3	4	4
C	3	3	0	3	3
D	4	4	3	0	2
E	4	4	3	2	0

Figure B.2: Schematic of how trees are converted to vectors. Numbers on branches in the unrooted tree are branch lengths. In this example, the tree is first converted to either a branch length-based dissimilarity map (matrix of distances between tips) or topological dissimilarity maps (matrix of number of edges between tips). Moving from left to right across rows in one half of a matrix, values are placed into a single column to yield a vector of distances between tips in the tree.

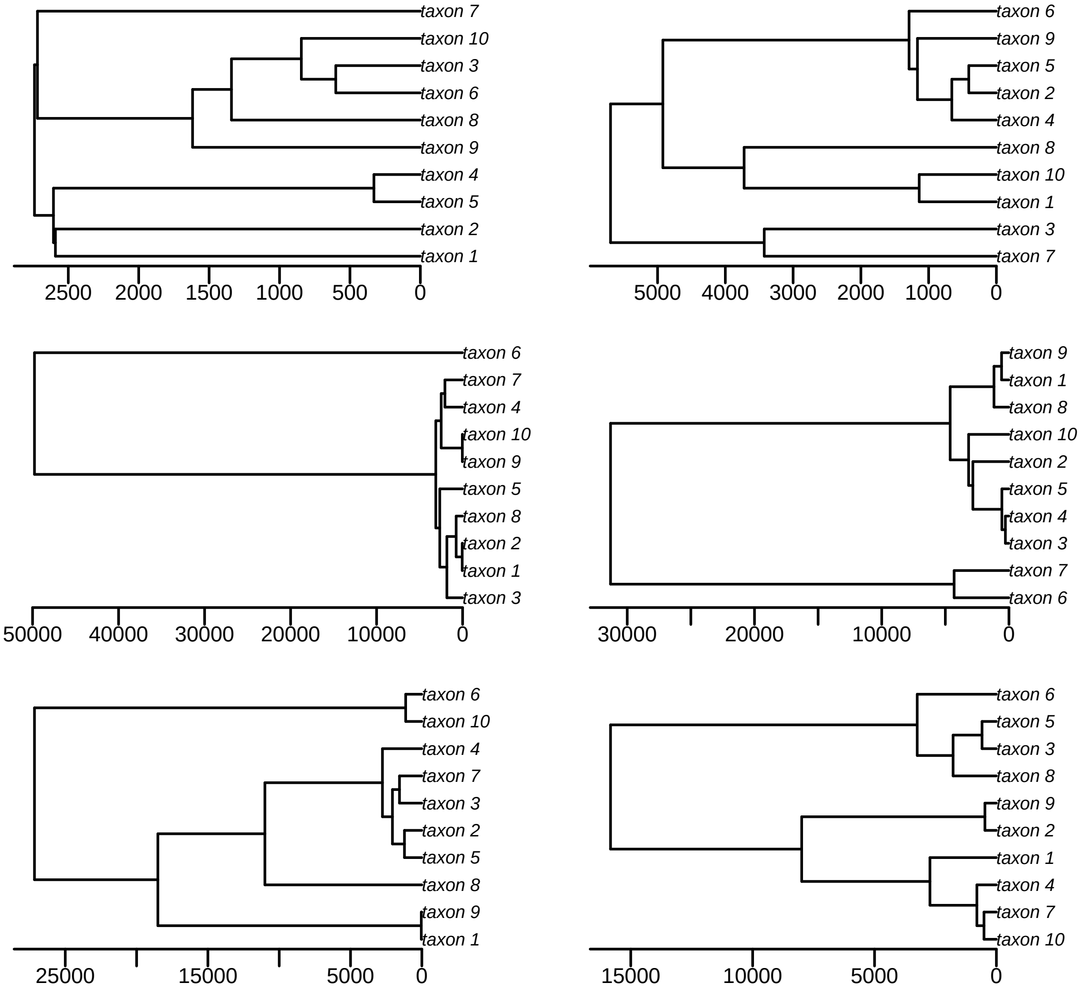


Figure B.3: The species trees used to generate gene trees under the coalescent model for the simulation experiments. At top-left is the tree used for the "single" coalescent distribution simulations, while the other trees are used in the "mixed" simulations.

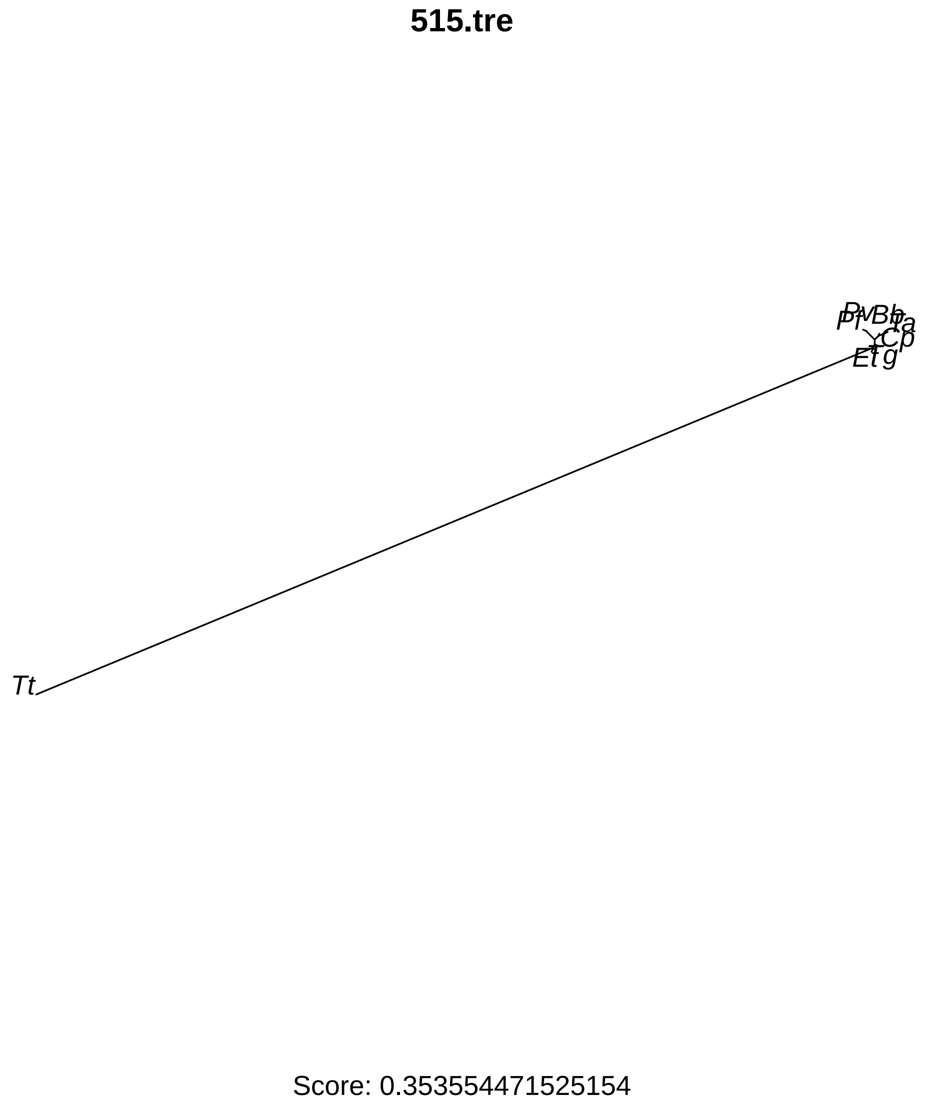


Figure B.4: Plot of the first Apicomplexa gene tree identified as an outlier. The extremely long branches lead to the identification as an outlier, and are likely the result of incorrect annotations of the original sequences.

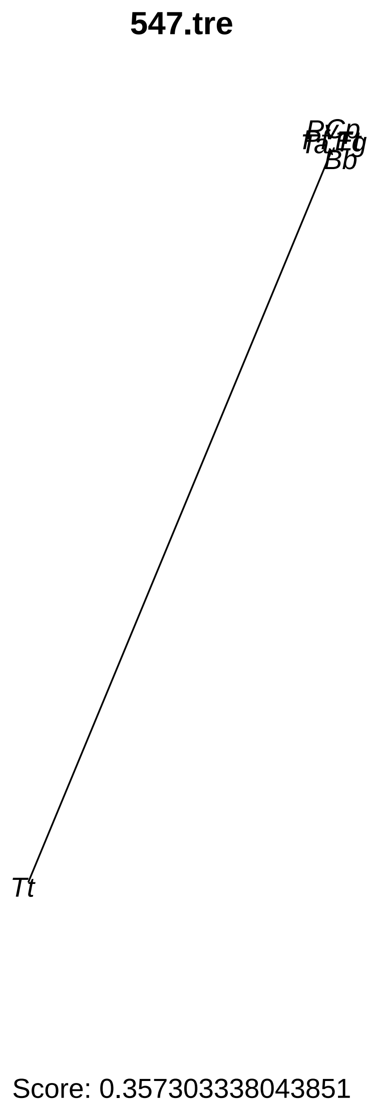


Figure B.5: Plot of the second Apicomplexa gene tree identified as an outlier.

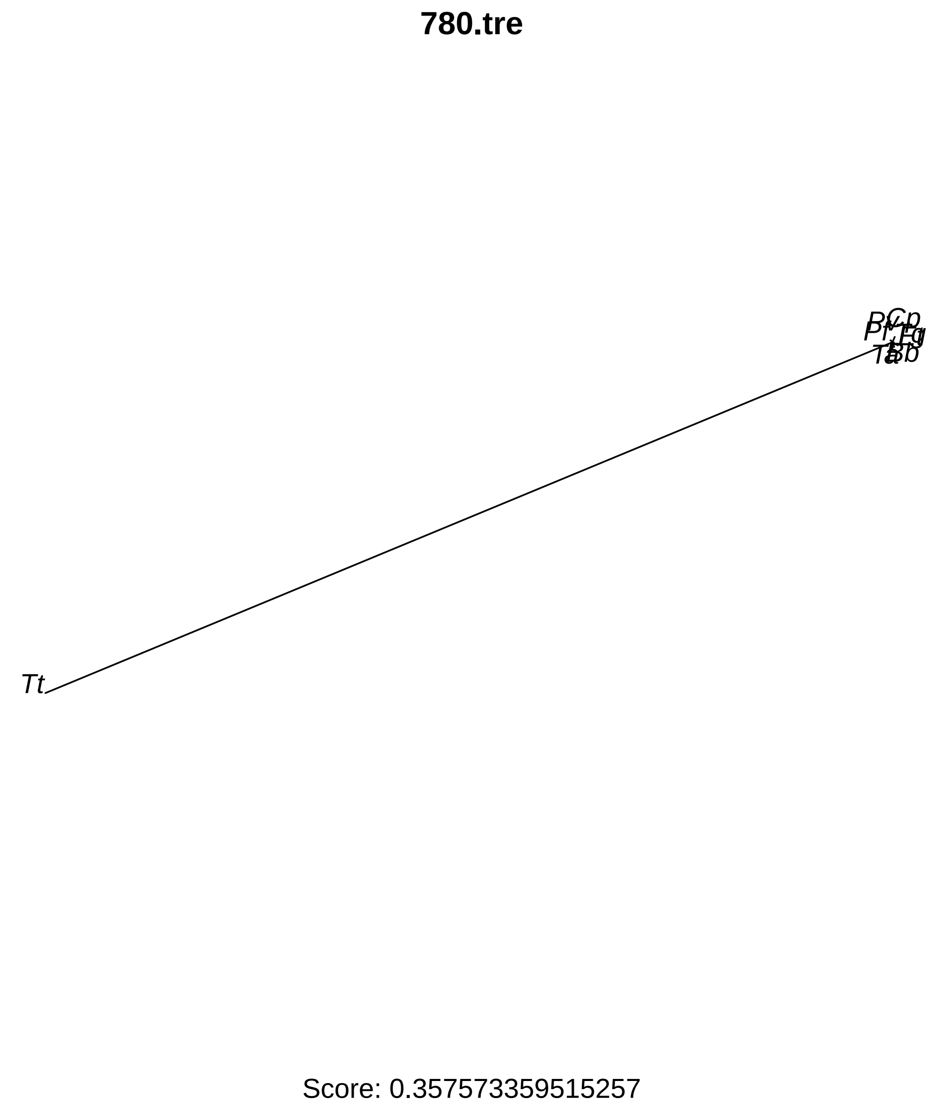
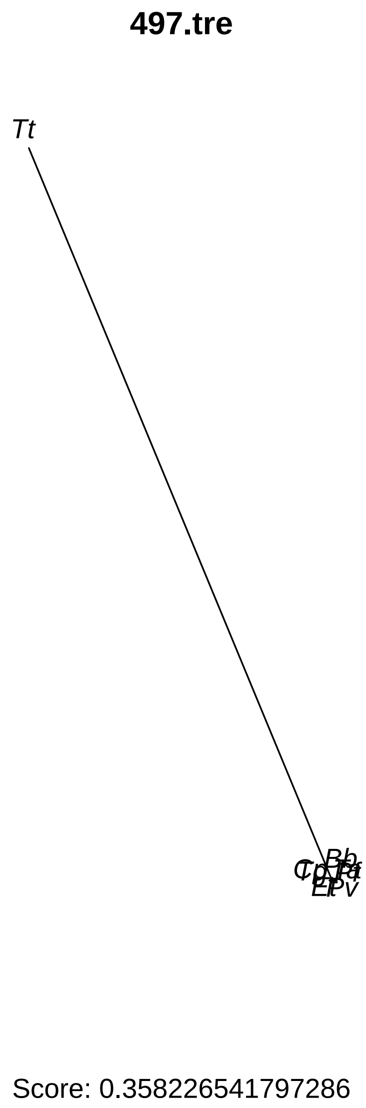


Figure B.6: Plot of the third Apicomplexa gene tree identified as an outlier.



497.tre

Figure B.7: Plot of the fourth Apicomplexa gene tree identified as an outlier.

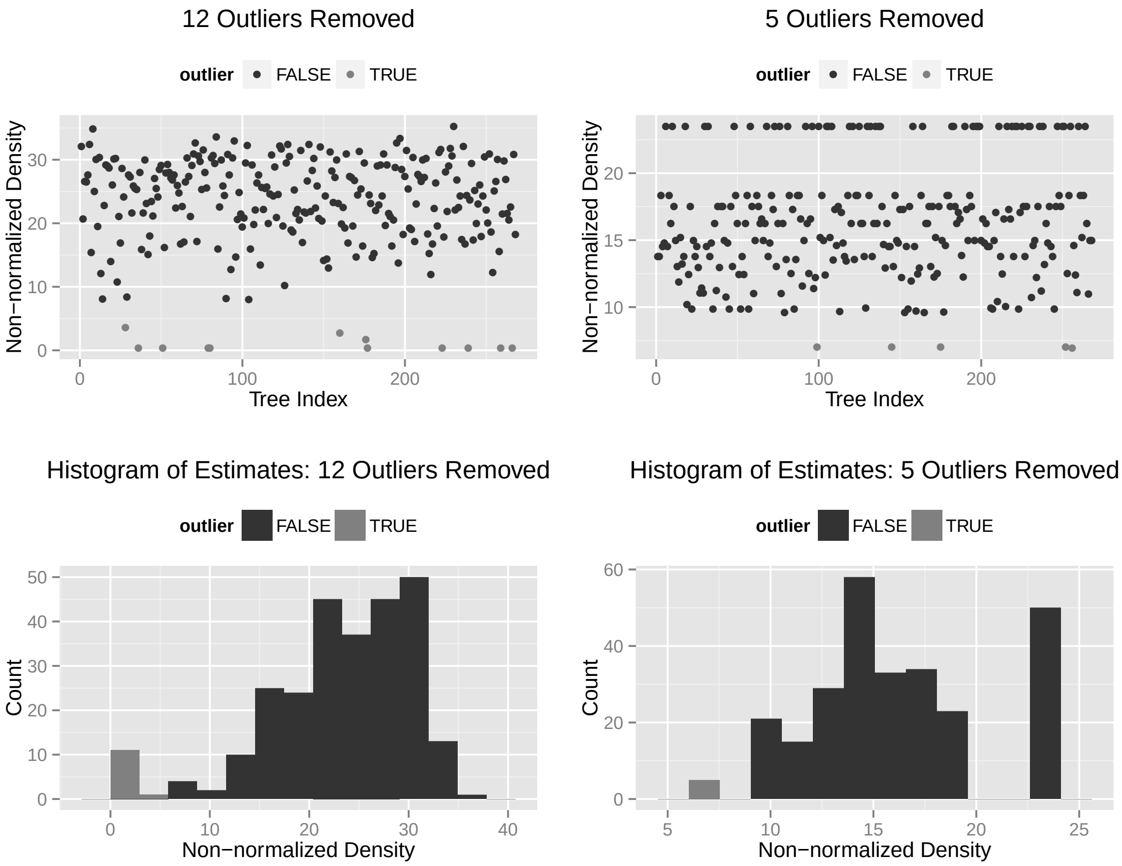


Figure B.8: Summary of tree scores for the Apicomplexa data set. In the top row the scores of individual trees are shown. "Tree Index" refers to the ordering of the trees in the input files. In the bottom row, the scores are summarized as a histogram. In the left column are the results computed with branch-length information, while the topology-only results are shown at right.

Appendix C

Supplementary Material for Chapter 3

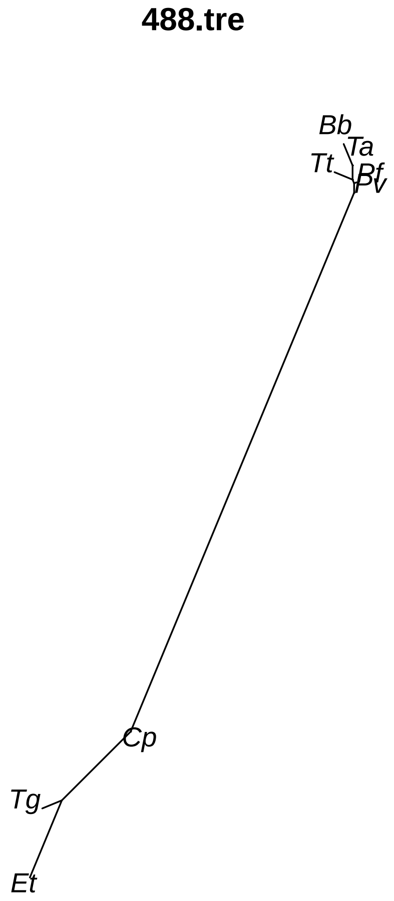


Figure C.1: A newly identified outlier from the Apicomplexa dataset.

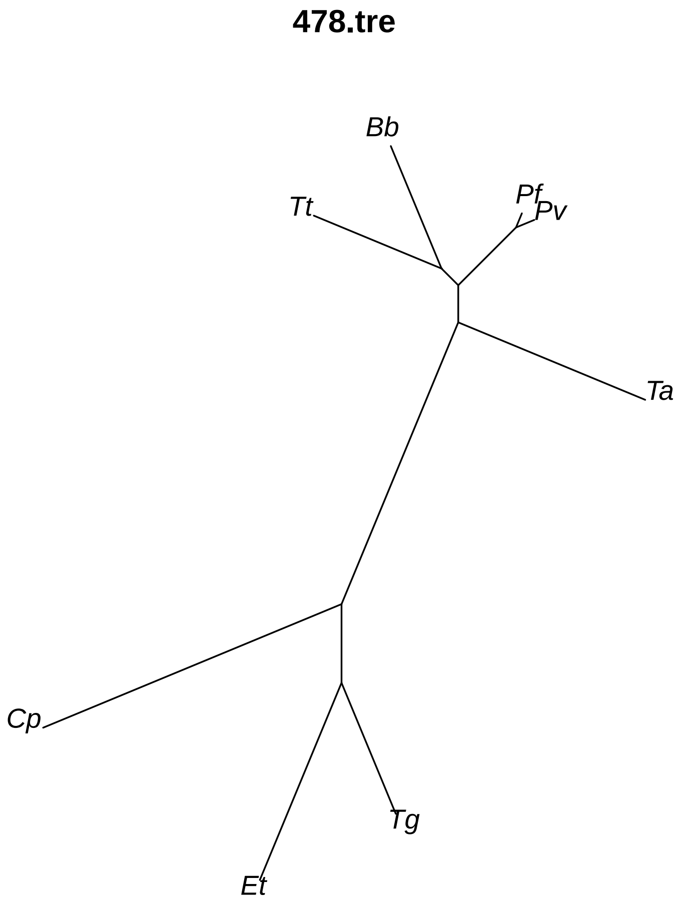


Figure C.2: A newly identified outlier from the Apicomplexa dataset.

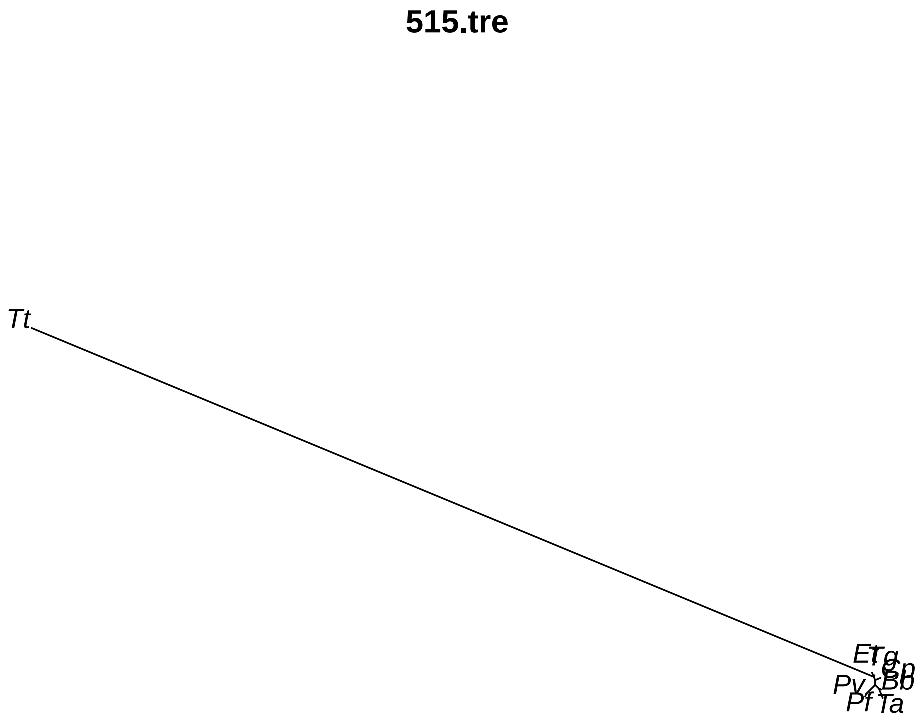


Figure C.3: A newly identified outlier from the Apicomplexa dataset.

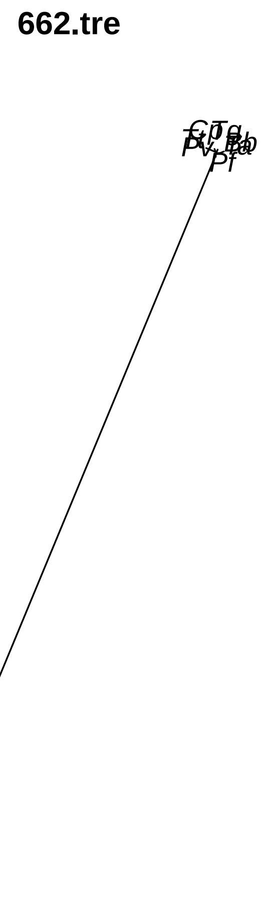


Figure C.4: A newly identified outlier from the Apicomplexa dataset.

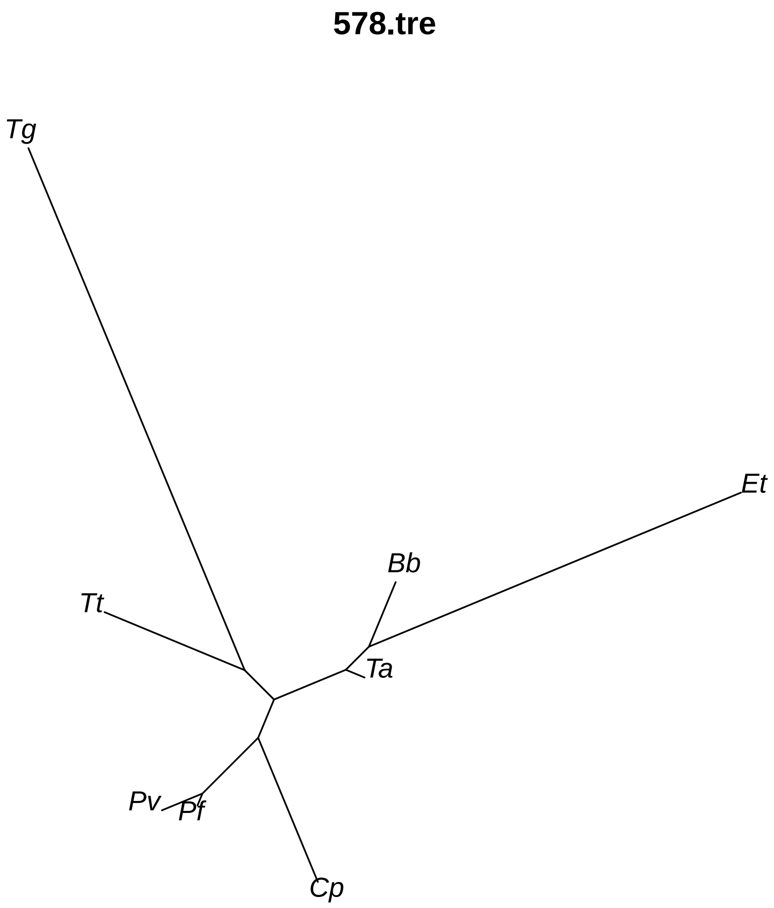


Figure C.5: A newly identified outlier from the Apicomplexa dataset.

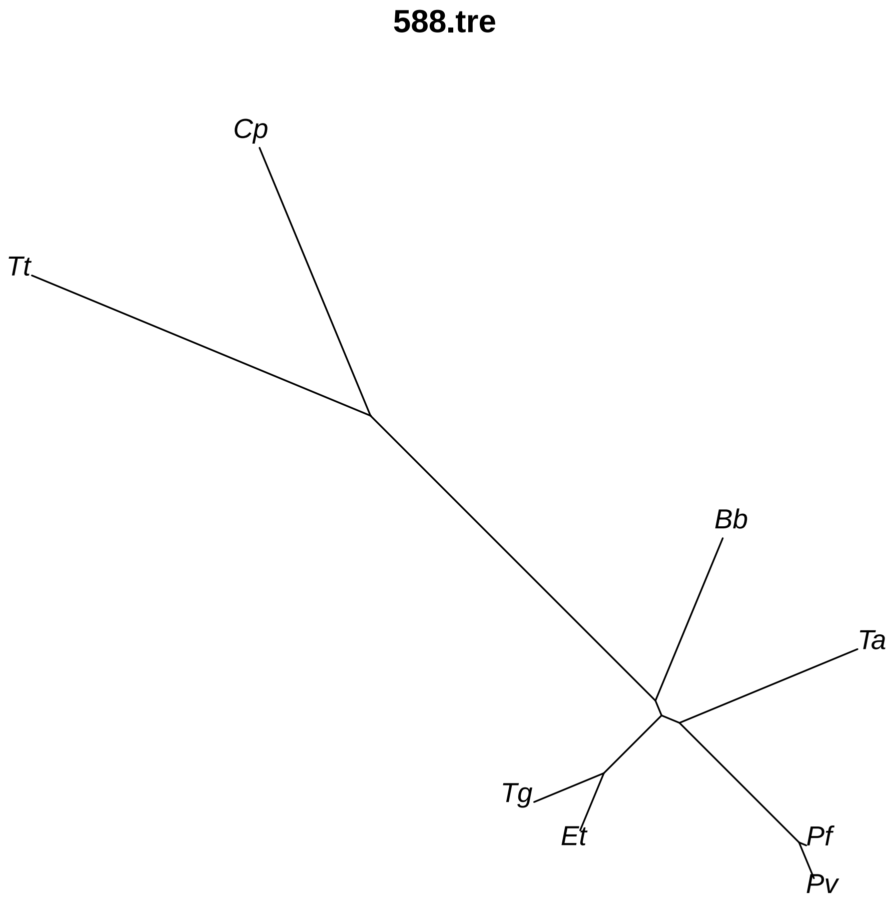


Figure C.6: A newly identified outlier from the Apicomplexa dataset.

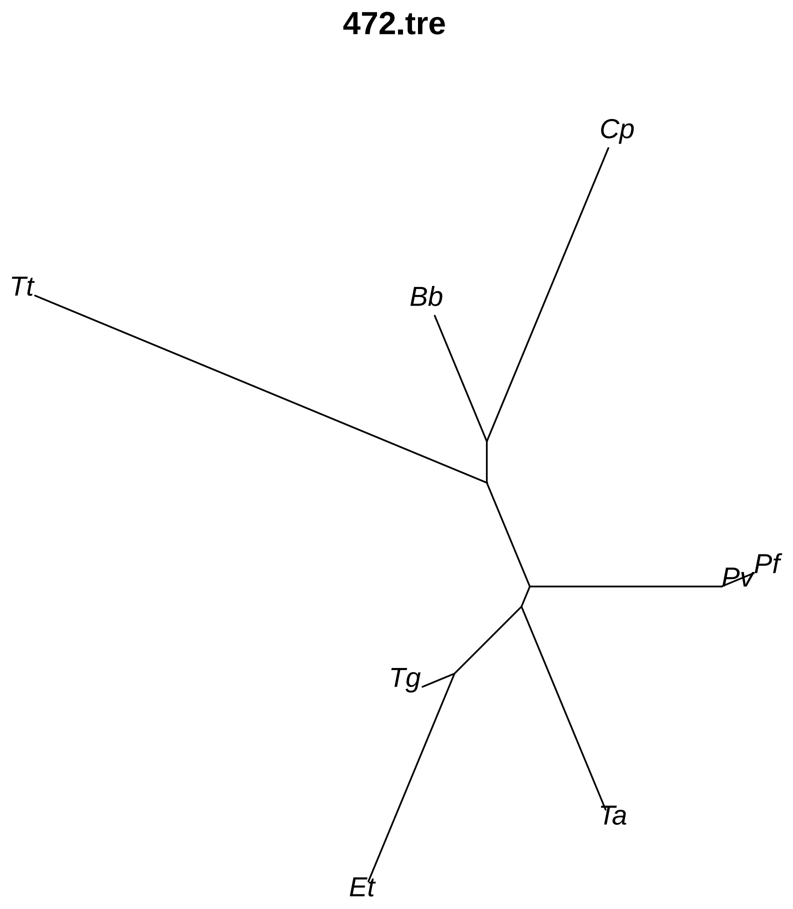


Figure C.7: A newly identified outlier from the Apicomplexa dataset.

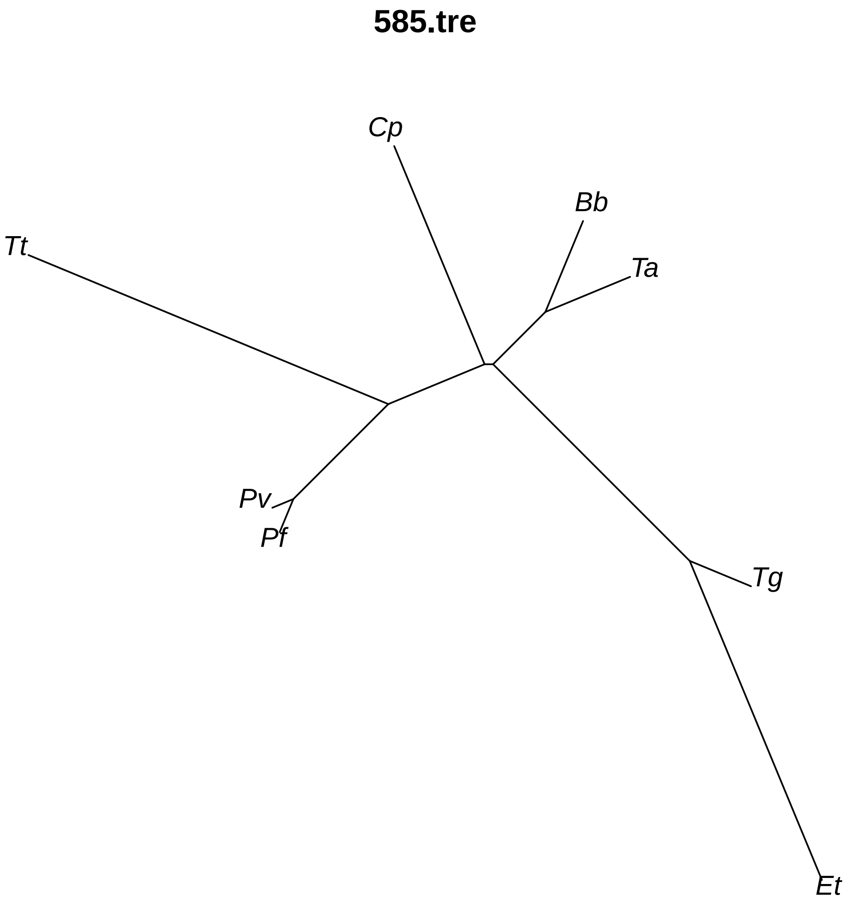


Figure C.8: A newly identified outlier from the Apicomplexa dataset.

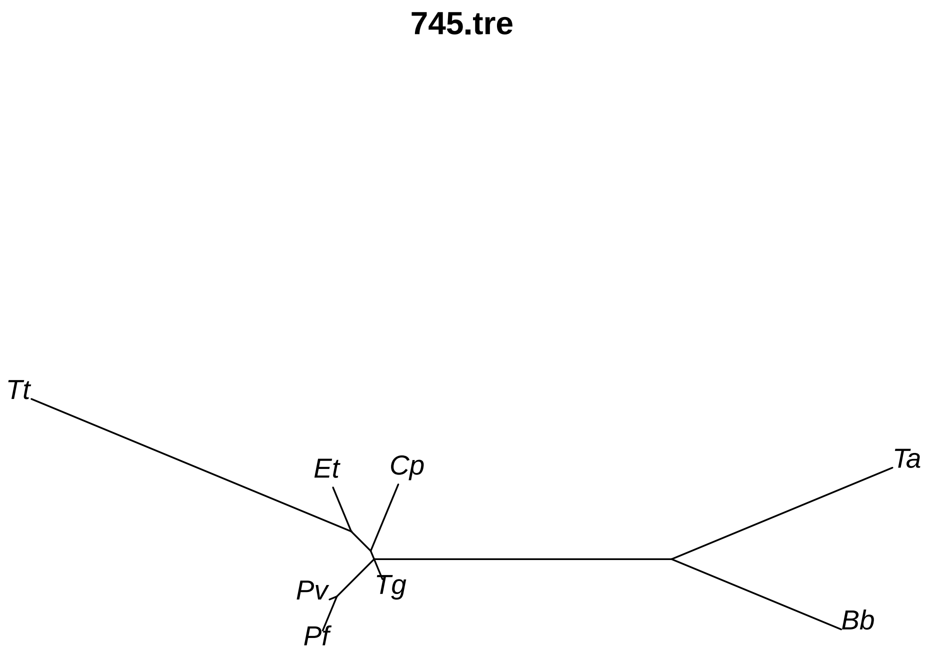


Figure C.9: A newly identified outlier from the Apicomplexa dataset.

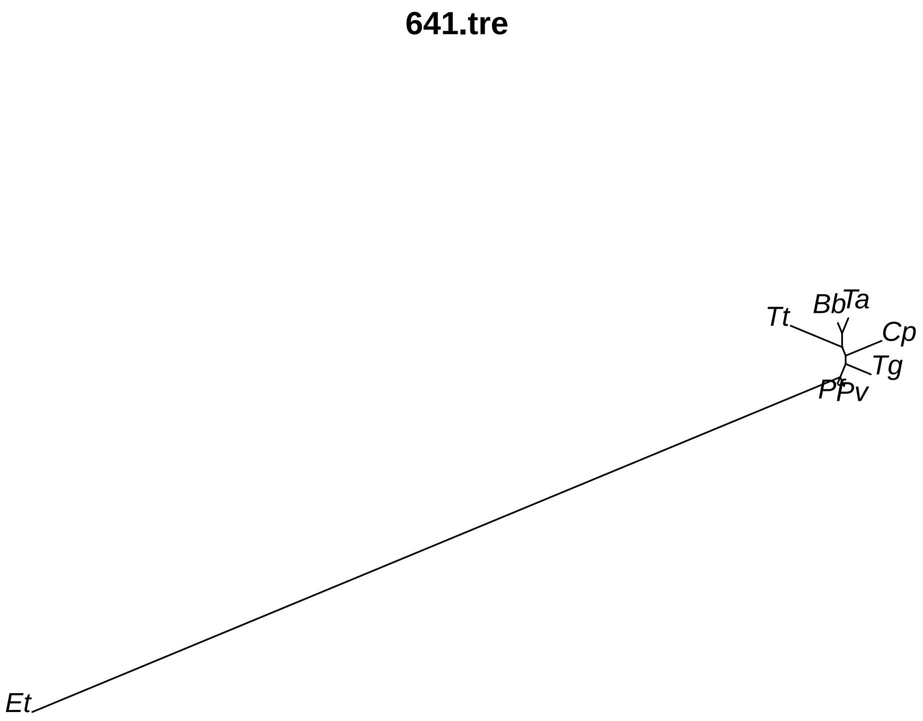


Figure C.10: A newly identified outlier from the Apicomplexa dataset.

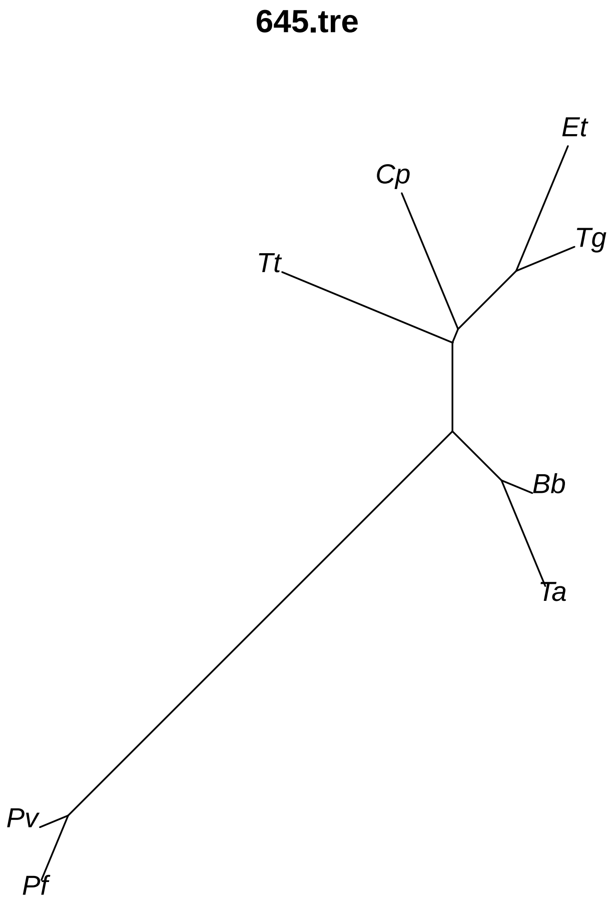


Figure C.11: A newly identified outlier from the Apicomplexa dataset.

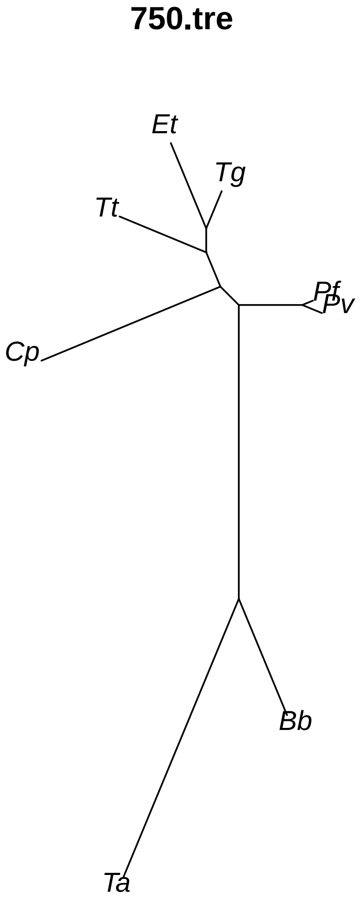


Figure C.12: A newly identified outlier from the Apicomplexa dataset.

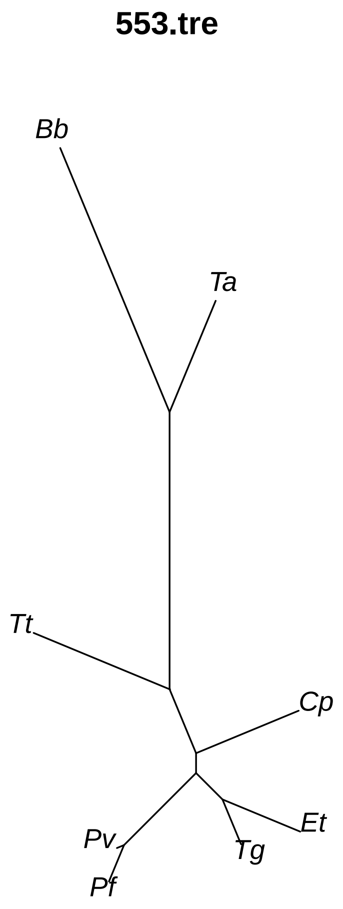


Figure C.13: A newly identified outlier from the Apicomplexa dataset.

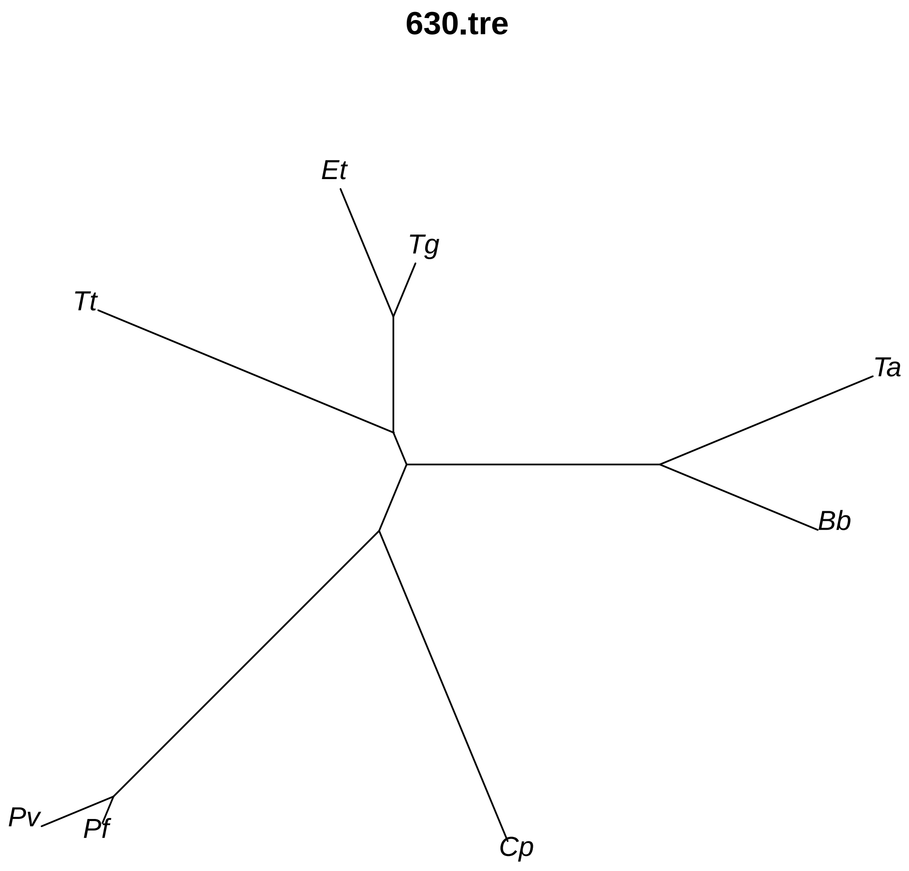


Figure C.14: A newly identified outlier from the Apicomplexa dataset.

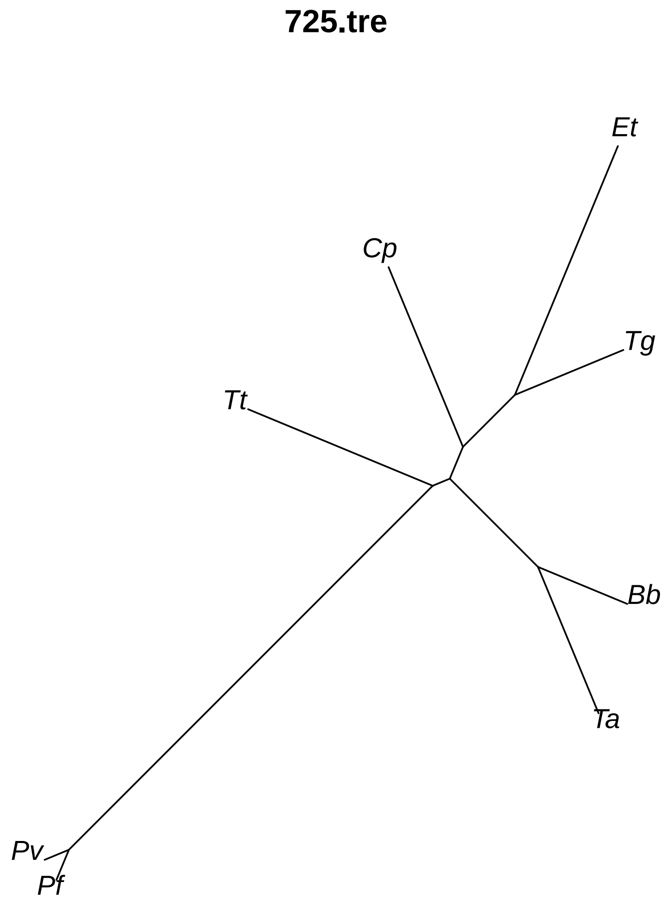


Figure C.15: A newly identified outlier from the Apicomplexa dataset.

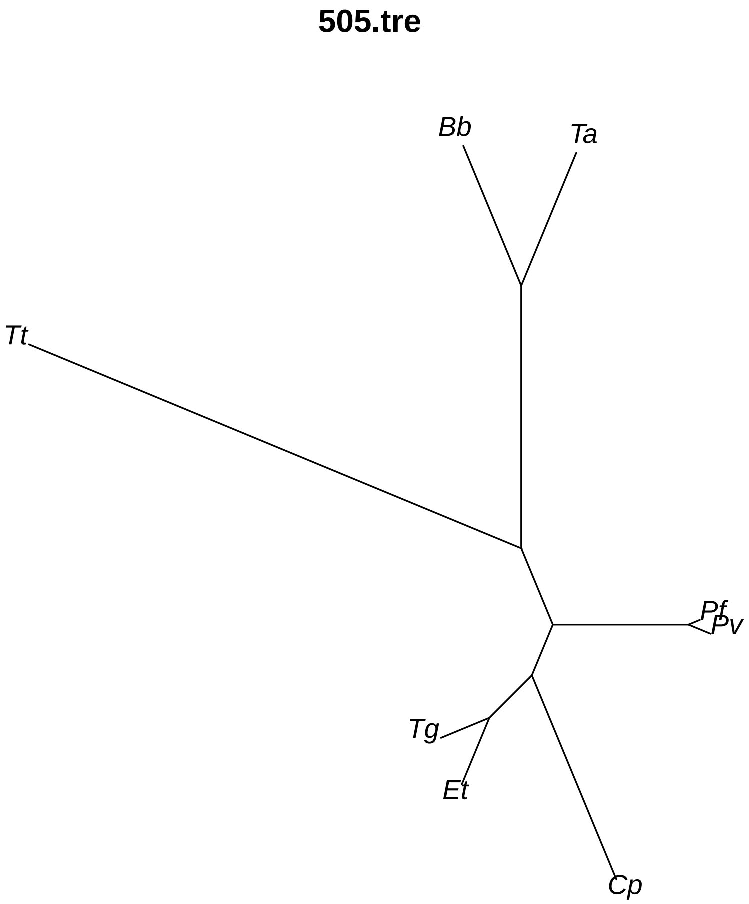


Figure C.16: A newly identified outlier from the Apicomplexa dataset.

Bibliography

[1] Mitchell S. Abrahamsen, Thomas J. Templeton, Shinichiro Enomoto, Juan E. Abrahante, Guan Zhu, Cheryl A. Lancto, Mingqi Deng, Chang Liu, Giovanni Widmer, Saul Tzipori, Gregory A. Buck, Ping Xu, Alan T. Bankier, Paul H. Dear, Bernard A. Konfortov, Helen F. Spriggs, Lakshminarayan Iyer, Vivek Anantharaman, L. Aravind, and Vivek Kapur. Complete genome sequence of the apicomplexan, cryptosporidium parvum. *Science*, 304:441–445, 2004.

[2] A.D.Leaché and B. Rannala. The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology*, 60(2):126–137, 2011.

[3] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

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

[5] C. Ane, B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, 24:412–426, 2007.

[6] C. Ané, B. Larget, D.A. Baum, S.D. Smith, and A. Rokas. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, 24(2):412–426, 2007.

[7] John M Archibald and Andrew J Roger. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *Journal of molecular biology*, 316(5):1041–1050, 2002.

[8] Amit Bahl, Brian Brunk, Jonathan Crabtree, Martin J. Fraunholz, Bindu Gajria, Gregory R. Grant, Hagai Ginsburg, Dinesh Gupta, Jessica C. Kissinger, Philip Labo, Li Li, Matthew D. Mailman, Arthur J. Milgram, David S. Pearson, David S. Roos, Jonathan Schug, Christian J. Stoeckert, and Patricia Whetzel. Plasmodb: the plasmodium genome resource. a database integrating experimental and computational data. *Nucleic Acids Res.*, 31:212–215, 2003.

[9] Miroslav Bačák. Computing medians and means in hadamard spaces. *SIAM Journal on Optimization*, 24(3):1542–1566, 2014.

[10] DA Benson, I Karsch-Mizrachi, DJ Lipman, J Ostell, and DL Wheeler. Genbank. *Nucleic Acids Res*, 36:D25–30, 2008.

[11] J. Bergsten. A review of long-branch attraction. *Cladistics*, 21:163–193, 2005.

[12] R. Betancur, C. Li, T.A. Munroe, J.A. Ballesteros, and G. Ortí. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (teleostei: Pleuronectiformes). *Systematic Biology*, page doi:10.1093/sysbio/syt039, 2013.

[13] L.J. Billera, S.P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv Appl Math*, 27(4):733–767, 2001. ISSN 0196-8858.

[14] Damian Bogdanowicz and Krzysztof Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 99 (PrePrints), 2011. ISSN 1545-5963. doi: http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.38.

[15] J.P. Bollback and J.P. Huelsenbeck. Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence. *Genetics*, 181(1):225–234, 2009.

[16] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[17] Kelly A. Brayton, Audrey O. T. Lau, David R. Herndon, Linda Hannick, Lowell S. Kappmeyer, Shawn J. Berens, Shelby L. Bidwell, Wendy C. Brown, Jonathan Crabtree, Doug Fadrosh, Tamara Feldblum, Heather A. Forberger, Brian J. Haas, Jeanne M. Howell, Hoda Khouri, Hean Koo, David J. Mann, Junzo Norimine, Ian T. Paulsen, Diana Radune, Qinghu Ren, Roger K. Smith Jr., Carlos E. Suarez, Owen White, Jennifer R. Wortman, Donald P. Knowles Jr.1, Terry F. McElwain, and Vishvanath M. Nene. Genome sequence of babesia bovis and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog*, 3:e148, 2007.

[18] Henner Brinkmann, Mark Van der Giezen, Yan Zhou, Gaëtan Poncelin De Raucourt, and Hervé Philippe. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic biology*, 54(5):743–757, 2005.

[19] P. Brito and S. Edwards. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, 135:439–455, 2009.

[20] D. R. Brooks. Parsimony analysis in historical biogeography and coevolution: methodological and theoretical update. *Syst. Zool.*, 39:14–30, 1990.

[21] D. R. Brooks and D. A. McLennan. *Phylogeny, Ecology and Behavior: A Research Program in Comparative Biology.* Univ. of Chicago Press, Chicago, 1991.

[22] D. R. Brooks and D. A. McLennan. *Parascript: Parasites and the Language of Evolution.* Smithsonian Institution Press, Washington, DC., 1993.

[23] D. R. Brooks and D. A. McLennan. *The Nature of Diversity: An Evolutionary Voyage of Discovery.* Univ. of Chicago Press, Chicago, 2002.

[24] D. R. Brooks, M. G. P. Van Veller, and D. A. McLennan. How to do BPA, really. *J. Biogeogr*, 28:343–358, 2001.

[25] P. Buneman. The recovery of trees from measures of similarity. In FR Hodson, DG Kendall, and P Tautu, editors, *Mathematics of the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, Edinburgh, 1971.

[26] Kenneth P Burnham and David R Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[27] F. Camastra and A. Verri. A novel kernel method for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):801–804, 2005.

[28] M. Carling and R. Brumfield. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in passerina buntings. *Genetics*, 178:363–377, 2008.

[29] B. C. Carstens and L. L. Knowles. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers. *Syst Biol*, 56: 400–411, 2007.

[30] T.A. Castoe, A.P.J. de Koning, H.M. Kim, W. Gu, B.P. Noonan, G. Naylor, C.L. Jiang, Z.J.and Parkinson, and D.D. Pollock. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 106:8986–8991, 2009.

[31] John Chakerian and Susan Holmes. *distory: Distance Between Phylogenetic Histories*, 2013. URL `http://CRAN.R-project.org/package=distory`. R package version 1.4.1.

[32] J.H. Cuthill and M. Charleston. Phylogenetic codivergence supports coevolution of mimetic heliconius butterflies. *PloS ONE*, 7(5):e36464. doi:10.1371/journal.pone.0036464, 2012.

[33] Charles Darwin. On the origins of species by means of natural selection. *London: Murray*, 1859.

[34] Charles Darwin. 1836-1844: geology, transmutation of species, metaphysical enquiries. In P H Barrett, editor, *Charles Darwin's Notebooks*. Cambridge University Press, 1987.

[35] Bhaskar Dasgupta, Xin He, Tao Jiang, Ming Li, John Tromp, and Louxin Zhang. On computing the nearest neighbor interchange distance. In *In: Proc. DIMACS Workshop on Discrete Problems with Medical Applications*, pages 125–143. Press, 1997.

[36] William HE Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of mathematical biology*, 49(4):461–467, 1987.

[37] D. M. de Vienne, S. Ollier, and G. Aguileta. Phylo-mcoa: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol Biol Evol*, 2012.

[38] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

[39] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9:687–705, 2002.

[40] T. R. Disotell and R. L. Raaum. Molecular timescale and gene tree incongruence in the guenons. *The Guenons: Diversity and Adaptation in African Monkeys Developments in Primatology: Progress and Prospects*, pages 27–36, 2004.

[41] A. P. G. Dowling. Testing the accuracy of treemap and brooks parsimony analyses of coevolutionary patterns using artificial associations. *Cladistics*, 18:416–435, 2002.

[42] A. P. G. Dowling, M. G. P. van Veller, E. P. Hoberg, and D. R. Brooks. A priori and a posteriori methods in comparative evolutionary studies of host-parasite associations. *Cladistics*, 19:240–253, 2003.

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

[44] R.C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

[45] S. Edwards. Is a new and general theory of molecular systematics emerging? *Evolution*, 63: 1–19, 2009.

[46] S. Edwards, L. Liu, and D. Pearl. High-resolution species trees without concatenation. *Proc Natl Acad Sci USA*, 104:5936–5941, 2007.

[47] Jonathan A Eisen, Robert S Coyne, Martin Wu, Dongying Wu, Mathangi Thiagarajan, Jennifer R Wortman, Jonathan H Badger, Qinghu Ren, Paolo Amedeo, Kristie M Jones, Luke J Tallon, Arthur L Delcher, Steven L Salzberg, Joana C Silva, Brian J Haas, William H Majoros, Maryam Farzad, Jane M Carlton, Roger K Smith, Jr., Jyoti Garg, Ronald E Pearlman, Kathleen M Karrer, Lei Sun, Gerard Manning, Nels C Elde, Aaron P Turkewitz, David J Asai, David E Wilkes, Yufeng Wang, Hong Cai, Kathleen Collins, B. Andrew Stewart, Suzanne R Lee, Katarzyna Wilamowska, Zasha Weinberg, Walter L Ruzzo, Dorota Wloga, Jacek Gaertig, Joseph Frankel, Che-Chia Tsao, Martin A Gorovsky, Patrick J Keeling, Ross F Waller, Nicola J Patron, J. Michael Cherry, Nicholas A Stover, Cynthia J Krieger, Christina del Toro, Hilary F Ryder, Sondra C Williamson, Rebecca A Barbeau, Eileen P Hamilton, and Eduardo Orias. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, 4:1620–1642, 2006.

[48] G. Estabrook, F. McMorris, and C. Meaeham. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Zool*, 34(2):193–200, 1985.

[49] M. A. Fares, K. P. Byrne, and K. H. Wolfe. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of saccharomyces species. *Mol Biol Evol*, 23(2):245–253, 2006.

[50] J Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Inc., 2003.

[51] Aasa Feragen, Megan Owen, Jens Petersen, Mathilde MW Wille, Laura H Thomsen, Asger Dirksen, and Marleen de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *Information Processing in Medical Imaging*, pages 74–85. Springer, 2013.

[52] M. Fischer and S. Kelk. On the maximum parsimony distance between phylogenetic trees, 2014.

[53] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995–1005, 2004.

[54] Bindu Gajria, Amit Bahl, John Brestelli, Jennifer Dommer, Steve Fischer, Xin Gao, Mark Heiges, John Iodice, Jessica C. Kissinger, Aaron J. Mackey, Deborah F. Pinney, David S. Roos, Christian J. Stoeckert, Haiming Wang, and Brian P. Brunk. Toxodb: an integrated toxoplasma gondii database resource. *Nucleic Acids Res.*, 36:D553–D556, 2008.

[55] Malcolm J. Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W. Hyman, Jane M. Carlton, Arnab Pain, Karen E. Nelson, Sharen Bowman, Ian T. Paulsen, Keith James, Jonathan A. Eisen, Kim Rutherford, Steven L. Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J. Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W. Mather, Akhil B. Vaidya, David M. A. Martin, Alan H. Fairlamb, Martin J. Fraunholz, David S. Roos, Stuart A. Ralph, Geoffrey I. McFadden, Leda M. Cummings, G. Mani Subramanian, Chris Mungall, J. Craig Venter, Daniel J. Carucci, Stephen L. Hoffman, Chris Newbold, Ronald W. Davis, Claire M. Fraser, and Bart Barrell. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419:498–511, 2002.

[56] O Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 1997. URL `http://mbe.oxfordjournals.org/content/14/7/685.abstract`.

[57] S. Ge, T. Sang, B. Lu, and D. Hong. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *PNAS*, 96(25):14400–14405, December 7, 1999.

[58] Pablo A Goloboff. Calculating spr distances between trees. *Cladistics*, 24(4):591–597, 2008.

[59] Alexander N Gorban and Andrei Y Zinovyev. Principal graphs and manifolds. *Ch*, 2:28–59, 2009.

[60] M. Graham and J. Kennedy. A survey of multiple tree visualisation. *Inf Vis*, 9:235–252, 2010.

[61] M. S. Hafner and S. A. Nadler. Cospeciation in host parasite assemblages: comparative analysis of rates of evolution and timing of cospeciation events. *Systematic Zoology*, 39:192–204, 1990.

[62] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer-Verlag, 2nd ed edition, 2009.

[63] D. Haws, P. Huggins, E. M. O'Neill, D. W. Weisrock, and R. Yoshida. A support vector machine based test for incongruence between sets of trees in tree space. *BMC Bioinformatics*, 13(210), 2012.

[64] D. Haws, T. Hodge, and R. Yoshida. Phylogenetic tree reconstruction: Geometric approaches. In Raina Robeva and Terrell Hodge, editors, *Mathematical concepts and methods in modern biology: using modern discrete models*. Academic Press, 2013.

[65] Jumpei Hayakawa and Akimichi Takemura. Estimation of exponential-polynomial distribution by holonomic gradient descent, 2014.

[66] Mark Heiges, Haiming Wang, Edward Robinson, Cristina Aurrecoechea, Xin Gao, Nivedita Kaluskar, Philippa Rhodes, Sammy Wang, Cong-Zhou He, Yanqi Su, John Miller, Eileen Kraemer, and Jessica C. Kissinger. Cryptodb: a cryptosporidium bioinformatics resource update. *Nucleic Acids Res.*, 34:D419–D422, 2006.

[67] J. Hein, M.H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005.

[68] J. Heled and A.J. Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580, 2011.

[69] W. Henning. *Phylogenetic Systematics*. Univ. of Illinois Press, Urbana, 1966.

[70] Christiane Hertz-Fowler, Chris S. Peacock, Valerie Wood, Martin Aslett, Arnaud Kerhornou, Paul Mooney, Adrian Tivey, Matthew Berriman, Neil Hall, Kim Rutherford, Julian Parkhill, Alasdair C. Ivens, Marie-Adele Rajandream, and Bart Barrell. Genedb: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, 32:D339–D343, 2004.

[71] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin. Spr distance computation for unrooted trees. *Evolutionary bioinformatics online*, 4:17–27, 2008.

[72] David M. Hillis, Tracy A. Heath, and Katherin St. John. Analysis and visualization of tree space. *Syst Biol*, 54(3):471–482, 2005.

[73] S. Holmes. Statistical approach to tests involving phylogenies. In O. Gascuel, editor, *Mathematics of Phylogeny and Evolution*, chapter 4, pages 91–117. Oxford University Press, New York, 2005.

[74] D. S. Horner and G. PesoleâĂă. Phylogenetic analyses: a brief introduction to methods and their application. *Expert Rev Mol Diagn*, 4(3):339–350, 2004.

[75] R. Hovmoller, L. L. Knowles, and L. S. Kubatko. Effects of missing data on species tree estimation under the coalescent. *Molecular Phylogenetics and Evolution*, 69:1057–1062, 2013.

[76] J. P. Huelsenbeck, B. Rannala, and B. Larget. A bayesian framework for the analysis of cospeciation. *Evolution*, 54(2):352–364, 2000.

[77] J. P. Huelsenbeck, B. Rannala, and Z. Yang. Statistical tests of host-parasite cospeciation. *Evolution*, 51(2), Apr. 1997.

[78] J.P. Huelsenbeck and F. Ronquist. Mrbayes: Bayesian inference in phylogenetic trees. *Bioinformatics*, 17:754—755, 2001.

[79] A. B. Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. A support vector method for clustering. *NIPS*, pages 367–373, 2000.

[80] A. B. Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *JMLR*, 2: 125–137, 2001.

[81] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[82] R. Jothi, E. Zotenko, A. Tasneem, and T.M. Przytycka. Coco-cl: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, 22:779–788, 2006. doi: 10.1093/bioinformatics/btl009.

[83] TH Jukes and C Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press, 1969.

[84] Winfried Just. Computational complexity of multiple sequence alignment with sp-score. *Journal of computational biology*, 8(6):615–623, 2001.

[85] Standley Katoh. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, 2013.

[86] P Keese. Risks from gmos due to horizontal gene transfer. *Environmental Biosafety Research*, 7:123–149, 2008.

[87] J.F.C Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[88] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data. *J Mol Evol*, 29:170–179, 1989.

[89] S Kishore, J Stiller, and K Deitsch. Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite plasmodium falciparum and other apicomplexans. *BMC Evolutionary Biology*, 13(37):doi 10.1186/1471–2148–13–37, 2013.

[90] S. P. Kishore, J. W. Stiller, and K. W. Deitsch. Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite plasmodium falciparum and other apicomplexans. *Evol Biol*, pages 13–37, 2013.

[91] L.L. Knowles. Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40:593–612, 2009.

[92] L.L. Knowles. Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology*, 58(5):463–467, 2009.

[93] L.L. Knowles and P.B. Klimov. Estimating phylogenetic relationships despite discordant gene trees across loci: the species tree of a diverse species group of feather mites (acari: Proctophyllodidae). *Parasitology*, 138(13):1750–1759, 2011.

[94] Tamio Koyama, Hiromasa Nakayama, Katsuyoshi Ohara, Tomonari Sei, and Nobuki Takayama. Software packages for holonomic gradient method. In *Mathematical Software–ICMS 2014*, pages 706–712. Springer, 2014.

[95] P. Kück, C. Mayer, J.-W. Wágele, and B. Misof. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS ONE*, page DOI: 10.1371/journal.pone.0036593, 2012.

[96] C. Kuo, J. P. Wares, and J. C. Kissinger. The apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees. *Mol Biol Evol*, 25(12):2689–2698, 2008.

[97] K Lange. *Numerical Analysis for Statisticians.* Springer, 2001.

[98] Gregory F Lawler. *Introduction to stochastic processes.* CRC Press, 2nd edition, 2006.

[99] D J Lawson and N M Adams. A general decision framework for structuring computation using data directional scaling to process massive similarity matrices. [preprint] `http://arxiv.org/abs/1403.4054v1`, 2014.

[100] M. S. Y. Lee and A. F. Hugall. Partitioned likelihood support and the evaluation of data set conflict. *Syst Biol*, 52(1):15–22, 2003.

[101] P. Legendre, Y. Desdevises, and E. Bazin. A statistical test for host-parasite coevolution. *Systematic Biology*, 51:217–234, 2002.

[102] ND Levine. Progress in taxonomy of the apicomplexan protozoa. *J Eukaryot Microbiol*, 35: 518–520, 1988.

[103] L. Li, C. J. Stoeckert, and D. S. Roos. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13:2178–2189, 2003.

[104] Dan Liang, Xing Xing Shen, and Peng Zhang. One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Molecular biology and evolution*, page mst072, 2013.

[105] J Lindsay and MG Holden. Understanding the rise of the superbug: investigation of the evolution and genomic variation of staphylococcus aureus. *Functional & Integrative Genomics*, 6:186–201, 2006.

[106] L. Liu and D. K. Pearl. Species trees from gene trees. *Syst. Biol.*, 2007. in press.

[107] L. Liu, D. Pearl, R. Brumfield, and S. Edwards. Estimating species trees using multiple-allele dna sequence data. *Evolution*, 62:2080–2091, 2008.

[108] L. Liua, L. Yub, L. Kubatkoc, D.K. Pearlc, and S.V. Edwards. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1):320–328, 2009.

[109] P. J. Lockhart, D. Penny, and Axel Meyer. Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. *J Mol Evol*, 41:666–674, 1995.

[110] W. P. Maddison. Gene trees in species trees. *Syst Biol*, 46(3):523–536, 1997.

[111] W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*, 55:21–30, 2006.

[112] A. P. Martin and T. M. Burg. Perils of paralogy: Using hsp70 genes for inferring organismal phylogenies. *Systematic Biology*, 51:570–587, 2002.

[113] Naoki Marumo, Toshinori Oaku, and Akimichi Takemura. Properties of powers of functions satisfying second-order linear differential equations with applications to statistics, 2014.

[114] J. Meloche. Asymptotic behaviour of the mean integrated squared error of kernel density estimators for dependent observations. *Can J Stat*, 18(3):205–211, 1990.

[115] E. Mossel and S. Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci, 2007. arXiv q-bio.PE.

[116] Elchanan Mossel and Sebastien Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 7(1):166–171, 2010.

[117] Hiromasa Nakayama, Kenta Nishiyama, Masayuki Noro, Katsuyoshi Ohara, Tomonari Sei, Nobuki Takayama, and Akimichi Takemura. Holonomic gradient descent and its application to the fisher-bingham integral. *Advances in Applied Mathematics*, 47(3):639 – 658, 2011. ISSN 0196-8858. doi: http://dx.doi.org/10.1016/j.aam.2011.03.001. URL `http://www.sciencedirect.com/science/article/pii/S019688581100008X`.

[118] T. Nepusz, R. Sasidharan, and A Paccanaro. Scps: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC Bioinformatics*, 11(120), 2010.

[119] Jerzy Neyman and Egon S Pearson. *On the problem of the most efficient tests of statistical hypotheses*. Springer, 1992.

[120] P Nordmann, T Naas, N Fortineau, and L Poirel. Superbugs in the coming new decade; multidrug resistance and prospects for treatment of staphylococcus aureus, enterococcus spp. and pseudomonas aeruginosa in 2010. *Current Opinion in Microbiology*, pages 436–440, 2007.

[121] Tom M. Nye. Trees of trees: An approach to comparing multiple alternative phylogenies. *Syst Biol*, 57(5):785–794, 2008.

[122] Tom M. Nye. Principal components analysis in the space of phylogenetic trees. *Ann Stat*, 39 (5):2716–2739, 2011.

[123] Tom MW Nye. An algorithm for constructing principal geodesics in phylogenetic treespace. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11(2):304–315, March 2014. ISSN 1545-5963. doi: 10.1109/TCBB.2014.2309599.

[124] M. Owen and J. S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE ACM T COMPUT BI*, 8(1):2–13, 2011.

[125] L. Pachter and B. Sturmfels. *Algebraic Statistics for Computational Biology.* Cambridge University Press, 2005. ISBN 9780521857000.

[126] R. Page. *Tangled trees.* The University of Chicago Press, 2003.

[127] R. D. M. Page. Component 2.0: Tree comparison software for Microsoft Windows. program and users manual, 1993.

[128] R.D.M. Page. Maps between trees and cladistic analysis of historical associations among genes,organisms, and areas. *Systematic Biology*, 43(1):58–77, 1994.

[129] R.D.M. Page. Treemap 1.0. program and users manual, 1995.

[130] R.D.M. Page and M.A. Charleston. Trees within trees: phylogeny and historical associations. *TREE*, 13(9), 1998.

[131] R.D.M. Page and E. C. Holmes. *Molecular Evolution : A Phylogenetic Approach.* Blackwell Publishing Ltd, 1998.

[132] A. Pain, H. Renauld, and et al. Genome of the host-cell transforming parasite theileria annulata compared with t. parva. *Science*, 309:131–133, 2005.

[133] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol Biol Evol*, 5: 568–583, 1988.

[134] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[135] R. Piaggio-Talice, J. G. Burleigh, and O. Eulenstein. Auqrtet supertrees. In O. R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 173—191. Kluwer Academic Publishers, Netherlands, 2004.

[136] Maria Poptsova. Testing phylogenetic methods to identify horizontal gene transfer. In Maria Boekels Gogarten, Johann Peter Gogarten, and Lorraine C. Olendzenski, editors, *Horizontal Gene Transfer*, volume 532 of *Methods in Molecular Biology*, pages 227–240. Humana Press, 2009. ISBN 978-1-60327-853-9.

[137] D. Posada and K.A. Crandall. The effect of recombination on the accuracy of phylogeny reconstruction. *Journal of Molecular Evolution*, 54:396–402, 2002.

[138] P. Puigbó, S. Garcia-Vallvé, and J. O. McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23(12):1556–2558, 2007.

[139] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[140] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, 13(3):235–238, 1997.

[141] TA Richards, DM Soanes, MDM Jones, O Vasieva, G Leonard, K Paszkiewicz, PG Foster, N Hall, and NJ Talbot. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proceedings of the National Academy of Sciences of the United States of America*, 108:15258–15263, 2011.

[142] M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA*, 95(11):6239–6244, 1998.

[143] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Math Biosci*, 53:131–147, 1981.

[144] D.F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory*, 11:105–119, 1971.

[145] A. Rokas and S.B. Carroll. Frequent and widespread parallel evolution of protein sequences. *Mol. Biol. Evol.*, 25:1943–1953, 2008.

[146] N. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol*, 61:225–247, 2002.

[147] N. A. Rosenberg. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, 57:1465–1477, 2003.

[148] A. RoyChoudhury, J. Felsenstein, and E. A. Thompson. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics*, 180:1095–1105, 2008.

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

[150] SL Salzberg, O White, J Peterson, and JA Eisen. Microbial genes in the human genome: Lateral transfer or gene loss? *Science*, 292:1903–1906, 2001.

[151] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, C-18(5):401–409, 1969.

[152] C. L. Schardl, K. D. Craven, A. Lindstrom, A. Stromberg, and R. Yoshida. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses, 2007. Preprint.

[153] C. L. Schardl, K. D. Craven, S. Speakman, A. Lindstrom, A. Stromberg, and R. Yoshida. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses. *Systematic Biology*, 57(3):483 – 498, 2008.

[154] C. L. Schardl, C. A. Young, U. Hesse, S. G. Amyotte, K. Andreeva, P. J. Calie, D. J. Fleetwood, D. C. Haws, N. Moore, B. Oeser, D. G. Panaccione, K. K. Schweri, C. R. Voisey, M. L. Farman, J. W. Jaromczyk, B. A. Roe, D. M. O'Sullivan, B. Scott, P. Tudzynski, Z. An, E. G. Arnaoudova, C. T. Bullock, N. D. Charlton, L. Chen, M. Cox, R. D. Dinkins, S. Florea, A. E. Glenn, A. Gordon, U. Güldener, D. R. Harris, W. Hollin, J. Jaromczyk, R. D. Johnson, A. K. Khan, E. Leistner, A. Leuchtmann, C. Li, J. Liu, J. Liu, M. Liu, W. Mace, C. Machado, P. Nagabhyru, J. Schmid J. Pan, K. Sugawara, U. Steiner, J. Takach, E. Tanaka, J. S. Webb, E. V. Wilson, J. L. Wiseman, R. Yoshida, and Z. Zeng. Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the clavicipitaceae reveals dynamics of alkaloid loci. *PLoS Genetics*, 9:e1003323, 2013.

[155] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003. ISBN 0-19-850942-1.

[156] C. Semple and M. Steel. *Oxford Lecture Series in Mathematics and its Applications*, volume 24, pages xiv+239. Oxford University Press, 2003.

[157] H. Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 51 (3):492–508, 2002.

[158] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applcations to phylogenetic inference. *Mol Biol Evol*, 16:1114 – 1116, 1999.

[159] M. J. Spiering, C. D. Moon, H. H. Wilkinson, and C. L. Schardl. Gene clusters for insecticidal loline alkaloids in the grass-endophytic fungus *neotyphodium uncinatum*. *Genetics*, 169:1403–1414, 2005.

[160] MJ Stanhope, A Lupas, MJ Italia, KK Koretke, C Volker, and JR Brown. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411: 940–944, 2001.

[161] M. Steel and D. Penny. Distributions of tree comparison metrics-some new results. *Syst Biol*, 42(2):126–141, 1993.

[162] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28:i409–i415, 2012.

[163] Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. *Contemporary mathematics*, 338:357–390, 2003.

[164] J Sukumaran and Mark T. Holder. Dendropy: A python library for phylogenetic computing. *Bioinformatics*, 26:1569–1571, 2010.

[165] D. L. Swofford. *PAUP*. Phylogenetic analysis using parsimony (* and other methods)*. Sunderland Mass., 1998.

[166] N. Takahata. Gene genealogy in 3 related populations: consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

[167] N. Takahata and M. Nei. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, 124:967–978, 1990.

[168] Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.

[169] K.L. Thompson and L. Kubatko. Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, 14:200, 2013.

[170] Y. Tian and L. Kubatko. Gene tree rooting methods give distributions that mimic the coalescent process. *Molecular Phylogenetics and Evolution*, 70:63–69, 2014.

[171] M. Vilaa, J. R. Vidal-Romani, and M. Björklund. The importance of time scale and multiple refugia: Incipient speciation and admixture of lineages in the butterfly Erebia triaria (Nymphalidae). *Molecular Phylogenetics and Evolution*, 36(2):249–260, August 2005.

[172] K. Voigt, E. Cicelnik, and K. O'Donnel. Phylogeny and PCR identification of clinically important zygomycetes based on nuclear ribosomal-DNA sequence data. *Journal of Clinical Microbiology*, 37(12):3957–3964, Dec. 1999.

[173] G. Weyenberg and R. Yoshida. Distances between trees. In Richard Kliman, editor, *Encyclopedia of Evolutionary Biology*. Elsevier (in press), 2015.

[174] G. Weyenberg, P. Huggins, C. Schardl, D.K. Howe, and R. Yoshida. kdetrees: Nonparametric estimation of phylogenetic tree distributions. *Bioinformatics*, 2014.

[175] G. Weyenberg, K. Yoshioka, and D.K. Howe. Normalizing kernels in the billera-holmes-vogtmann treespace. `http://arxiv.org/abs/1506.00142`, 2015.

[176] Grady Weyenberg. kdetrees development repository. `http://github.com/grady/kdetrees/`, 2015.

[177] Grady Weyenberg and Ruriko Yoshida. Reconstructing the phylogeny: Computational methods. In Raina Robeva, editor, *Algebraic and Discrete Mathematical Methods for Modern Biology*, chapter 12. Elsevier, 2015.

[178] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

[179] Z. Yang and B. Rannala. Bayesian species delimitation using multilocus sequence data. *PNAS*, 107(20):9264–9269, 2009.

[180] Y. Yu, J.H. Degnan C. Than, and L. Nakhieh. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2): 138–149, 2011.

[181] Y. Yu, T. Warnow, and L. Nakhleh. Algorithms for mdc-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J Comput Biol*, 18(11):1543–1559, 2011.

[182] Doron Zeilberger. A holonomic systems approach to special functions identities. *Journal of Computational and Applied Mathematics*, 32(3):321 – 368, 1990. ISSN 0377-0427. doi: http://dx.doi.org/10.1016/0377-0427(90)90042-X. URL `http://www.sciencedirect.com/science/article/pii/037704279090042X`.

[183] J. Zhang and M. Nei. Evolutionary distance: estimation. In D. N. Cooper, editor, *Encyclopedia of the Human Genome*. Wiley, 2003.

[184] Zhaolei Zhang and Mark Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18):5338–5348, 2003. doi: 10.1093/nar/gkg745. URL `http://nar.oxfordjournals.org/content/31/18/5338.abstract`.

Vita

Grady Weyenberg received Bachelors degrees in Physics and Mathematics from the University of Arizona, and a masters degree in Statistics from the University of Kentucky. He was both a teaching and research assistant throughout graduate school, and was awarded the R.L. Anderson Research Award for 2014 by the Department of Statistics at the University of Kentucky. He has accepted a postdoctoral position with the Integrative Epidemiology Unit at the University of Bristol, United Kingdom.

Publications

- G. Weyenberg, P. Huggins, C. Schardl, D.K. Howe, and R. Yoshida. kdetrees: Nonparametric estimation of phylogenetic tree distributions. *Bioinformatics*, 2014

- Grady Weyenberg and Ruriko Yoshida. Reconstructing the phylogeny: Computational methods. In Raina Robeva, editor, *Algebraic and Discrete Mathematical Methods for Modern Biology*, chapter 12. Elsevier, 2015

- G. Weyenberg and R. Yoshida. Distances between trees. In Richard Kliman, editor, *Encyclopedia of Evolutionary Biology*. Elsevier (in press), 2015

- G. Weyenberg, K. Yoshioka, and D.K. Howe. Normalizing kernels in the billera-holmes-vogtmann treespace. `http://arxiv.org/abs/1506.00142`, 2015