Averages in Tree Space

Mercedes Coleman, Cody FitzGerald, Amber Holmes, and Emily Smith
Mentors: Colby Long and Seth Sullivant

July 30, 2014

Abstract

Tree space is a subset of Euclidean space in which every point represents a tree. In this space, there exist numerous metrics, each of which lead to a different notion of a mean tree. We have implemented the algorithm for the $(1,1)$, $(2,2)$ and $(1,\infty)$-mean in MATLAB.

The $(1,1)$-mean is a weighted majority rule consensus tree, the $(2,2)$ mean is generated using Sturm's algorithm. Unlike the $(1,1)$ and $(2,2)$-mean, the $(1,\infty)$ mean is not as well studied. The $(1,\infty)$-mean algorithm we propose incorporates traversing the L_2 geodesic using the L_1 midpoint as a mechanism for obtaining the mean tree's orthant. Once the orthant has been found, the mean tree can be optimized using linear programming.

Introduction

Phylogeny is the branch of biology that is focused on studying and illustrating the relationships between a set of species. Different software packages can generate one set of trees while partitioning the genome itself can generate a completely different set of trees; one of the goals in this project is to establish algorithms for constructing averages of trees. Mathematicians invest in this concentration of science by exploring the properties and structures that accompany each of these trees. We use ideas from graph theory and combinatorics to discuss the space of all possible trees; this will enable us to see how to traverse a path between two trees in this space. Computing the average of a set of trees is useful because it provides a way to choose the most representative tree for a set of relationships. The ultimate goal in this project is to study different averages of trees in several metrics to establish one most representative mean tree for the original set of data.

Trees, Their Properties, and Their Residency

A *tree* is defined to be a connected acyclic graph, meaning that there exists a path from any given vertex to every other vertex in the set and no cycles exist. To be subcategorized into a phylogenetic tree, the vertices of degree one in the graph must be labelled and no vertex of degree two can exist. The vertices of degree one are commonly referred to as the *leaves* of the tree. Trees are often represented by a partitioning of the labelling set. This partitioning is intuitively called an *X-split* and denoted as $A \mid B$, where X is the labelling set, A is a subset of the labelling set, and B is the complement of A. Visually, a split is defined by "removing" an edge and thus splitting the tree into two pieces. The initial way to compare a set of trees, or a tree to itself, is to determine if a pair of splits are *compatible*. Splits $A_1 \mid B_1$ and $A_2 \mid B_2$ are compatible if at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, and $B_1 \cap B_2$ is the empty set. This definition of compatibility is the root of the Splits Equivalence Theorem, which says that given a tree all of its splits are compatible and, similarly, a set of compatible splits can construct one unique tree.

If one half of the partition is a single element, a leaf, the split is called a trivial split. Since trivial splits exist in every tree they are often disregarded, while non-trival splits build the space in which the set of all possible trees reside. In Euclidean space, a point resides in a quadrant of the coordinate plane; however,

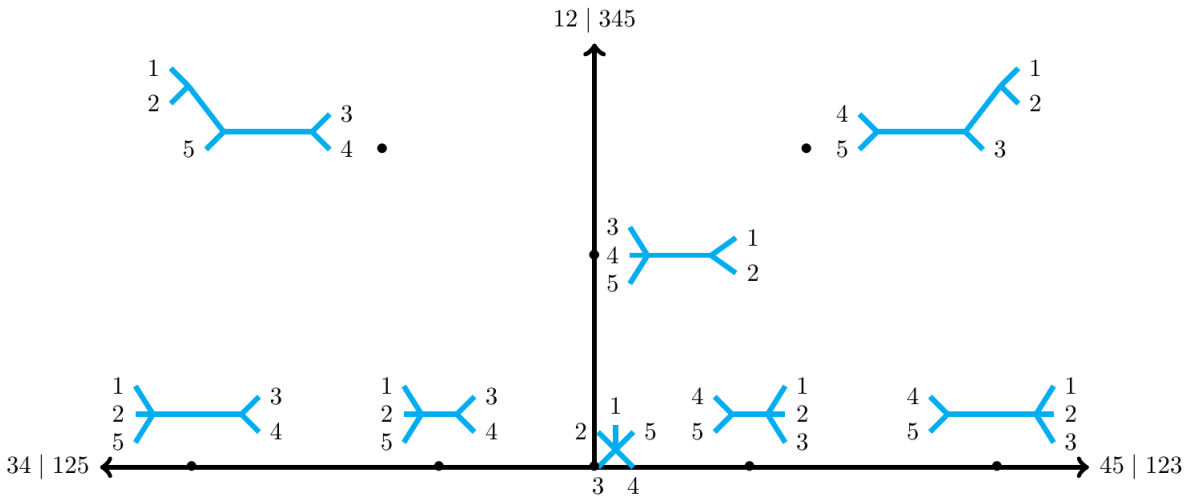

Figure 1: Each orthant represents all possible n-leaf trees with interior edges specified by the boundary rays.

trees create their own space called *tree space*. While tree space is similar to Euclidean space in that a tree is represented by a single point this point resides in an open space called an *orthant* that is bounded by the non-trivial splits of the tree called *boundary rays*. Also similar to the axes in Euclidean space, the intersection of the boundary rays is called the *origin*. As a point travels away from the origin along a specific boundary ray the corresponding edge becomes longer. An orthant defines a space in which a tree of the same topology and labelling set can change with varying interior edge lengths. To compare distinct trees on the same labelling set we define their orthants and "glue" them together by their common split, if it exists. Thus, we can traverse a path from a tree in one orthant to another by traveling through tree space.

A next step to consider is computing the shortest path between any two trees called the *geodesic*. This path can be constructed and measured in a variety of metrics. In general, p-distance or L_p-metric for $p \in [1, \infty]$ is given by the following definition.

Definition 1 *For two points* $x = (x_1, x_2, ..., x_n)$ *and* $y = (y_1, y_2, ..., y_n)$, $d_p(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$.

This definition is for Euclidean space, but a similar one could be written for tree space by placing a p-metric on each orthant traveled through and gluing the pieces together to get the overall distance. We are most concerned with the L_2 geodesic. Depending on the construction of the orthants the shortest path may be a path that connects one tree to the origin and the origin to the second tree; this path is called the *cone path*. A cone path can be the geodesic but the geodesic is not always the cone path. To determine if the geodesic is the cone path we must compare the slopes from each tree to the origin.

Now that we have built this understanding of tree space, we can calculate the mean of a set of points within it. There are a number of different means we can take, each defined by parameters p and q.

Definition 2 *Given a set of points,* $X_1, X_2, ..., X_m$, $p \in [1, \infty]$, *and* $q \in [1, \infty]$ *the (p,q)-mean is given by:*
$argmin \sum_{i=1}^{m} d_p(X_i, X)^q$ *for* $q \in [1, \infty)$ *and*
$argmin\ max_{i=1}^{m} d_p(X_i, X)$ *for* $q = \infty$.

In this project, we explored the $(1, 1), (2, 2)$, and $(1, \infty)$-means.

The (1,1)-Mean

In 1-dimensional Euclidean space, the (1,1)-mean of a set of points is the median, which may or may not be unique; if there is not a unique middle point then the mean is any point between the middle pair of points, inclusive. The 1-dimensional (1,1)-mean generalizes to higher dimensions very well. In n-dimensional Euclidean space, the (1,1)-mean is determined by finding the median value of each coordinate. The (1,1)-mean can then be generalized to tree space analogously by allowing the coordinates to denote splits and coordinate values to denote the edge length of the associated splits.

Proposition 3 *The splits of the (1,1)-mean tree are the splits contained in the Majority Rule Consensus Tree.*

Since the splits of the mean tree are known, the edge lengths can be calculated by taking the median of the edge lengths of each split from the trees in the set.

The (2,2)-Mean

The (2,2)-mean is the point that minimizes the sum of the squared L_2 distances from itself to all points in the set; in Euclidean space this is the usual mean. The orthant structure of tree space complicates finding the (2,2)-mean tree. Instead of the ordinary optimization problem executed in Euclidean space, Sturm's algorithm is used to approximate the (2,2)-mean tree (Ref. 4). Sturm's alogrithm is an iterative approximation that involves numbering the trees in the set, choosing the first tree, T_1, as the first estimate, M_1, and finding the k^{th} estimate by taking the point $\frac{1}{k}$ of the length of the geodesic between the previous estimate and the next tree in the set. This process can be represented by the following equations:

$$\begin{aligned} M_1 &= T_1 \\ M_2 &= \frac{1}{2}M_1 + \frac{1}{2}T_2 \\ M_3 &= \frac{2}{3}M_2 + \frac{1}{3}T_3 \\ &\vdots \\ M_k &= \frac{k-1}{k}M_{k-1} + \frac{1}{k}T_{k\,mod(m)}. \end{aligned}$$

The higher the value of k the closer this estimate is to the $(2, 2)$-mean tree.

The (1,∞)-Mean

Generating the (1,∞)-mean tree incorporates ideas from the L_2 geodesic, L_1 distance, and a Sturm-like algorithm. Given a set of trees, select a tree, T_1, to be the original estimate for the (1,∞)-mean tree. Next find the tree, T_2, in the set that maximizes the L_1 distance between T_1 and T_2. Compute the L_2 geodesic between T_1 and T_2. Travel along the geodesic until the $\frac{1}{k}$ L_1 point is reached (along the L_2 geodesic) where k is the iteration number. Let P(x, y, λ) denote the point that is $\lambda\, d_1(x, y)$ along the L_2 geodesic between x and y. Using this notation and letting $\lambda_k = \frac{1}{k}$, we can generate a Sturm-like algorithm represented formulaically:

$$\begin{aligned} M_1 &= T_1 \\ M_2 &= P(M_1, T_2, \lambda_2) \\ M_3 &= P(M_2, T_3, \lambda_3) \\ &\vdots \\ M_k &= P(M_{k-1}, T_k, \lambda_k). \end{aligned}$$

This point is the updated estimate. This Sturm-like algorithm continues until orthant convergence occurs. Once the mean trees orthant has been located, linear programming techniques can be utilized to find the $(1,\infty)$ mean tree.

Conclusion and Future Work

We implemented algorithms for finding the geodesic and the mean trees in each of these metrics in MATLAB. Tree data is stored as a matrix of splits for all trees in the set and MATLAB efficiently computes the mean tree in each of the previously defined metrics. Future work on this project includes establishing proof for the $(1, \infty)$-mean tree algorithm and exploring the $(2, \infty)$ mean.

References

[1] L. J. Billera, S. P. Holmes, K. Vogtmann, Geometry of the Space of Phylogenetic Trees. Adv. Appl. Math., V.27, pages 733-767, 2001.

[2] E. Miller, M. Owen, J. S. Provan, Polyhedral computational geometry for averaging metric phylogenetic trees. Cornell University Library, V.1, 2012.

[3] M. Owen, J. S. Provan, A Fast Algorithm for Computing Geodesic Distances in Tree Space. Cornell University Library, V.1, 2009.

[4] C. Semple and M. Steel, Phylogenetics. Oxford Lecture Series in Mathematics and Its Applications, 2003.