Mathematical Phylogenetics and The Space of Trees

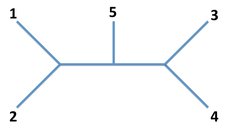
{ Mercedes Coleman, Cody FitzGerald, Amber Holmes, Emily Smith }
Graduate Assistant: Colby Long and Mentor: Seth Sullivant.

Introduction

Phylogeny is the branch of biology that focuses on illustrating and studying the relationships between a set of species. Mathematical phylogenetics studies the structure and properties of these relationships which we represent as trees. Using ideas from graph theory, we discuss the space of all possible trees and how to traverse a path between two trees in this space. We can also compute the average of a set of trees in a variety of metrics; these mean trees are interesting because variations in data from the same species or even different software packages using the same data often build different trees to represent one set of relationships. It is then useful to examine multiple averages to determine the most representative tree for a particular set of relationships.
Our project focuses on implementing algorithms for these different mean trees in MATLAB and developing an algorithm for the (1,∞)-mean tree.

Splits Equivalence Theorem

- An **X-split** is a partition of X into two non-empty sets. We denote the X-split whose blocks are A and B by A|B.
- Any given pair of X-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.



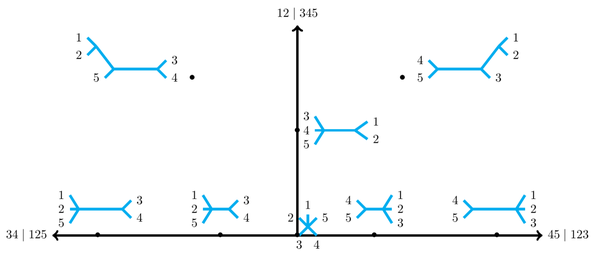
Theorem 1 ***Splits-Equivalence Theorem.*** *Let $\sum$ be a collection of X-splits. Then, there is a tree, T, such that $\sum = \sum(T)$ if and only if the splits in $\sum$ are pairwise compatible. Moreover, if such a tree exists, then, up to isomorphism, T is unique.*

1	2345
2	1345
3	1245
4	1235
5	1234
12	345
34	125

Tree Space

Tree space for n-leaf trees:

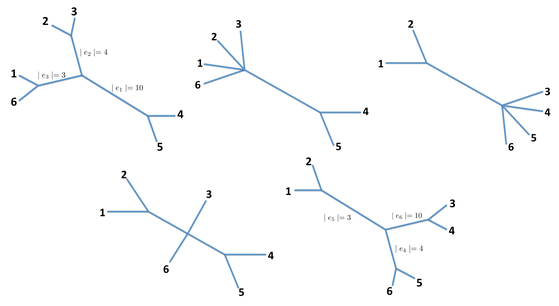
- **Tree space** is a space constructed such that every point represents a possible tree.
- Non-trivial splits determine rays to form orthants.
- Orthants are "glued" together along common boundary rays.
- Distance away from the origin represents edge length.



Geodesic in Tree Space

Geodesic:

- The unique shortest path from one point to another.
- Exists for any two points in tree space because the space is CAT(0).



For any two trees, one tree (T_1) will have a set of edges A and the other tree (T_2) will have a set of edges B. These sets can be partitioned into disjoint sets of splits denoted $A_1, A_2, \ldots, A_k$ from T_1 and $B_1, B_2, \ldots, B_k$ from T_2, such that the sequence of orthants traverse from T_1 to T_2.

$$A = \{e_1, e_2, e_3\} \qquad A_1 = \{e_2, e_3\}$$
$$B = \{e_4, e_5, e_6\} \qquad A_2 = \{e_1\}$$
$$B_1 = \{e_6\}$$
$$B_2 = \{e_4, e_5\}$$

Properties of the Geodesic

We define a **path space** to be the sequence of orthants traversed from one tree to another.

Theorem 2 (Owen, 2011) *If a path satisfies the following two properties, it is called a **proper path**:*

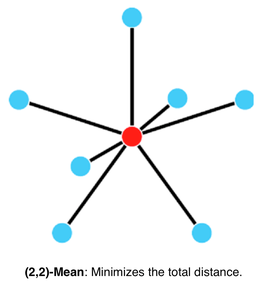
- *For each $i \geq j$, A_i and B_j are compatible.*
- $$\frac{\| A_1 \|}{\| B_1 \|} \leq \frac{\| A_2 \|}{\| B_2 \|} \leq \cdots \leq \frac{\| A_k \|}{\| B_k \|}.$$

Theorem 3 (Owen-Provan, 2011) *If this additional property is satisfied, the proper path is the geodesic.*

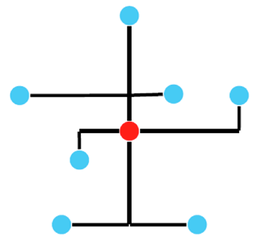
- *For each support pair (A_i, B_i), there is no nontrivial partition $C_1 \cup C_2$ of A_i and partition $D_1 \cup D_2$ of B_i, such that C_2 is compatible with D_1 and $\frac{||C_1||}{||D_1||} < \frac{||C_2||}{||D_2||}$.*

(p,q) mean trees

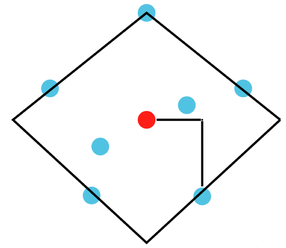
In Euclidean space, the p-distance is given by $d_p(x,y) = \sqrt[p]{\sum_{i=1}^{m}(x_i - y_i)^p}$. The (p,q)-mean is given by $\arg\min \sum_{i=1}^{m} d_p(X_i, X)^q$. When $q = \infty$, the mean is given by $\arg\min \max_{i=1}^{m} d_p(X_i, X)$. These concepts generalize naturally to tree space.



(2,2)-Mean: Minimizes the total distance.



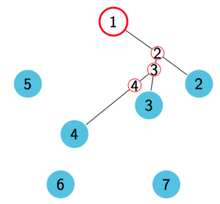
(1,1)-Mean: Finds the median of each coordinate.



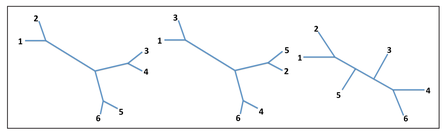
(1,∞)-Mean: $\arg\min \max_{i=1}^{m} d_1(X_i, X)$

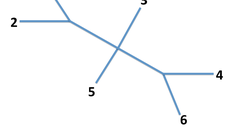
$(2,2)$: The (2,2)-mean tree is found using Sturm's algorithm; an iterative algorithm that takes the form

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



$(1,1)$: The majority consensus tree for a set of trees T is the set of splits that occurs in more than half of the elements of T. The majority consensus tree is essential to the (1,1)-mean.





The trees enclosed in the box above represent some given data while the tree excluded from the grouping represents the majority consensus tree of the data.

$(1,\infty)$: Let $\frac{d^{\lambda_k}}{d^{\lambda_{k-1}}}x + d^{\lambda_k}y$ denote the point that is $\lambda_k\, d_1(x,y)$ along the L_2 geodesic between x and y where $\lambda_k = \frac{1}{k}$.

$$\begin{aligned} M_1 &= T_1 \\ M_2 &= \frac{d^{\lambda_2}}{d^{\lambda_1}}M_1 + d^{\lambda_2}T_2 \\ M_3 &= \frac{d^{\lambda_3}}{d^{\lambda_2}}M_2 + d^{\lambda_3}T_3 \\ &\vdots \\ M_k &= \frac{d^{\lambda_k}}{d^{\lambda_{k-1}}}M_{k-1} + d^{\lambda_k}T_k. \end{aligned}$$

MATLAB Implementation

(2,2)-mean:
A set of trees and the desired number of iterations is input. Approximations are computed as described by Sturm's algorithm (Ref. 4) with help of geodesic and intermediate tree calculations described in Ref. 3. The final approximation is output as the mean tree.
(1,1)-mean:
A set of trees is input. The majority consensus tree (Ref. 1) is determined because these splits are the only ones with nonzero edge length in the mean tree. The median edge length of each split is computed, and the majority consensus tree along with these edge lengths is output as the mean tree.
(1,∞)-mean:
We are in the process of developing an algorithm to compute the (1,∞)-mean tree. In its current form, our algorithm is similar to Sturm's algorithm in that it performs an iterative approximation starting with a tree in the set. However, instead of simply computing the geodesic from the previous approximation to the next tree in the list, our algorithm computes the geodesic from the previous approximation to the tree in the set the greatest L_1 distance from it. The k^{th} approximation is the tree $\frac{1}{k}$ of the L_1 distance from the $k-1^{st}$ approximation to the farthest tree that falls on the L_2 geodesic path.
Note: Trees are generally stored in Newick format, in which a semicolon concludes a tree. The following string represents one tree:
((((2:346872.0,((15:315457.0,(13:281656.0,(14:278519.0,8:278518.0):3138.0):33801.0):18976.0,(12:261023.0,5:261022.0):73411.0):12439.0):288572.0,(6:282673.0,(1:250669.0,9:250669.0):32004.0):352772.0):266701.0,((3:629216.0,7:629216.0):137260.0,(11:623222.0,((10:46777.0,4:46777.0):284276.0,16:331054.0):292168.0):143255.0):135669.0):3209566.0,O:4111714.0); .

References

1. Semple and Steel, *Phylogenetics*.
2. Billera, Louis J., Holmes, Susan P., Vogtmann, Karen. Geometry of the space of phylogenetic trees.
3. Megan Owen, Scott Provan. A Fast Algorithm for Computing Geodesic Distances in Tree Space.
4. Ezra Miller, Megan Owen, J. Scott Provan. Polyhedral computational geometry for averaging metric phylogenetic trees.

Acknowledgements

- Funded by: NSA and NSF grants.
- Supported by: North Carolina State University.

Contact Information

M. Coleman mcoleman@lamar.edu.
C. FitzGerald cel92@wildcats.unh.edu.
A. Holmes abholmes@student.lagrange.edu.
E. Smith smithe1@kenyon.edu.