Phylogenetic Treespace

Alex Sheng, Bowen Li, Claire Chang, Yash Rastogi

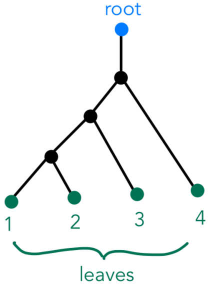
IAS PCMI 2023

August 3, 2023

Phylogenetic Trees

Definition

A **leaf-labeled tree with** n **leaves** is a tree with a distinguished vertex, called the *root*, and n vertices with degree 1, called *leaves*, that are labeled from 1 to n.



Phylogenetic Trees

Definition

A **leaf-labeled tree with** n **leaves** is a tree with a distinguished vertex, called the *root*, and n vertices with degree 1, called *leaves*, that are labeled from 1 to n.

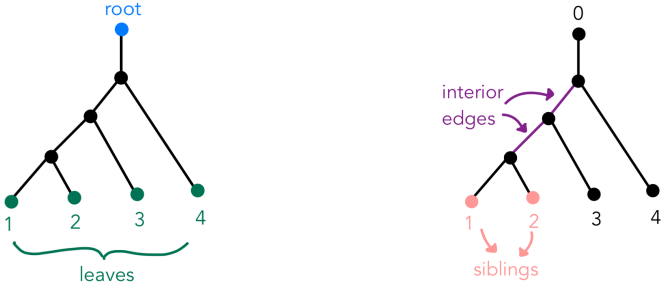


Figure: Parts of a phylogenetic tree.

Problem

- Homology is imprecise!
- Examine the topology to better describe this uncertainty.

Goal

Given a set of leaves, construct the phylogenetic treespace containing all possible trees with a metric defined upon it. Then, study the distances and probability distribution across the treespace to better understand these evolutionary relationships.

Orthants

For a tree with n interior edges with lengths $l_1, l_2, \dots l_n$, the coordinates of a tree in an orthant are determined by $(l_1, l_2, \dots l_n)$. If there are n leaves and the tree is binary, then there are $n-2$ interior edges, and the orthants are $(n-2)$-dimensional.

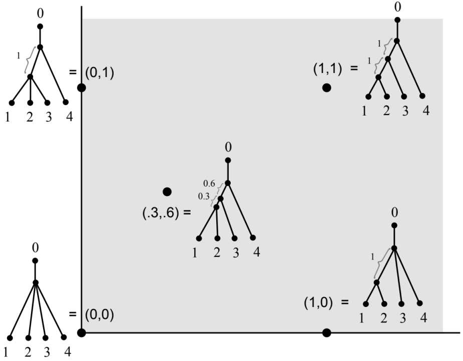


Figure: The 2-dimensional quadrant corresponding to a metric 4-tree, reproduced from [2].

Rotations

Definition

A rotation (or nearest neighbor interchange) is a move which collapses an interior edge to zero and then expands the resulting degree 4 vertex into an edge and two degree 3 vertices in a new way.

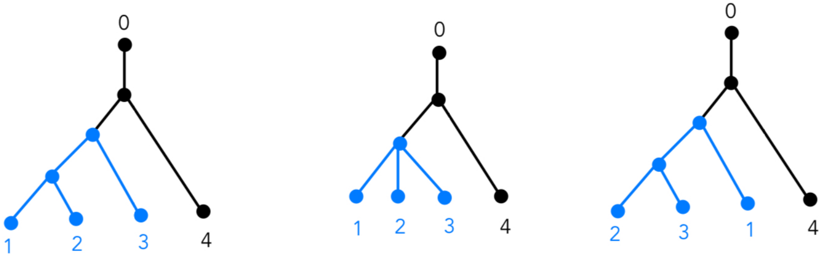


Figure: Example of a tree rotation.

Connecting Orthants

Each orthant represents a different rotation.

Note: non-binary trees are a degenerate case (just the edges).

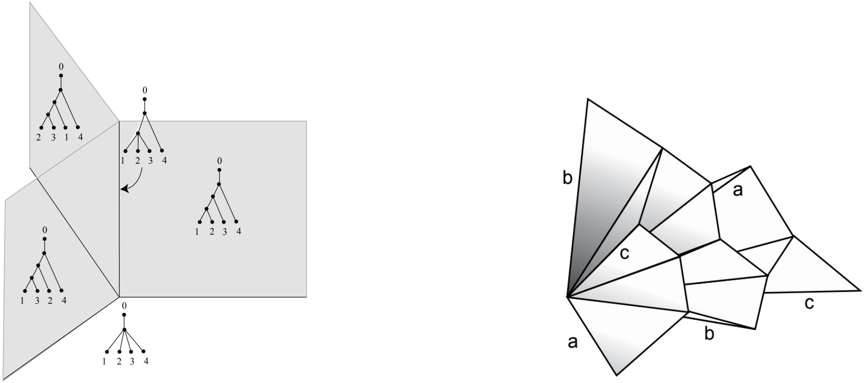


Figure: Connected orthants for the treespace $\mathcal{T}_4$, reproduced from [2].

CAT(0)

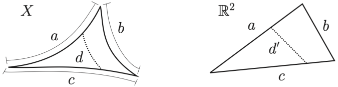


Figure: A chord in a triangle in X, and the corresponding chord in the comparison triangle in $\mathbb{R}^2$. The triangle in X is at least as thin as a Euclidean triangle if $d \leq d'$ for all such chords. Figure from [1].

Definition

A metric space X is CAT(0) if:

- between any two points there is a unique geodesic, and
- every triangle is "at least as thin" as a Euclidean triangle.

Theorem (Billera 2001 [2])

$\mathcal{T}_n$ is a CAT(0) space.

Geodesic and Cone Path

- Since the tree space $\mathcal{T}_n$ is CAT(0), it follows by Gromov (1987) that there exists a unique geodesic connecting any two points of $\mathcal{T}_n$ (nontrivial!)
- Let us define the *cone path*:

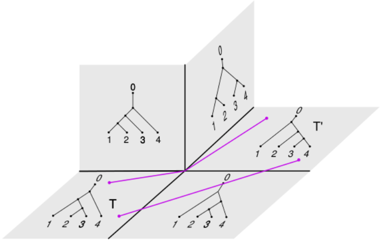


Figure 18: Cone path may or may not be geodesic

- Question: is the cone path the geodesic? (*it's so easy to compute*)

Internal Edge and Edge Partition

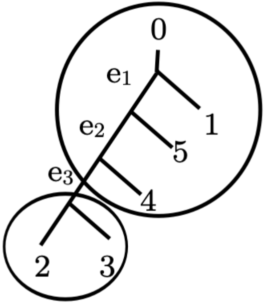


Figure: [4]The internal edge corresponding to partition $\{2,3\} \cup \{0,1,4,5\}$

The partition corresponding to e_3 is $\{23 \mid 0145\}$.
The partition corresponding to e_2 is $\{234 \mid 015\}$.

Is the Cone Path the Geodesic?

- Bridson & Haefliger (1999) shows that for a CAT(0) space, the cone path between two points T and T' is a geodesic iff the angle between is at least π.
- Proposition: if no edge of T is **compatible** with any edge of T', then the cone path *is* the geodesic.
- Corollary: trees that share common edges (i.e., from two neighboring orthants) does not have cone path as the geodesic, which makes sense.
- Proposition: suppose T and T' have no edges in common, but a set of edges E of T and a set of edges F of T' are compatible. If $||T(E)|| \cdot ||T'(E)|| - ||T/E|| \cdot ||T'/F|| > 0$, then the cone path is *not* the geodesic.

Length of the Geodesic

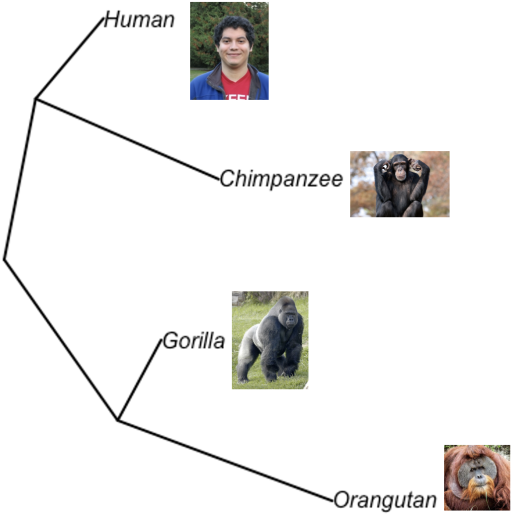
- If two trees are in the same orthant, or if the geodesic is the cone path, then it's easy!
- If not, we have the following:

Theorem

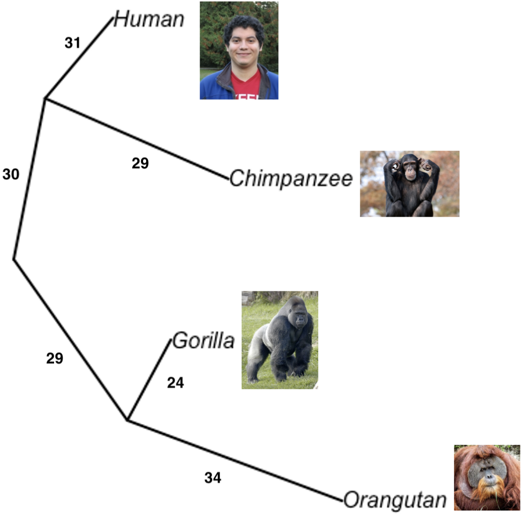
Let T and T' be binary trees with no edges in common. Suppose the edges $\{e_i\}$ of T and $f_i\}$ of T' can be ordered in such a way that $E_i = \{e_1, \cdots, e_i\}$ and $F_i = \{f_1, \cdots, f_i\}$ are compatible for all i. If for all $i < j$ we have $|e_i| \cdot |f_j| - |e_j| \cdot |f_i| > 0$, then the geodesic from T to T' contains trees with edge sets $E_i \cup F_i$ for all i, and the geodesic from T to T' has length the length of the vector

$$(|e_1| + |f_1|, \cdots, |e_{n-2}| + |f_{n-2}|).$$

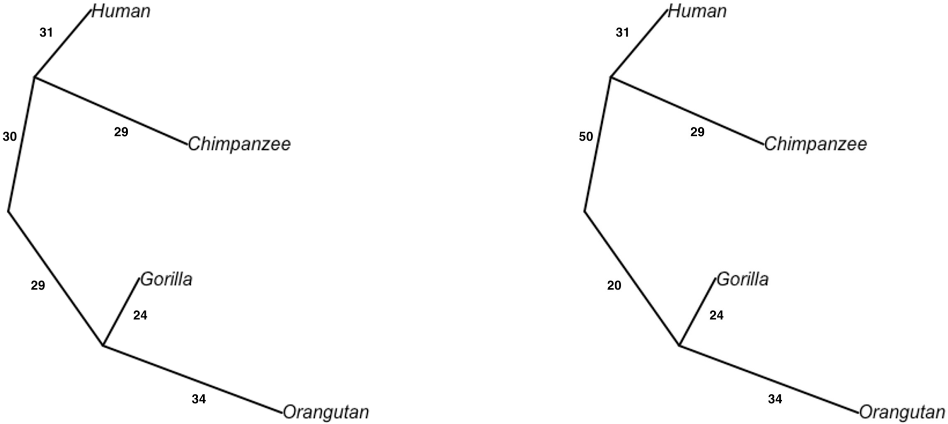
Example of Evolutionary Tree



Example of an Evolutionary Tree



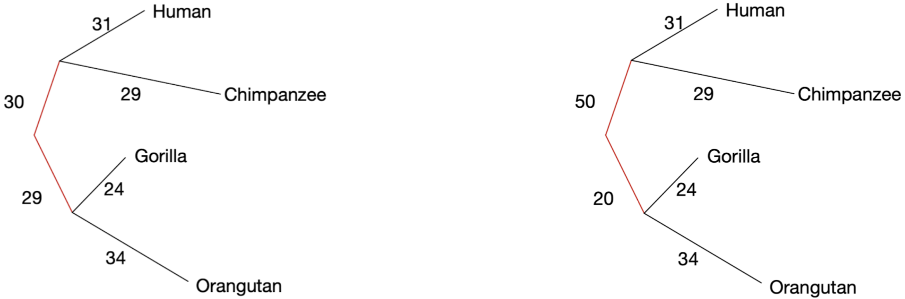
Computing Distances by Hand I: Euclidean Distances



(a) Model A

(b) Model B

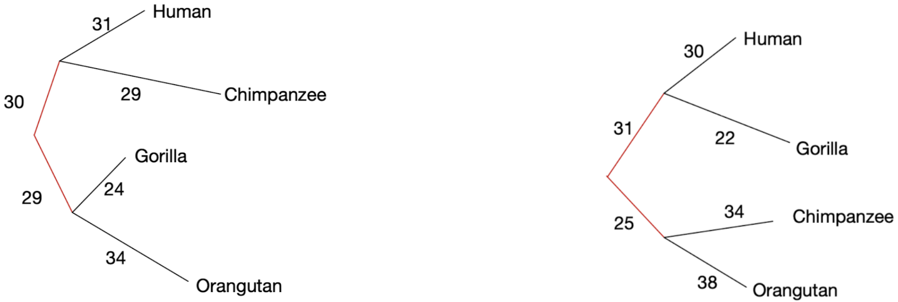
Computing Distances by Hand I: Euclidean Distances



(a) Model A (b) Model B

Two models represent the same tree, so we only need to calculate the Euclidean distance. The Model A has coordinate $(30, 29)$ and the model B has coordinate $(50, 20)$. The euclidean distance would be $\sqrt{(50-30)^2+(29-20)^2} \approx 22$.

Computing Distances by Hand II: Cone Path



(a) Model A

(b) Model C

One can check that for Model A and Model C no two edges are compatible, then we only need to calculate the cone path. The Model A has coordinate $(30, 29)$ and Model C has coordinate $(31, 25)$. Thus, the length of cone path is $\sqrt{30^2 + 29^2} + \sqrt{31^2 + 25^2} \approx 81.5$

Computing Distance by Hand III: Geodesic

Computing Distance by Hand III: Geodesic

Well, I am not going to do that by hand.

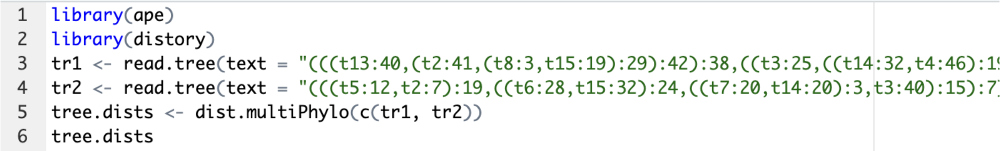
Computing Distance by Hand III: Geodesic

Well, I am not going to do that by hand. Calculating the length of geodesic on CAT(0) is actually NP-hard [3]!

Computing Distance by Hand III: Geodesic

Well, I am not going to do that by hand. Calculating the length of geodesic on CAT(0) is actually NP-hard [3]!

- R Package: ape, distory
- code snippets



```r
library(ape)
library(distory)
tr1 <- read.tree(text = "(((t13:40,(t2:41,(t8:3,t15:19):29):42):38,((t3:25,((t14:32,t4:46):1
tr2 <- read.tree(text = "(((t5:12,t2:7):19,((t6:28,t15:32):24,((t7:20,t14:20):3,t3:40):15):7
tree.dists <- dist.multiPhylo(c(tr1, tr2))
tree.dists
```

Figure: Code for calculating distances

Larger Examples

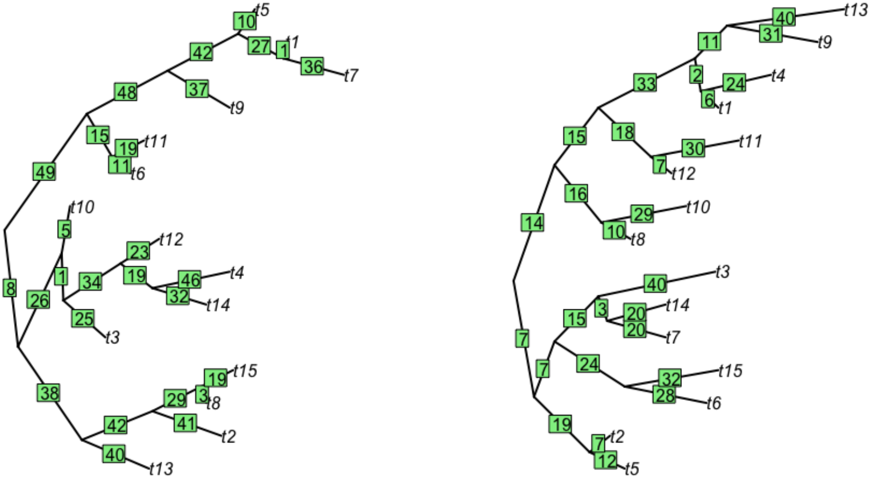


Figure: Two trees with 15 leaves

Larger Examples

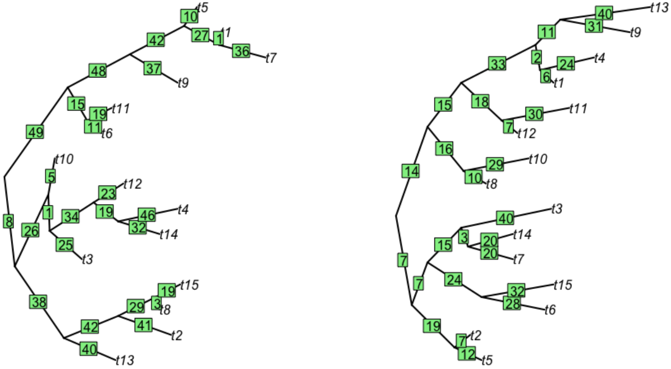


Figure: Two trees with 15 leaves

Using computer codes presented above, we calculated that the distance is approximately 184.

Key Takeaways

Key Takeaways

- Leaf-labeled trees are important to biologists.

Key Takeaways

- Leaf-labeled trees are important to biologists.
- We embeds the set of phylogenetic trees into a CAT(0) space, which has a well-defined notion of distance.

Key Takeaways

- Leaf-labeled trees are important to biologists.
- We embeds the set of phylogenetic trees into a CAT(0) space, which has a well-defined notion of distance.
- Having quantitative metric also allows biologists to statistically evaluate the credibility of evolutionary models.

Biological Problems Amenable to Mathematical Approaches

Biological Problems

The problem we have investigated is quite similar mathematically to other biological problems:

- Protein Folding Mutagenics
- Chromosomal translocations
- Comparisons to determine the degree of biological similarity (of e.g. biomolecules, neural structures)

Mathematical Characterization of These Problems

- Embeddings of graph structures into metrizable topological groups
- Simple automorphism groups of trees determined by their actions on finite subtrees

Trees and the Theory of Free Groups

The following was the first result on the structure of discrete subgroups of p-adic groups:

Theorem (Ihara 1966 [5])

Every torsion-free subgroup of $SL_2(\mathbf{Q}_p)$ is a free group.

The proof was difficult and *ad hoc*. Trees allow us to systematize and simplify such proofs (i.e. the tree of SL_2 over the field $\mathbf{Q}_p$).

Upshot

To prove a group is free, show that it acts freely on a tree.

Biological Application of Bass-Serre Theory: Step 1

Bass-Serre Theory

- The study of groups acting by automorphisms on simplicial trees (c.f. Serre [5]).
- Motivation: Understanding structure of certain algebraic groups (those whose Bruhat-Tits buildings are trees)
- Key Object of Study: Fundamental group of a graph of groups; a one-dimensional version of orbifold theory

Reducing Biological Trees to Cell Complexes

- Every connected graph such that each vertex has finite degree (e.g. biological trees) can be viewed as a one-dimensional cell complex.
- Correspondence between finitely generated groups and their associated cell complex.
  - Stalling's Theorem characterizes the ends of finitely generated groups through the ends of the cell complex associated to the corresponding graph.

Biological Application of Bass-Serre Theory: Step 2

Apply Combinatorics to Achieve Biological Comparisons

- Each biological difference is an action of a tree's automorphism group.
- Bass-Serre theory decomposes group actions as compositions of
  - Free products with amalgamation (pushouts in the category of groups as seen in the Seifert-van Kampen Theorem)
  - HNN Extensions (group embeddings such that all isomorphic subgroups are conjugate)
- Count the number of each type of automorphism and use it as a "distance" to predict likelihood of biological relationships

Open Question

What are the biological meanings of amalgamated free products and HNN extensions?

References

[1] Federico Ardila-Mantilla. CAT(0) Geometry, Robots, and Society. *Notices of the American Mathematical Society*, 67(07):1, Aug 2020.

[2] Louis J. Billera, Susan P. Holmes, and Karen. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001.

[3] Anne Kupczok, Arndt Von Haeseler, and Steffen Klaere. An exact algorithm for the geodesic distance between phylogenetic trees. *Journal of Computational Biology*, 15(6):577–591, 2008.

[4] Megan Owen. Computing geodesic distances in tree space, 2011.

[5] J.-P. Serre. *Trees (Translated from the French by J. Stillwell)*. Springer-Verlag, 1980.