A path between trees in the phylogenetic forest

Marzieh (Tara) Khodaei and Peter Beerli
Department of Scientific Computing, Florida State University, Tallahassee FL

Introduction

Phylogenetic trees are fundamental for understanding the evolutionary history of a set of species. Understanding the local neighborhoods of a phylogenetic tree is essential, but since trees are high-dimensional objects, discussing these neighborhoods is difficult. We use different distance methods to explore the phylogenetic tree landscape. Based on the geodesic distance between pairs of trees, we developed a method to generate trees on the shortest path between two arbitrary trees.

Method

We generate a *geodesic* $\Gamma = (\gamma(\lambda) : 0 \leq \lambda \leq 1)$ between two arbitrary trees using Theorem 2.4., by Owen and Provan, 2010:

- Let $T = (X, \varepsilon, \Sigma)$ and $T' = (X, \varepsilon', \Sigma')$ are two disjoints trees.
- A *Support* (A, B) s.t. $A = (A_1, A_2, ..., A_k)$ and $B = (B_1, B_2, ..., B_k)$
- The geodesic distance $L(\Gamma) = \|(\|A_1\|, ..., \|A_k\|) + (\|B_1\|, ..., \|B_k\|)\|$
- $k + 1$ legs of geodesic:

$$\Gamma^i = \begin{cases} \left[\gamma(\lambda) : \frac{\lambda}{1-\lambda} \leq \frac{\|A_1\|}{\|B_1\|}\right] & i = 0 \\ \left[\gamma(\lambda) : \frac{\|A_i\|}{\|B_i\|} \leq \frac{\lambda}{1-\lambda} \leq \frac{\|A_{i+1}\|}{\|B_{i+1}\|}\right] & i = 1, ..., k-1 \\ \left[\gamma(\lambda) : \frac{\lambda}{1-\lambda} \geq \frac{\|A_k\|}{\|B_k\|}\right] & i = k \end{cases}$$

- A sample tree $T_i = (X, \varepsilon^i, \Sigma^i)$ for i^{th} leg, s.t.

$$\varepsilon^i = B_1 \cup ... \cup B_i \cup A_{i+1} \cup ... \cup A_k \quad , \quad |e|_{T_i} = \begin{cases} \frac{(1-\lambda)\|A_j\| - \lambda\|B_j\|}{\|A_j\|} |e|_T & e \in A_j \\ \frac{\lambda\|B_j\| - (1-\lambda)\|A_j\|}{\|B_j\|} |e|_{T'} & e \in B_j \end{cases}$$

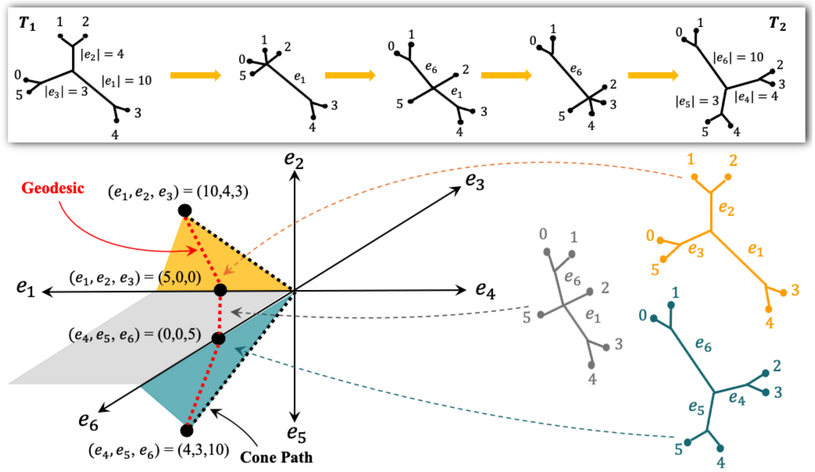

Figure 1. Geodesic distance of $L(\Gamma) = \|(\|A_1\|, \|A_2\|) + (\|B_1\|, \|B_2\|)\| = 15\sqrt{2}$ with support (A, B) s.t. $A = (A_1, A_2) = (\{e_2, e_3\}, \{e_1\})$ and $B = (B_1, B_2) = (\{e_6\}, \{e_4, e_5\})$.

Results

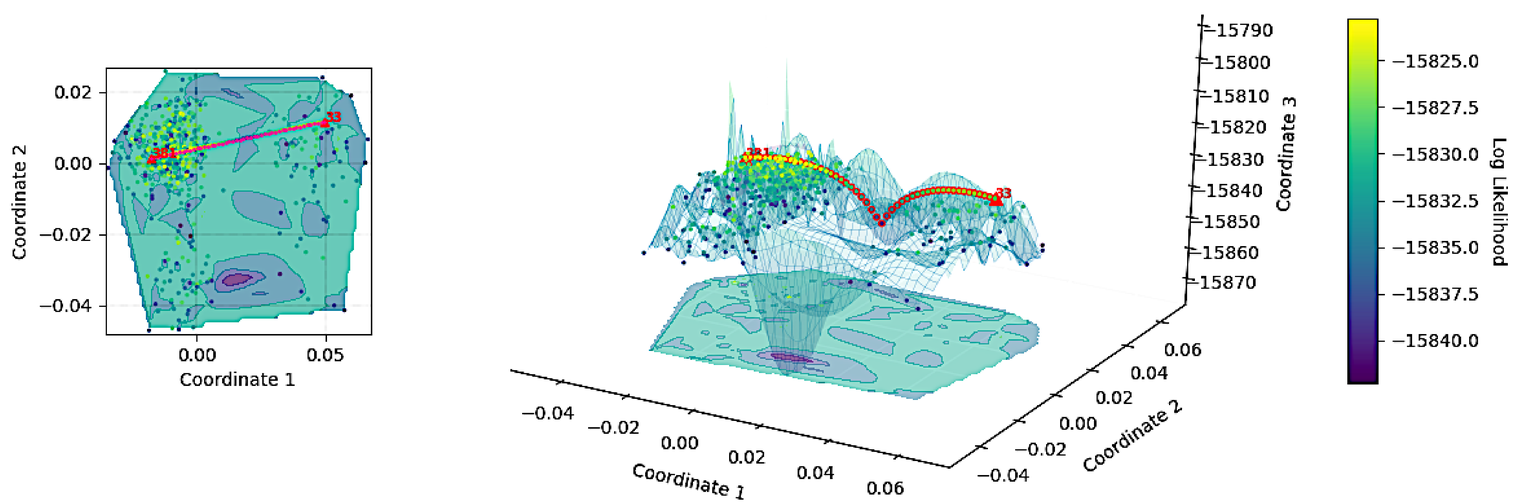

Figure 2: Cubic interpolation of the tree space, displaying both the contour and the surface of tree space by applying the geodesic distance method of GTP (Owen and Provan, 2010). Each dot is a tree; the lighter the dot, the higher the likelihood of the tree. The red dots are the trees generated on the shortest path (geodesic) between two arbitrary trees. We generated a sample of 100,000 trees using the Bayesian Phylogeny inference program REVBAYES and their tutorial dataset primates_cytb_JC. We extracted 500 rooted trees collected during the MCMC run after removing half of the trees as burn-in. We used the geodesic distance metric among all trees to visualize the relationship among them using Multidimensional Scaling.

References

- Megan Owen and J. Scott Provan. 2010. "A fast algorithm for computing geodesic distances in tree space."*EEE/ACM Transactions on Computational Biology and Bioinformatics*, 8.1 : 2–13.
- T. F. Cox, and M. A. A. Cox. "Multidimensional Scaling." *CRC Press, 2000*.
- Höhna, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, Ronquist. 2016. "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language." *Systematic Biology*, 65:726-736.

The relationship among all species is treelike. We do not know the actual tree but infer likely trees using sequence data. Establishing the neighborhood of a tree is difficult. Our method allows us to present and evaluate trees on the shortest geodesic path between two arbitrary trees. Extensions of these path-trees will improve visualization of tree space and also improve the heuristic search for the best tree.