PATHTREES: A Python package to explore the tree landscape

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

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 developed the Python package PATHTREES, which uses the geodesic distance between pairs of trees and their likelihoods to build a continuous tree landscape. We use this tool to find the best tree and describe its neighborhood.

Methodology

Based on the Billera-Holmes-Vogtmann (BHV) distance between pairs of trees, we describe a method to generate intermediate trees on the shortest path between two arbitrary trees, called pathtrees. These pathtrees give a structured way to investigate intermediate neighboorhoods between trees of interest in the BHV tree space and can also be used to find high likelihood trees independently of traditional heuristic search mechanisms. We implemented our algorithm in the Python package PATHTREES which enables the construction of the continuous tree landscape interior of the convex hull of starting trees, low-dimensional visualization of the generated tree landscape, identifying clusters of trees with the same topologies, and searching for the best tree.

Fig. 1 shows an example of pathtrees (red dots) along the shortest path (geodesic) between three arbitrary trees (colored triangles) in the tree space and their corresponding optimized trees (middle-size black dots). Each pathtree and the corresponding optimized tree are connected with a dashed line. We applied multidimensional scaling (MDS) and an interpolation method (cubic spline) to visualize a space of 1000 trees and 90 pathtrees. The log-likelihood was used for the contour color of the surface. Each dot is a tree; the lighter the dot, the higher the likelihood of the tree.

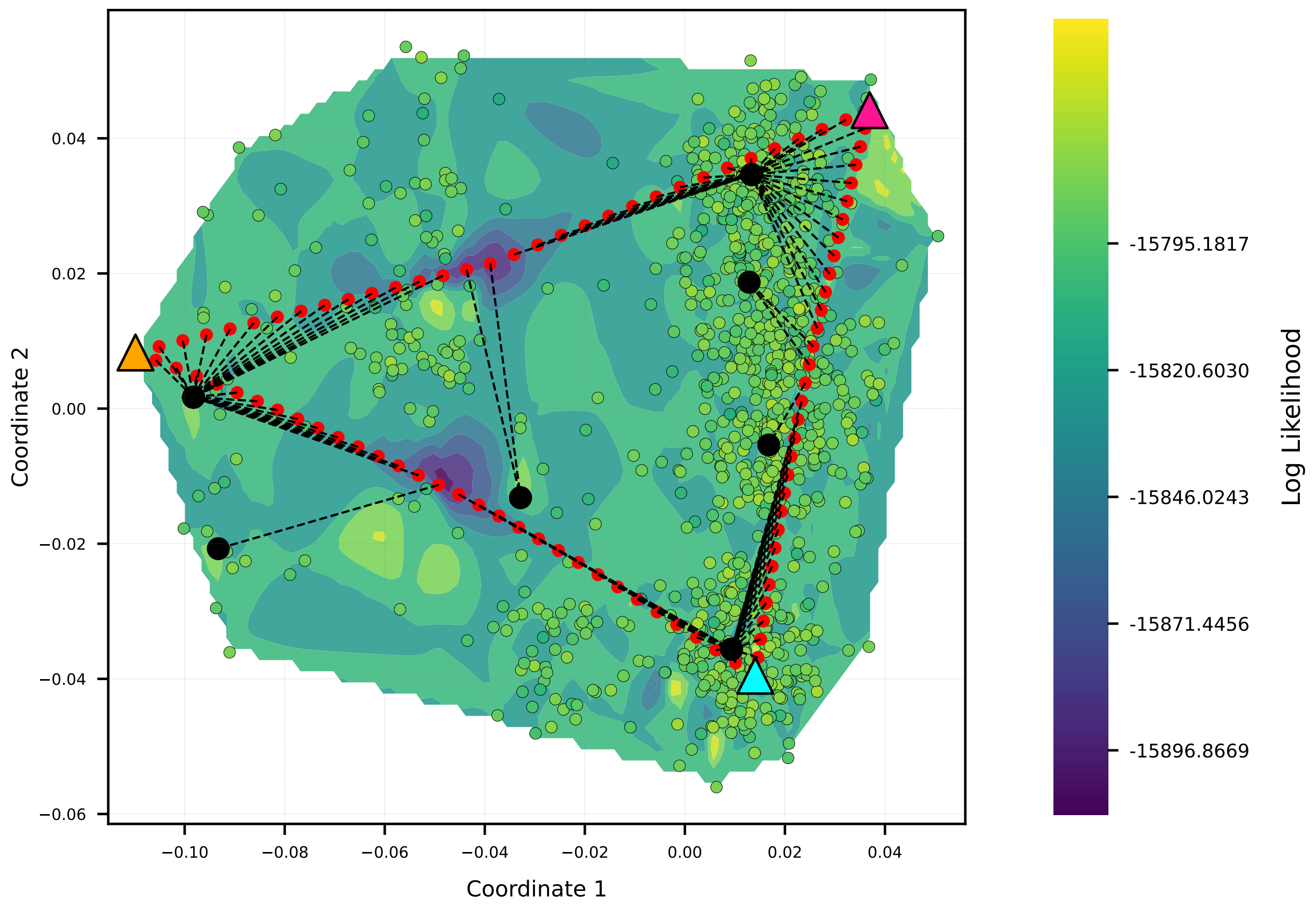

Figure 1. An example of PATHTREES.

Application to Real Data

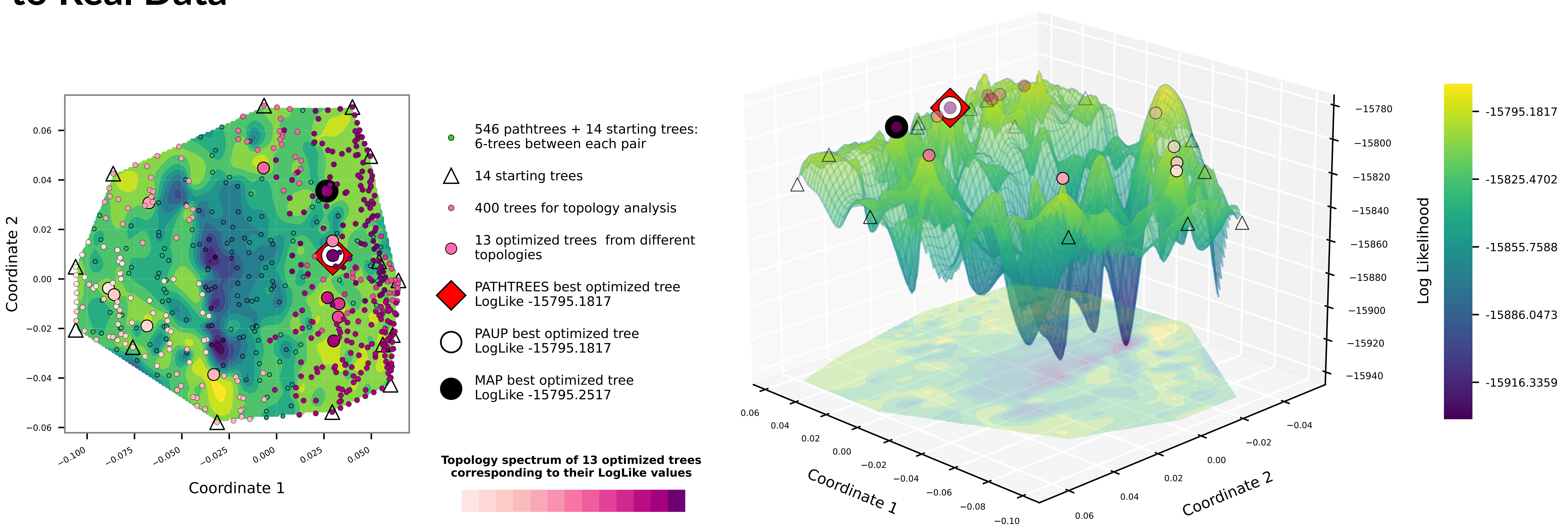

Figure 2: We generated 50,000 trees using the program RevBayes and their tutorial dataset primates_cytb_JC of 1141 base pairs and 23 primate taxa. We selected around 1000 of generated trees, and then extracted the trees on the vertices of the convex hull of these sample trees (14 trees) as starting trees for PATHTREES. This figure shows the contour and surface plots of PATHTREES by generating 6 pathtrees per anchor tree pair, and applying MDS and the RBF thin-plate spline interpolation defined by the BHV distance matrix. Six trees were generated on the geodesic between each pair of starting trees (546 pathtrees). 400 trees were selected from the 546+14 trees and classified based on their topologies (13 topologies with different colors of the purple spectrum). Each medium-size circle shows the optimized tree with the corresponding topology (local optimal trees). The red square shows the best likelihood tree in the tree space found by PATHTREES. We compared our results with those generated by the maximum likelihood software Paup*, and the maximum a posteriori (MAP) phylogeny in RevBayes. The large white circle shows the best tree of Paup* which is identical to PATHTREES' optimal tree. One of PATHTREES' local optimal trees matches the MAP tree (big black circle).

References

- Owen, M. Provan, J. S. (2011), "A fast algorithm for computing geodesic distances in tree space", *IEEE/ACM Trans. Comput. Biol* 8, 2–13.
- Billera, L. J., Holmes, S. P. Vogtmann, K. (2001), "Geometry of the space of phylogenetic trees", *Adv. Appl. Math* 27, 733–767.
- Swofford, D. (2003), 'PAUP*. phylogenetic analysis using parsimony (*and other methods). version 4.'.
- Höhna,, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P. Ronquist, F. (2016), "Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language", *Syst. Biol* 65, 726–736.
- Cox, M. A. Cox, T. F. (2008), Multidimensional scaling., in "Handbook of data visualization", *Springer, Heidelberg, Berlin*, pp. 315–347.
- Buhmann, M. D. (2003), "Radial basis functions: theory and implementations", *Cambridge university press*.
- De Boor, C. (1978), "Piecewise cubic interpolation", in 'A practical guide to splines', *Springer-Verlag, New York*, pp. 40–47.