Explorations in Tree Space: Where are the good trees?

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

Introduction

Bayesian inference is commonly used to find the best phylogeny credibility sets of phylogenies of samples of species. We are interested to understand this search space better to develop improvements of the search algorithms.

Methods

We explore the distribution of trees in tree space; we use the tree-distance developed by Kendall and Colijn(2015). The distance measure can use the *topology* and the *branchlengths*. The contribution of each is adjusted using a weight λ; we set the λ to 0.0 (topology only), 1.0 (branchlengths only), and 0.5 (equal weight of branchlengths and topology). The original distance is favoring the topology component when total branchlengths are small; to fix this, we reweighted the contribution of topology and branchlengths by their respective sums of edges and total branchlengths.

For our experiment, we generated a sample of 100,000 trees using the program REVBAYES (Höna et al. 2016) and their tutorial dataset *p*rimates_cytb_JC. From this sample, we extracted 100 rooted trees that were collected during the MCMC run after removing half of the trees as burn-in. We then used the distances among all pairs to visualize the relationship among them using Multidimensional Scaling (MDS – Trevor et al. 2000). We used the *splitstree* framework (Huson and Bryant 2006) to study the best likelihood-trees and their topological congruence.

Results

The trees fell into several different groups depending on λ ; λ=0, favoring topology, produces tight clusters of trees, each group has the same topology, with an λ=1 that ignores topology only one large group containing all trees can be seen, intermediate leads to groupings by topology, but even with λ=0.5 these groups are tight.

We marked the 6 trees with the highest likelihoods (Fig. 1): they are labeled 30, 34, 66, 79, 83, and 93 . It turns out that these 6 best trees have the same topologies and thus are in the same group with λ=0; with λ=1 they are widespread in the cluster. The *neighbornet* (Fig. 2 left) based on the data shows that there is considerable potential of uncertainty near the center of the unrooted tree. The combined six best tree (Fig. 2 right) show that none of the nodes is conflicting.

Discussion and Conclusion

The analyses corroborate that treespace is complex. There can be many groups depending on λ. Group recognition depends on the weighting scheme of branch lengths and topology, where topology only seems to have a better resolution but little information beyond that each group has different topology; This corroborates that developing algorithms that may calculate gradients for trees to improve Markov chain Monte Carlo tree-searches may be difficult because incorporating information about branch length we loose information about the topology. Finding the maximum likelihood tree may be more difficult than we like.

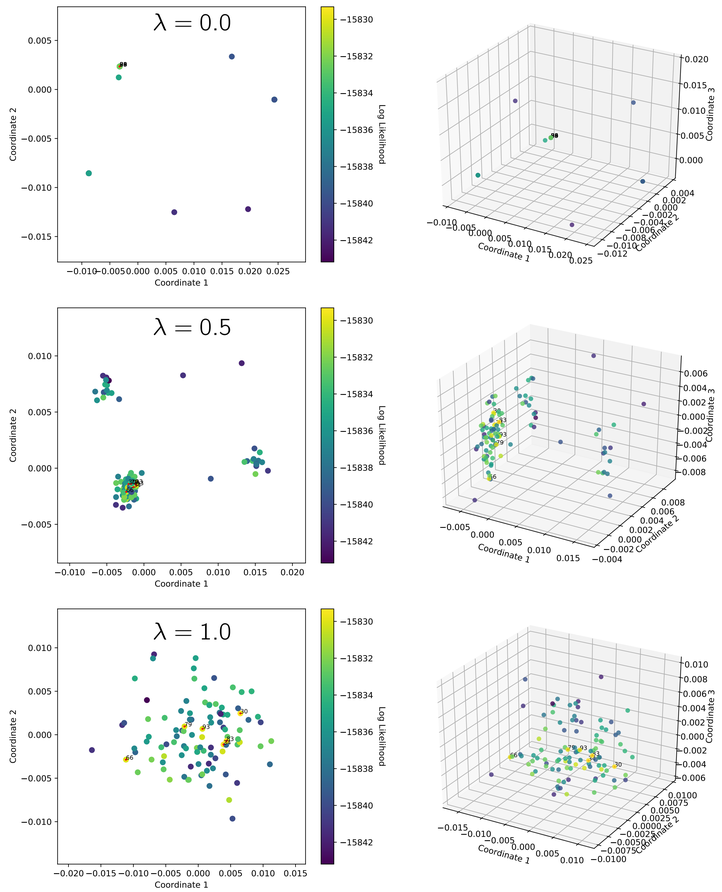

Figure 1: Visualization of tree space in the first two dimensions (left) and three (right) dimensions of the MDS plot; top row topology only: λ=0; middle row: topology+branchlengths: λ=0.5; bottom: branchlengths only:λ=1; each dot is a tree, the lighter the dot the higher the likelihood of the tree, the 6 best trees are marked with red.

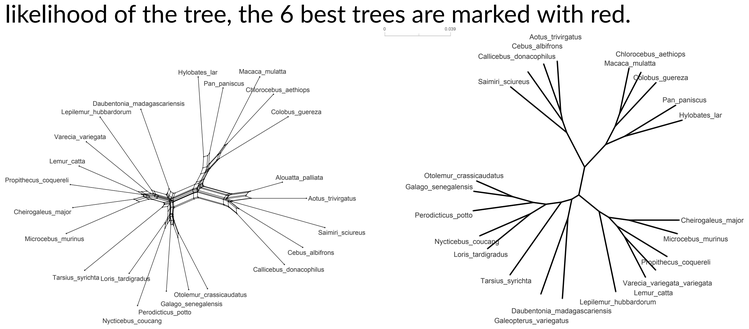

Figure 2: *Splitstree* analysis: left: *Neighbornet* of the of the REVBAYES primate cytochrome B dataset showing considerable uncertainty in the center of the unrooted tree; right: combination of all 6 best likelihood trees, showing the same topology.

References

M. Kendall, and C. Colijn. 2016. "Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution."*Molecular Biology and Evolution*, 33(10):2735–2743.
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.
D.H. Huson and D. Bryant.2006. "Application of Phylogenetic Networks in Evolutionary Studies. " *Molecular Biology, and Evolution*, 23(2):254-267. Software available from www.splitstree.org

Tree space has a complex structure.

We corroborate that trees on the same tree-island defined by a distance metric that is based on the topology have similar likelihoods.

In contrast, incorporating branch lengths into the distance measure fails to show clear groupings.

Take a picture to download a copy of the poster.

[Originally presented at the SSB meeting Gainesville January, 2020, edited and reanalyzed the trees with a corrected distance program for the presentation at FSU Scientific Computing XPO, April 2020]