Robust Analysis of Phylogenetic Tree Space

Pawan Kumar

Centurion University, Department of EEE, Parlakhemundi, Odisha

Data Structure using C++

Abstract

Phylogenetic analyses often produce large numbers of trees. Mapping trees' distribution in "tree space" can illuminate the behavior and performance of search strategies, reveal distinct clusters of optimal trees. Here, I explore the consequences of this transformation in phylogenetic search results from 128 morphological data sets, using stratigraphic congruence Distances mapped into two or even three dimensions often display little correspondence with true distances, which can lead to profound misrepresentation of clustering strcture. My recommendations for tree space validation and visualization are implemented in a new graphical user interface in the "TreeDist" R package.

Materials and Methods

Wright and Lloyd (2020) used a selection of 128 morphological data sets to demonstrate how tree space analysis can facilitate the interpretation of phylogenetic results. A single MCMC run was executed in "RevBayes" (Höhna et al. 2016) for 300,000 generations. To minimize the risk of artifacts due to non-convergence of chains, I conservatively discard the first 50 of Bayesian trees as burn-in, and sample 2500 of the remaining trees at uniform intervals to represent the posterior distribution. Wright and Lloyd (2020) identified most parsimonious trees using TNT (Goloboff and Catalano 2016) under equal-weights parsimony, using exhaustive searches for data sets with ¡25 leaves, and heuristic searches for larger data sets. I include all most parsimonious trees reported, with an upper limit of 1000 trees for each data set.I treat all trees as cladograms, discarding branch length information in order to focus exclusively on the evolutionary relationships contained within each tree. The underlying paleontological data sets contain 4–88 (median: 15) terminal taxa and 8–540 (median: 57) This broad suite of tree sets with disparate properties helps to illuminate, if incompletely, the nature of tree spaces constructed from typical morphological data sets.

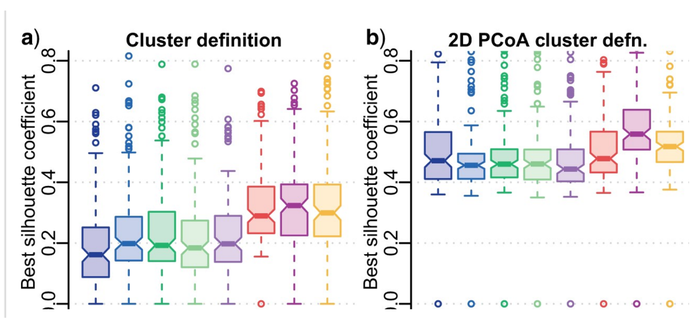
Distances

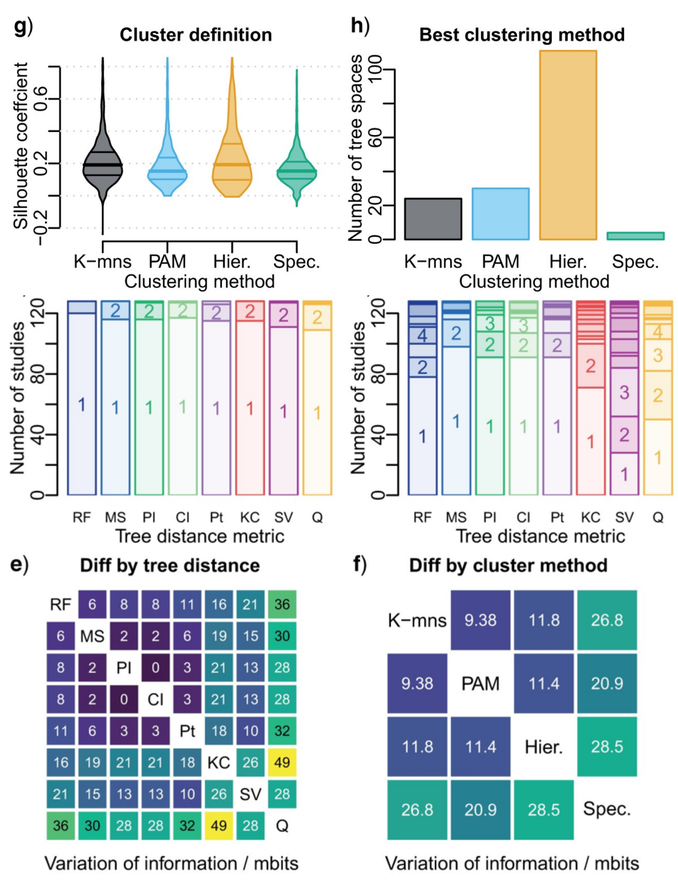
This study considers distances that purport to quantify the similarity of relationships between cladograms: the Robinson–Foulds (RF), matching split information (MS), phylogenetic information (PI), clustering information (CI), path (Pt), Kendall–Colijn (KC), and quartet (Q) metrics, and a new metric (SV) derived from vector representations of trees.The quartet distance (Estabrook et al. 1985) counts whether the relationships between each possible combination of four leaves are the same or different between two trees; it has a similar objective to information-theoretic distances but is slower to calculate.

Clustering

I identify clusters of unique tree topologies using:

- the Hartigan–Wong K-means algorithm (Hartigan and Wong 1979, R function kmeans(), with 3 random starts and up to 42 iterations
- partitioning around medoids (cluster::pam(), Maechler et al. 2019), using 3 random starts and the algorithmic shortcuts of Schubert and Rousseeuw (2021).
- hierarchical clustering with minimax linkage (Murtagh 1983) (protoclust::protoclust(), Bien and Tibshirani 2011) (chosen after outperforming other linkage methods in initial informal analyses)
- spectral clustering (using custom function Tree Dist::Spectral Eigens () alongside cluster::pam()).





Mapping

PCoA is a simple approach, which essentially rotates a high-dimensional space such that as much of the variance of the data as possible falls within the plotted dimensions (Thrun 2018). PCoA requires Euclidean distances, and converting distances between phylogenetic trees into a Euclidean space entails a loss of information (Nye 2011). To make the distances Euclidian, I follow the standard practice of adding a constant to each distance (Cailliez 1983; Jombart et al. 2017), while noting that this might distort the relative magnitude of individual distances. Finally, t-SNE constructs a probability distribution whereby trees that lie close to a specified tree are more probable. A low-dimensional mapping is selected in order that the equivalent treatment of mapped distances replicates this probability distribution as closely as possible.

Distortion

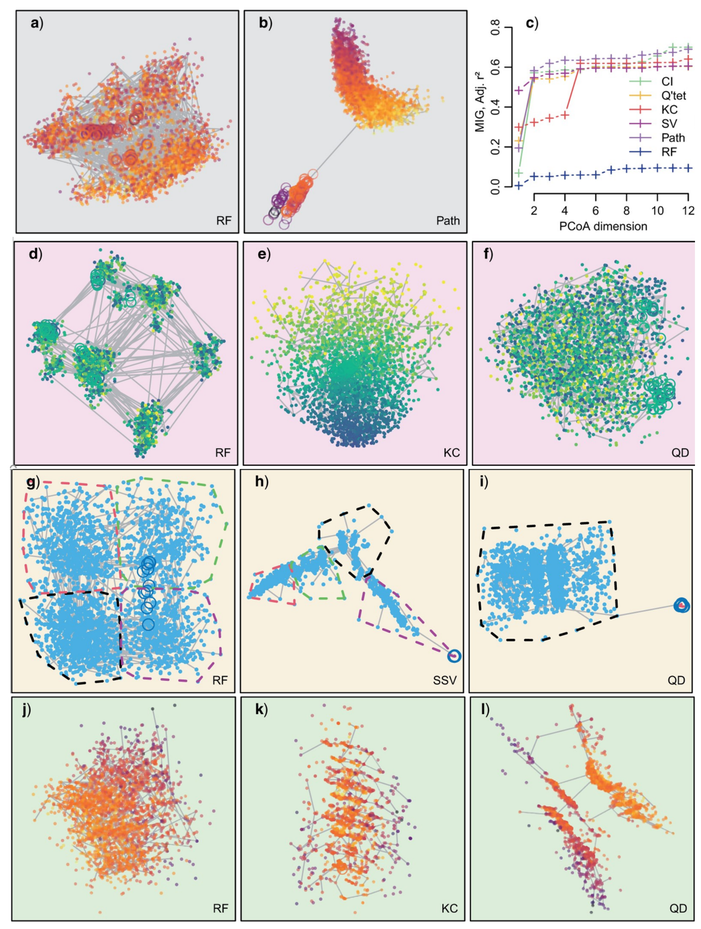
To evaluate the susceptibility of a tree space to distortion on mapping, I calculate its correlation dimension.That is, the number of dimensions necessary in order to reproduce all the structure present in the tree space.

Results

Six-dimensional mappings for each data set, tree distance method, and mapping method, with evaluation of clusterings and depiction of stratigraphic fit, are provided in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.kh1893240 (Smith 2021). Results obtained under the CI distance when trees were rooted do not materially differ from those when trees are treated as unrooted (Smith 2021).

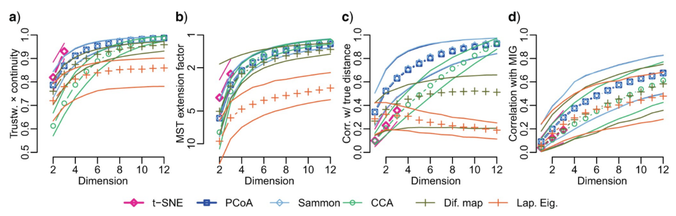
Tree Distance Metric

The results of the tests devised to compare tree distances by Smith (2020a), plus the new "balance independence" test, are presented in Table 1. Of the metrics examined, only the quartet and information-theoretic tree distances consistently reflect differences in the evolutionary relationships within trees (Table 1). Relative to these distances, Euclidian vector-based distances—the path, KC, and sv metrics—do a poor job of representing pre-defined structures in sets of trees.



Mapping Method

Despite being a petroleum- and gas-rich country, Algeria is making efforts to exploit its renewable energies. The Algerian government has adopted new renewable energy laws and financial support for the investors to facilitate the exploitation of the renewable energies for electricity production and direct utilizations. Algeria has relatively abundant geothermal resources especially in the northeastern parts but not totally used.



Visualizing Tree Spaces

Different mapping techniques have different motivations, and thus differ markedly in the structure they depict. Mapping has an order of magnitude more impact on the clustering structures perceived—the easiest aspect of structure to quantify—than the measurement of tree distance or the method of cluster detection.

References

[1] J. Fabre. *Introduction a la geologie du Sahara Algerien et des regions voisines*. Societe Nationale d'Edition et de Diffusion, 1976.

[2] A. Fekraoui and M. Abouriche. Algeria country update report. *Proceedings of the WGC*, pages 31–34, 1995.

[3] A. Fekraoui and F. Kedaid. Geothermal resources and uses in algeria: a country update report. *Proceedings of the WGC*, pages 1–8, 2005.

[4] F. Kedaid. Algerian geothermal country report. *Geothermal and Volcanological Research Report of Kyushu University*, 11:4–6, 2002.

[5] H. Saibi. Geothermal resources in algeria. *Renewable and Sustainable Energy Reviews*, 13:2544–2552, 2009.

[6] D. Takherist and A. Lesquer. Mise en evidence d'importantes variations regionales du flux de chaleur en algerie. *Can. J. Earth Sci.*, 26:615–626, 1989.