Robust Analysis of Phylogenetic Tree Space

Pawan Kumar

Centurion University, Department of EEE, Parlakhemundi, Odisha

Registration No-190101150015 , Subject-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. 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 used a selection of 128 morphological data sets to demonstrate how tree space analysis can facilitate the interpretation of phylogenetic results.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 identified most parsimonious trees using TNT under equal-weights parsimony, and heuristic searches for larger data sets. if incompletely, the nature of tree spaces constructed from typical morphological data sets.

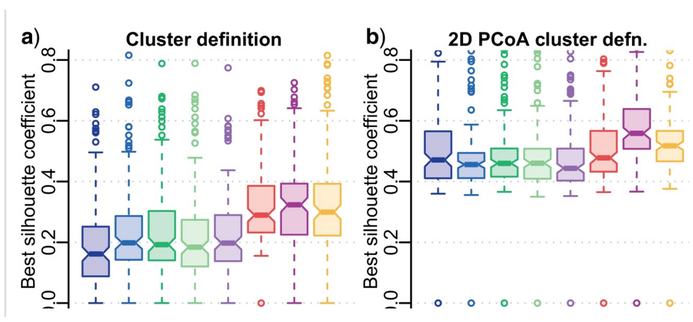
Distances

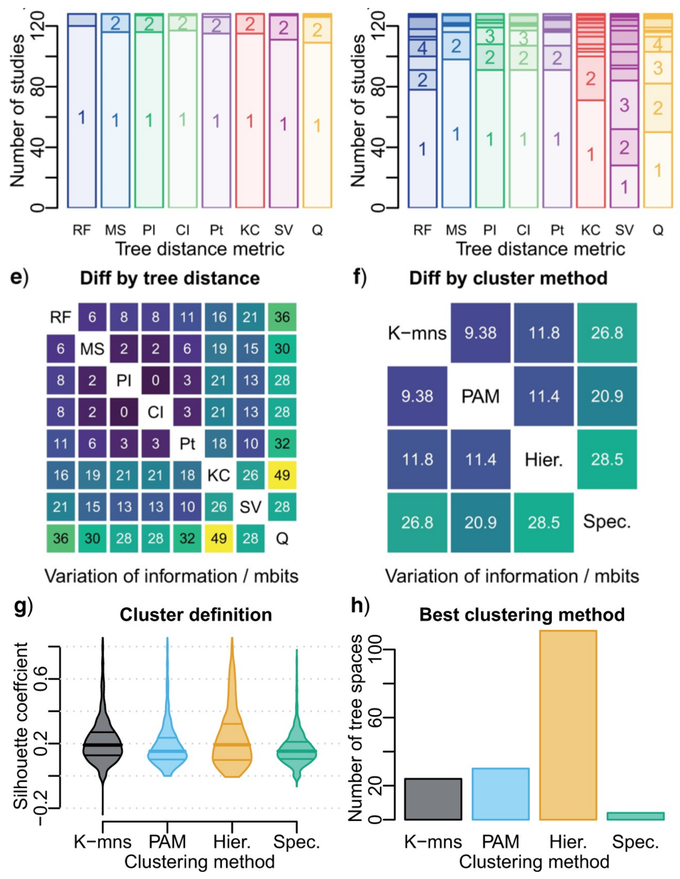
This study considers distances that purport to quantify the similarity of relationships between cladograms: the Robinson–Foulds, matching split information, phylogenetic information, clustering information, path, Kendall– Colijn,and quartet metrics, and a new metric derived from vector representations of trees.The quartet distance 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, using 3 random starts and the algorithmic shortcuts of Schubert and Rousseeuw.
- hierarchical clustering with minimax linkage,(Bien and Tibshirani 2011) (chosen after outperforming other linkage methods in initial informal analyses)





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 . PCoA requires Euclidean distances, and converting distances between phylogenetic trees into a Euclidean space entails a loss of information . To make the distances Euclidian, I follow the standard practice of adding a constant to each distance , while noting that this might distort the relative magnitude of individual distances. Finally,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

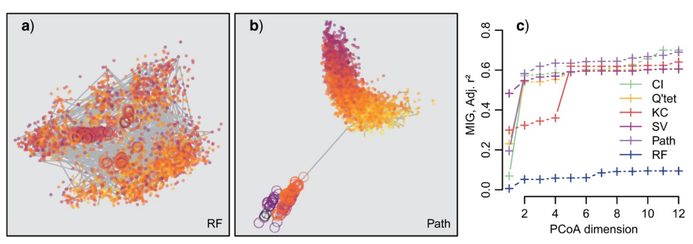
To evaluate the susceptibility of a tree space to distortion on mapping, I calculate its correlation dimension.That is, the number of dimensions necessary in order to reproduce all the structure present in the tree space.

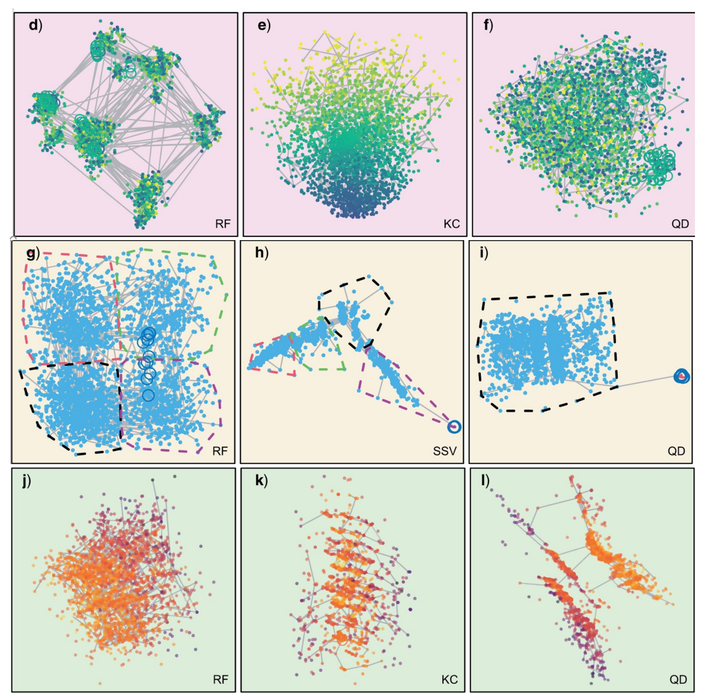
Results

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. Results obtained under the CI distance when trees were rooted do not materially differ from those when trees are treated as unrooted .

Tree Distance Metric

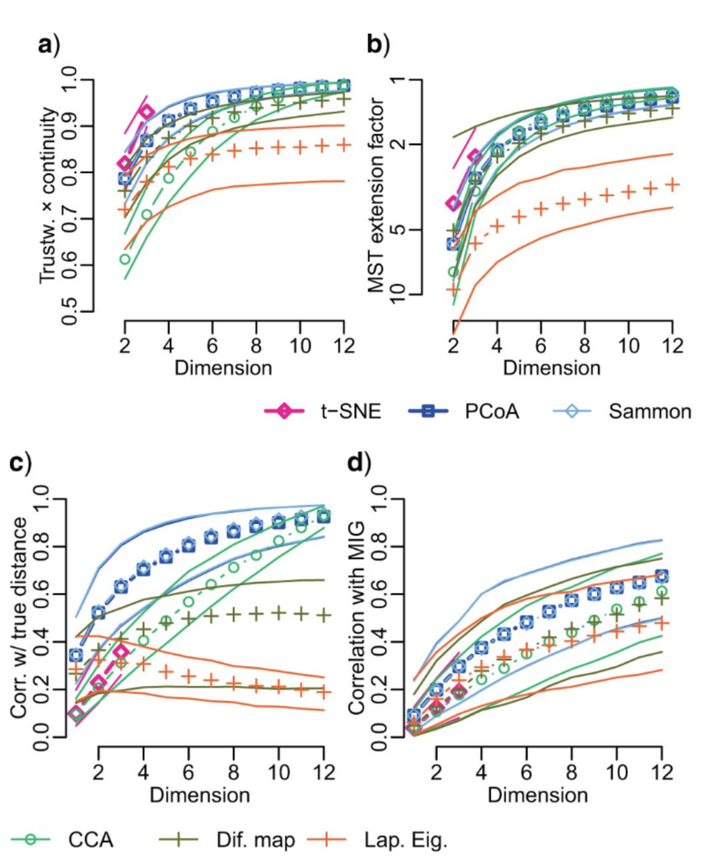
The results of the tests devised to compare tree distances by Smith , plus the new "balance independence" test, are presented in Table 1. Of the metrics examined, only the quartet and information-theoretic tree distances consistently reflect differences in the evolutionary relationships within trees .





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.



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] MARTIN R. SMITH. Alrobust analysis of phylogenetic tree space. *Oxford University Press*, pages 01–16, 2021.