Clustering gene trees into topological classes

Phylogenomics is the application of phylogenetic methods to genomic data to uncover the relationships between species. Not every genetic locus in a genomic dataset need support the same relationship, however. Effects such as incomplete lineage sorting, introgression and horizontal transfer can cause incongruence to occur between gene trees. Here we attempt to identify multiple topologies present in the data by applying hierarchical clustering to the inferred maximum-likelihood trees from each individual locus. Once clusters are identified the constituent loci are concatenated, and this is used to infer an overall cluster tree.

Hierarchical clustering data points requires two things: a metric for determining the distances between the data points, and a linkage criterion for assigning distances when data are merged to form clusters. We investigate which combination of tree distance metric and linkage criterion is most useful for identifying multiple topologies. We use the sum of log-likelihoods of the resulting cluster trees as a scoring function to assess the relative performance of each method.

We investigate three commonly used tree distance metrics: Robinson-Foulds, Weighted Robinson-Foulds and Felsenstein’s Branch Length Distance. We also vary the linkage method used between single-linkage, complete-linkage, UPGMA and Ward’s method. We test these approaches on simulated data, for which the underlying trees and the number of underlying classes are known. When choosing the number of clusters to use we show that plotting the score against the number of clusters used is a good indicator of the number of underlying classes in the data. On simulated data our method is successful in recovering the correct underlying substructure of the data, with the combination of Robinson-Foulds distance and Ward’s method giving the best performance.

We applied our method to a widely studied dataset of 106 genes from eight yeast species. This method shows the data to cluster into four similar classes, each resulting in a distinct class topology.