

Clustering Genes into Topological Classes

Kevin Gori,¹ Christophe Dessimoz,² Nick Goldman

1: email to kgori@ebi.ac.uk 2: Christophe Dessimoz is funded by the Swiss National Science Foundation

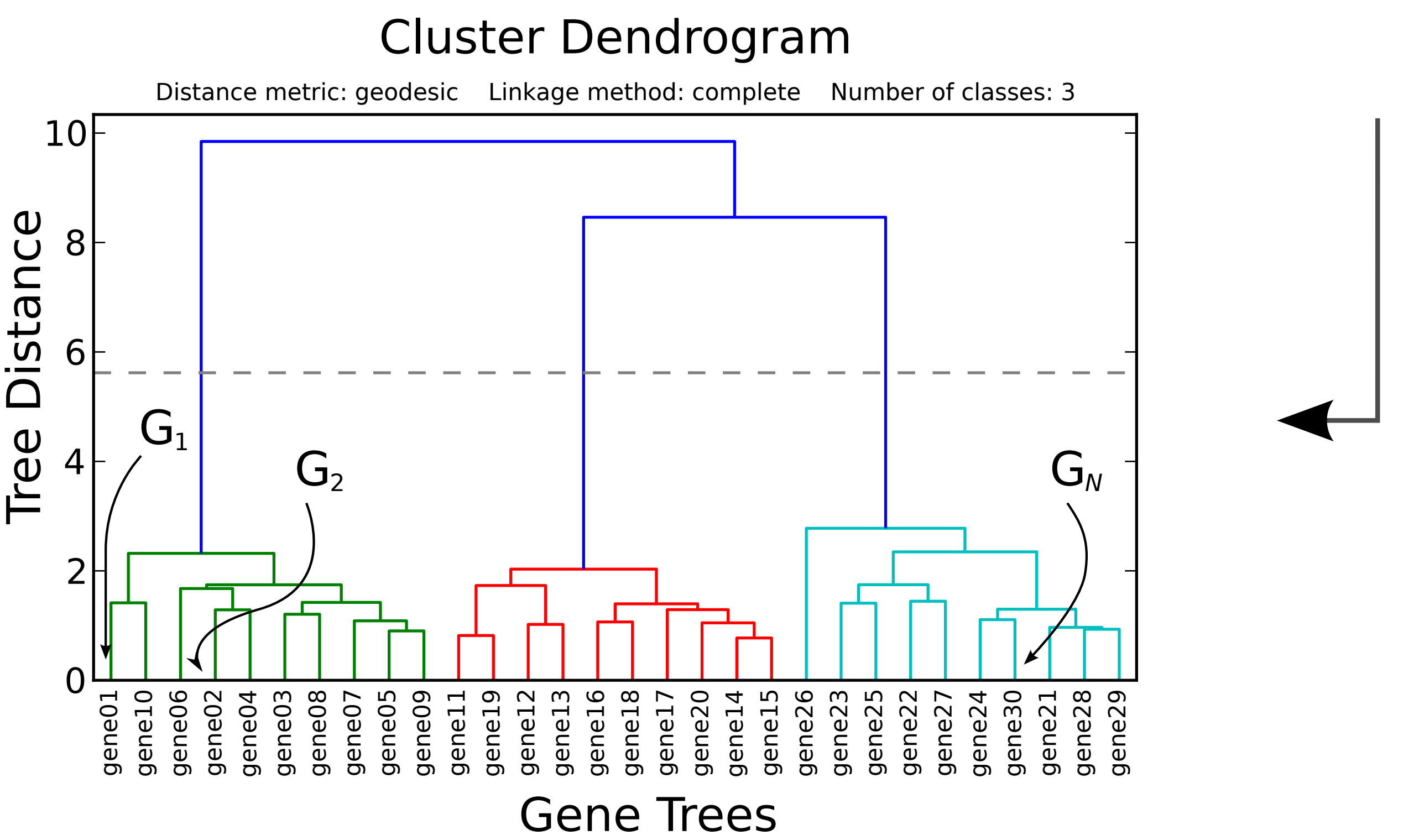
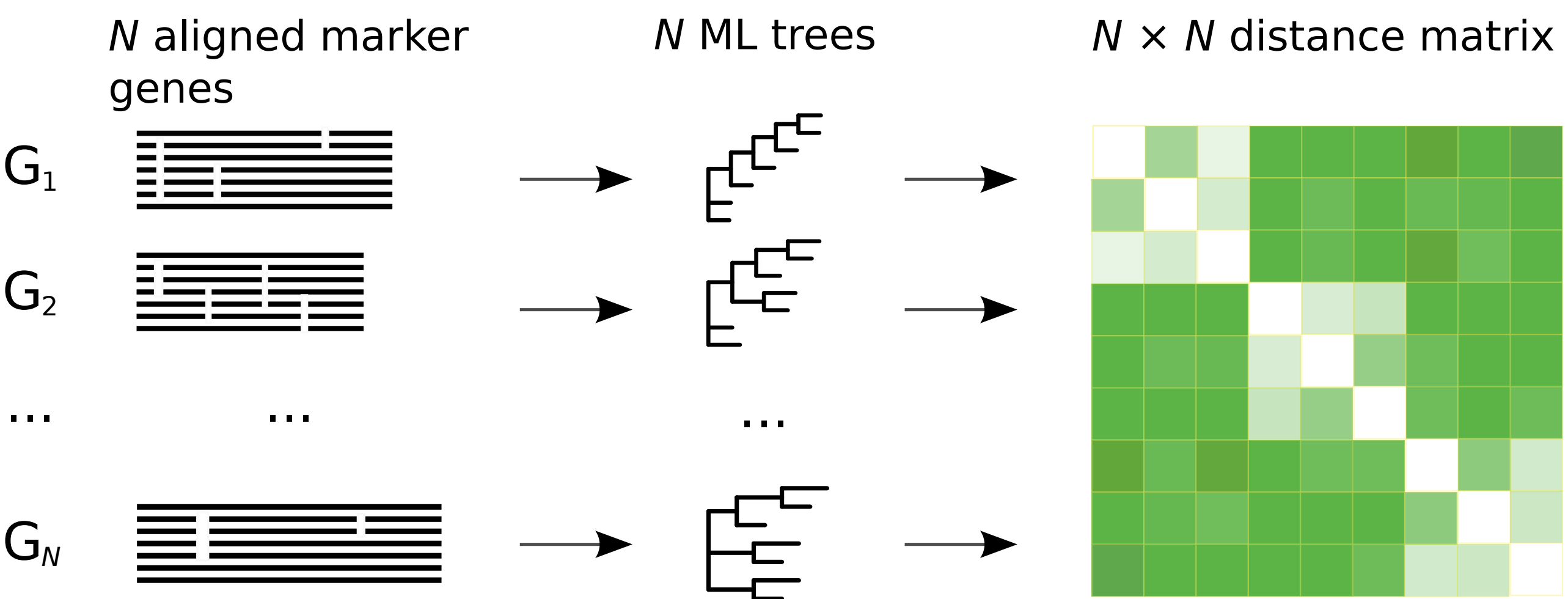


Outline

To uncover evolutionary relationships between species, we should not assume that all genetic loci in a genomic dataset support the same underlying tree topology. Effects such as incomplete lineage sorting, introgression and horizontal transfer can cause incongruence to occur between gene trees. Here, we explore ways to identify multiple topologies present in the data by clustering trees reconstructed from individual loci into classes with common underlying topologies.

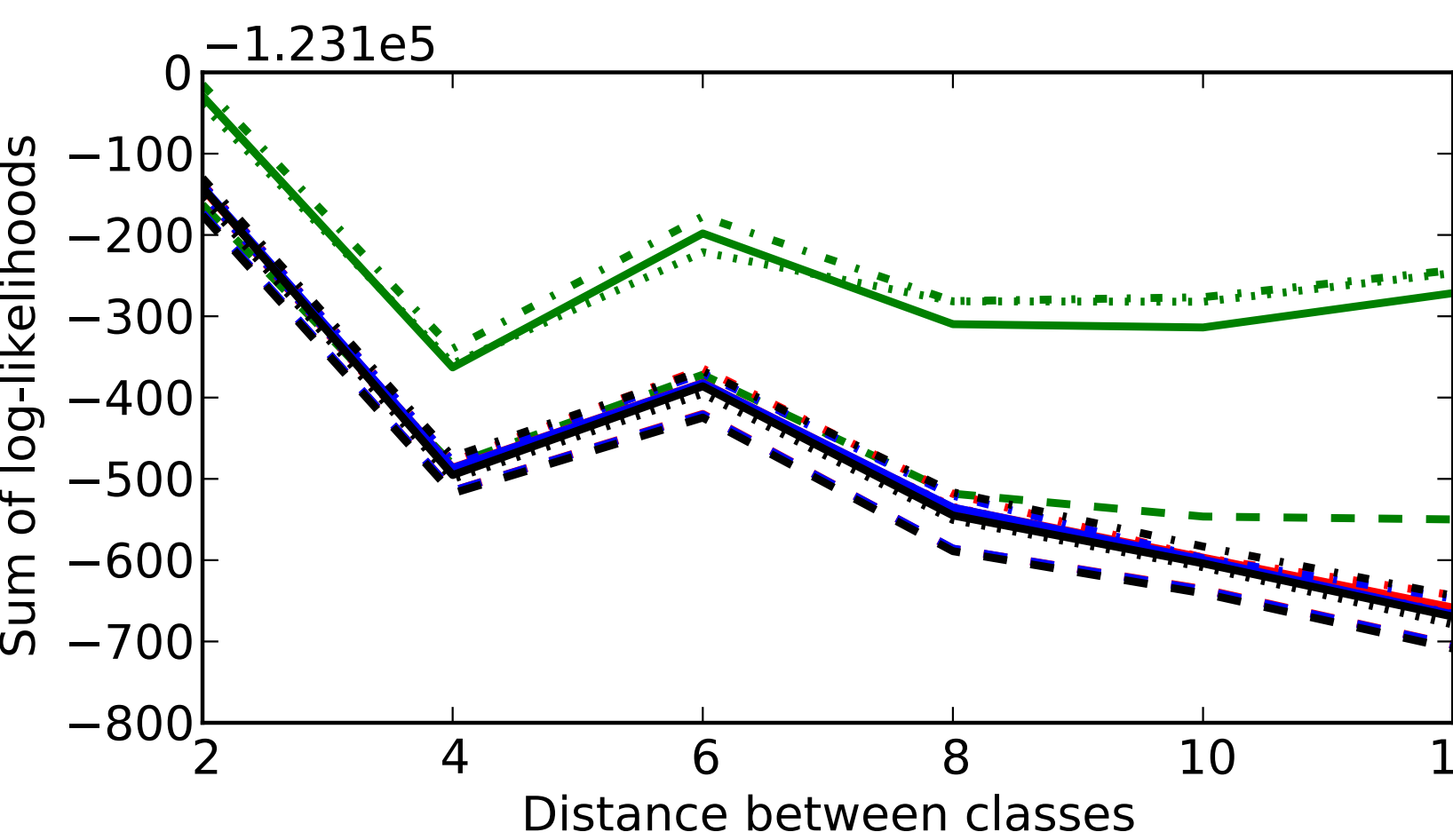
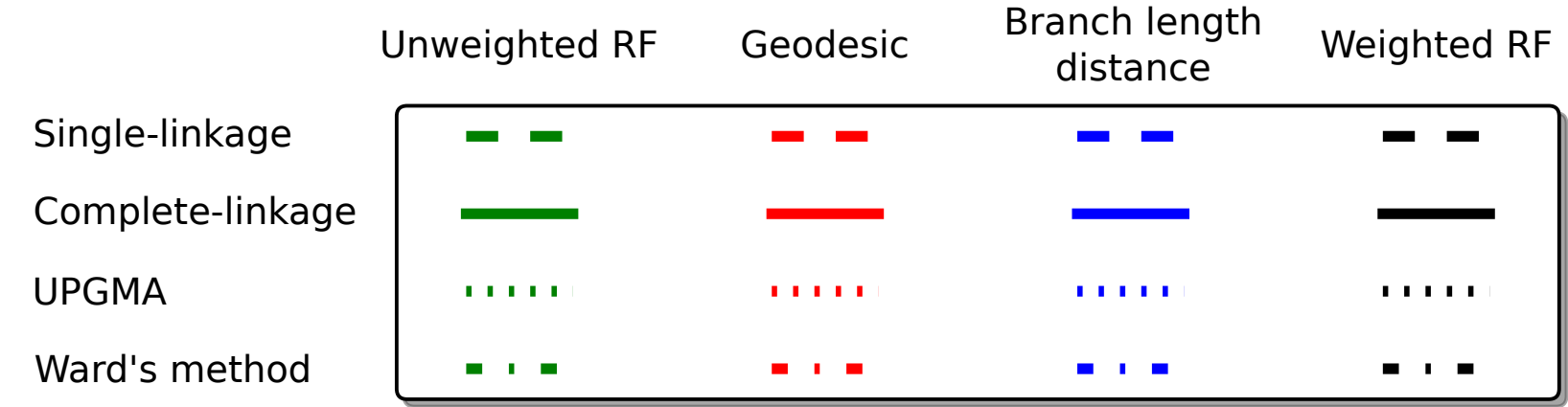
Tree distance and clustering methods

Our method uses the distances between the genes' inferred trees as a basis for hierarchical clustering. For each alignment the tree is inferred by maximum likelihood using PhyML. Distances between the trees are estimated according to one of a choice of commonly used tree distance metrics: the weighted and unweighted Robinson-Foulds distances (Robinson and Foulds, 1979, Robinson and Foulds, 1981), Felsenstein's Branch Length distance (Kuhner and Felsenstein, 1994), and the Geodesic distance (Billera, Holmes and Vogtmann, 2001). The resulting distance matrix is used as input to hierarchical clustering based on one of single-linkage, complete-linkage, UPGMA, and Ward's method.



Assessment of clustering performance

Using simulated data we assessed which combination of distance metric and clustering method was best able to recover underlying class structure.

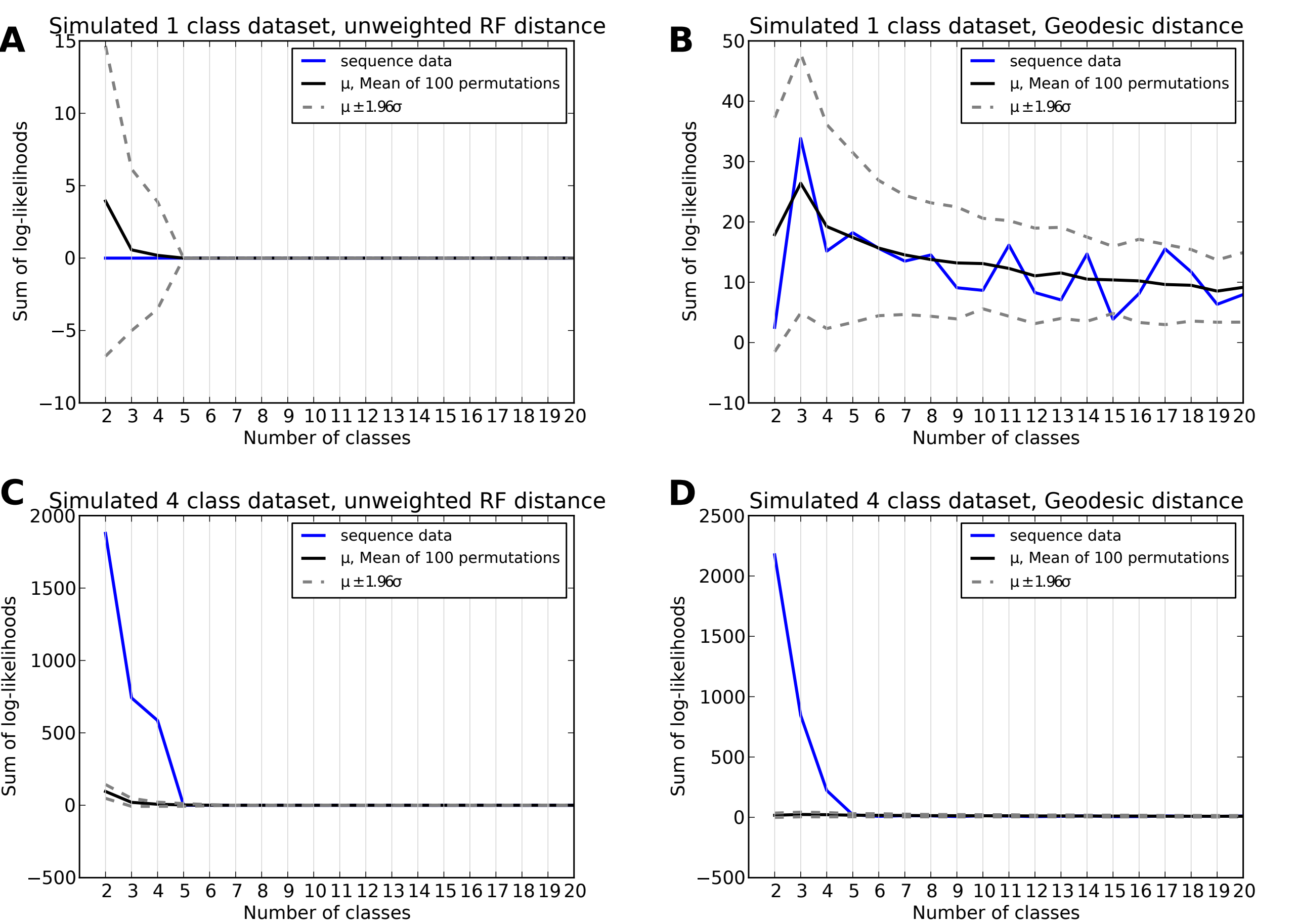


- Simulate 100 sets of 30 alignments from 3 trees; 3 topological classes of ten genes each.
- Cluster the alignments using all combinations of distance measures and clustering methods.
- Assess results by the sum of the likelihood scores for each class after tree inference with PhyML.
- Done 6 times for different distances between the underlying trees (unweighted RF distance), on the assumption that close trees represent a more difficult case.
- In this study the unweighted RF distance performs better than metrics that use the branch lengths. In each case, single-linkage clustering performs the worst, with there being little to choose between the other methods.

Determining the number of classes

Simulated Data

We used a permutation procedure to assess the number of classes present in a dataset. We simulated two datasets - 106 genes with 1 underlying topology, and 106 genes from 4 underlying topologies. For each dataset we also produced 100 permuted copies; all the aligned columns from all the alignments were shuffled, and partitioned back into 106 alignments to remove any class structure present in the dataset. We clustered the datasets and their permuted copies into 1 to 20 classes, and recorded their summed likelihood scores.

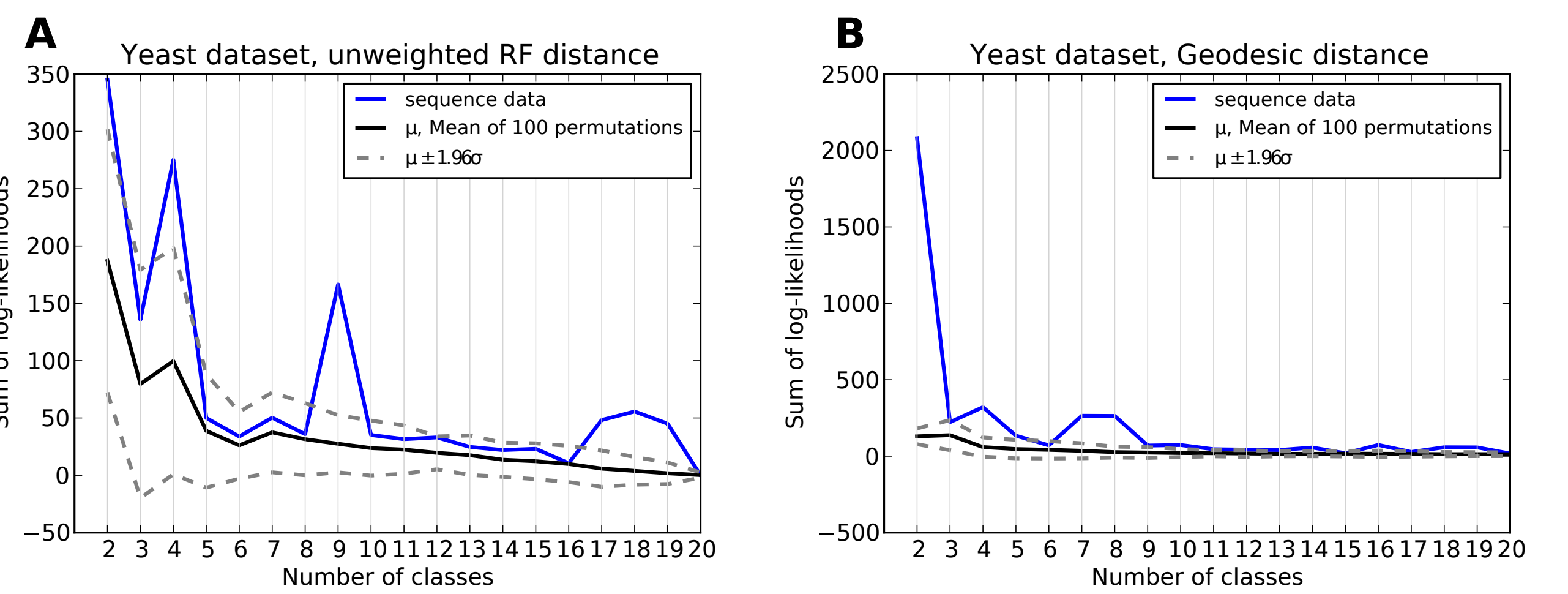


Plots of the difference in likelihood scores when moving from K - 1 classes to K
A: 1 underlying class, unweighted RF distance; B: 1 underlying class, geodesic distance; C: 4 underlying classes, unweighted RF distance; D: 4 underlying classes, geodesic distance.

These plots show the improvement in likelihood score when increasing the number of classes. When only one class is present the improvement in the score of the unshuffled data set is contained in the distribution of the shuffled copies. When four classes are present the improvement remains above the shuffled distribution for 2, 3 and 4 classes, and falls in amongst for 5 classes or greater, suggesting the correct number of classes is 4.

Real Data

We analysed a set of 106 genes from yeasts (Rokas et al., 2003) using this hierarchical clustering procedure, and used the permutation approach to investigate the number of classes present.



Difference plots for the yeast dataset
A: Unweighted RF distance; B: Geodesic distance.

These data appear to show some class structure, but it is not clear from this approach how many classes are present.

This work is a preliminary investigation into the possibility of explaining phylogenetic data with more than one tree. Our approach works well on the simulated cases we have examined, although more scenarios must be looked at before we can be confident of our method. The analysis of the yeast dataset has shown that real data is more complex than simulated data. Future work will tackle the problem of choosing the number of classes using a Bayesian model-selection approach, and explore alternative clustering strategies.

References

Billera, Holmes and Vogtmann, 2001. "Geometry of the Space of Phylogenetic Trees." *Advances in Applied Mathematics* 27 (4) (November): 733-767.
Kuhner and Felsenstein, 1994. "A Simulation Comparison of Phylogeny Algorithms Under Equal and Unequal Evolutionary Rates." *Molecular Biology and Evolution* 11 (3) (May): 459-468.
Robinson and Foulds, 1979. "Comparison of Weighted Labelled Trees." *Lecture Notes in Mathematics* Vol. 748.
Robinson and Foulds, 1981. "Comparison of Phylogenetic Trees." *Mathematical Biosciences* 53 (1-2): 131-147.
Rokas, Williams, King and Carroll, 2003. "Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies." *Nature* 425 (6960): 798-804.