Applications of Clustering and Visualization to Phylogeny

Ruchi Kalra and Katherine St. John
Dept. of Mathematics & Computer Science
Lehman College– CUNY
Bronx, NY 10468
{kalrr192,stjohn}@lehman.cuny.edu

Luay Nahkleh and Tandy Warnow
Dept. of Computer Sciences
University of Texas
Austin, TX 78712
{nakhleh,tandy}@cs.utexas.edu.

[**Keywords:** clustering, visualization, phylogeny.]

Abstract

Phylogenic trees play a major role in representing the interrelationships among biological entities. While trees are effective for many biological processes, processes such as hybridization and horizontal gene transfer result in networks of relationships rather than trees of relationships. Few methods have been developed for inferring phylogenetic networks, or for the simpler problem of determining that the hybridization has occurred. In this poster, we describe new techniques for detecting hybridization, based upon visualization and clustering of sets of phylogenies. These techniques show promise for inferring the underlying network model when hybridization has occurred.

1. Introduction

Understanding the evolutionary history of species, or of biomolecular sequences, is an important part of much biological research. Current phylogenetic reconstruction methods focus on the reconstruction of trees, rather than more complex models of speciation. A large proportion of the speciation is not treelike [5, 7], and accurate representations of their evolutionary histories will require networks rather than trees. In addition, emerging biological evidence indicates that different chromosomes and different parts of chromosomes in a single species may have different evolutionary histories [10] One example of this is the presence of bacterial and viral sequences in the human genome [16]. In such an evolutionary scenario, two species must be able to combine their genomes to produce new species (hybrid speciation), and species must be able to contribute genetic material to other species (via horizontal gene transfer and introgression).

In hybridization, two lineages recombine to create a new species. The true evolutionary history is best represented by gene trees in a phylogenetic network, or *directed acyclic graph*, rather than by a tree.

2. Prior Work

There are roughly two approaches for detecting network evolutionary mechanisms, such as hybridization, and reconstructing appropriate evolutionary histories when these mechanisms occur. The first class of approaches uses *Combined Analysis* and allows the input dataset to have its evolutionary history be truly a network. A method for detecting horizontal gene transfer is given in [8], while SplitsTree [2], Median Networks [3] and Tcs [15] are generic methods for inferring graphs, rather than trees, from either sequence data or distance data. With the exception of [8], these methods do not explicitly refer to any underlying model of network evolution. The second class uses the *Separate Analysis* approach. The separate analysis approach separates the sites within the biomolecular sequence dataset into subsets, so that within each subset the sites all evolve down a single tree within the true phylogenetic network. Because short sequences can cause data analysis problems, biologists often have looked for indications that it is safe to "combine" (i.e. concatenate) sequence datasets to get longer sequences, and a number of statistical tests exist for determining whether such a combination is "safe" (e.g. [6, 11, 14]). Roughly speaking, these tests seek to determine whether the two datasets come from the same underlying evolutionary tree. A popular test is the Incongruence Length Difference Test (ILD)) [6] which measures the likelihood that the combined sequences evolved from the same underlying evolutionary tree.

3. Visualizing Hybrid Events

We studied the effectiveness of clustering and visualization to detect hybrid events that have occurred. From our visualization, it was surprisingly easy to distinguish when

a events occurred. The popular k-agglomerative clustering was also effective tool. We used PAUP* [13] for our Maximum Parsimony analyses, as well as for the ILD. The resulting set of trees for the set of sequences A, for the set of sequences B, and the set of concatenated sequences of A and B, are called T_A, T_B, and T_{AB} below.

For visualizing large sets of trees, we used our Tree Set Visualization module [1] which runs under Mesquite [9]. Mesquite is a java-based framework for phylogenetic analysis written by Wayne and David Maddison. The results were generated with version 0.992 of Mesquite [9] and version 2.0 of the visualization module (both freely available). To display large sets of tree on the screen in a meaningful way, a standard technique, Multidimensional Scaling (MDS) [4] was used. MDS seeks to minimize the "stress" between the true distances between trees and the distances displayed on the screen by repeated incrementing the displayed distances by small "steps" (see [1] for details of our implementation). MDS, like other heuristic searches, can get stuck on on local optima, and our implementation allows the user to "scramble" the initial displayed distances and search again. It has worked extremely well in practice, grouping trees into clusters that have highly resolved consensus trees.

To automate the clustering process, we also implemented a k-agglomerative clustering with single and complete linkage. The implementation allows the optimization criteria to be customized for the data and desired goals (i.e. clusters that minimize diameter, maximize resolution of the strict consensus tree, etc) (see [12] for cluster quality statistics for evolutionary trees). For this initial study, we used the criterion of maximizing the minimum distance between clusters; this worked well in approximating the clustering defined by the input set of trees, and of the MDS defined clustering. We observed that these trees clearly clustered into two separate clusters in our MDS visualization (and these two clusters were centered around the true trees underlying the network). This showed that the two separate analyses of the data (one for each part) would yield a reasonable estimate of the underlying tree from which it evolved. On the other hand, if we concatenated the datasets and *then* performed a phylogenetic analysis, the resultant set of trees (i.e. T_{AB}) would *not* form two clear clusters.

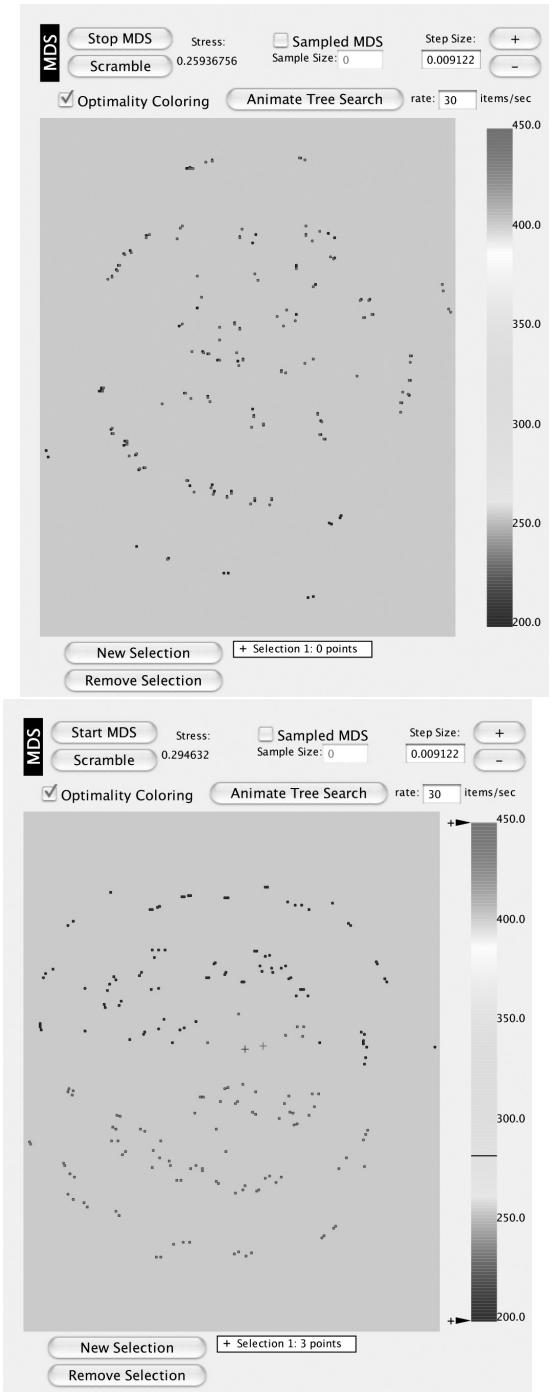

Figure 1. Two sets of trees $T_A \cup T_B$ for sequence length 2000. The lighter and darker shades of gray correspond to trees from the different sets. Left: Network contained zero hybrid events (i.e. a tree), and the underlying sequences that generated the trees have an ILD P-value of 1.0, indicating very high compatibility. The ILD value correctly predicts the tree-like structure. Note many trees occur in both sets and appear medium gray in the picture. Right: Network contained one hybrid event. The underlying sequences have an ILD P-value of 0.28, incorrectly indicating compatibility.

4. Acknowledgments

We thank Jeff Klingner for the tree set visualization module, Wayne and David Maddison for Mesquite, Randy Linder for helpful discussions and insights. This project was supported by NSF-ITR 0121651/0121682.

References

[1] N. Amenta and J. Klingner. Case study: Visualizing sets of evolutionary trees. In *8th IEEE Symposium on Information Visualization (InfoVis 2002)*, pages 71–74, 2002.

[2] H.-J. Bandelt and A. W. M. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, 1:242–252, 1992.

[3] H.-J. Bandelt et al. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phyl. Evol.*, 16:8–28, 2000.

[4] I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Springer, 1997.

[5] N. Ellstrand et al. Distribution of spontaneous plant hybrids. *PNAS*, pages 935090–935093, 1996.

[6] J. S. Farris, M. KŁllersj, A. G. Kluge, and C. Bult. Testing significance of incongruence. *Cladistics*, 10:315–319, 1994.

[7] V. Grant. *Plant Speciation*. Columbia University Press, New York, 1971.

[8] M. T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proc. 5th Ann. Int'l COnf. Comput. Mol. Biol. RECOMB2001*, New York, 2001. ACM Press. To appear.

[9] W. P. Maddison and D. R. Maddison. Mesquite: a modular system for evolutionary analysis. version 0.992, 2002. Available from `http://mesquiteproject.org`.

[10] L. Rieseberg and C. Linder. Hybrid classification: Insights from genetic map-based studies of experimental hybrids. *Ecology*, 80:361–370, 1999.

[11] A. G. Rodrigo et al. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *NZ J Botany*, 31:257–268, 1993.

[12] C. Stockham et al. Statistically based postprocessing of phylogenetic analysis by clustering. In *Proceedings of 10th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB'02)*, pages 285–293. Edmonton, Canada, 2002.

[13] D. L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.

[14] A. R. Templeton. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution*, 37:221–244, 1983.

[15] A. R. Templeton, K. A. Crandall, and C. F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data iii: Cladogram estimation. *Genetics*, 132:619–633, 1992.

[16] J. C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.