Clustering Trees into Topological Classes
Project Plan

Kevin Gori

July 19, 2012

Supervisors: Nick Goldman, Christophe Dessimoz

1 Objective

Phylogenetic inference on modern datasets containing large numbers of genes from a group of taxa will in most cases produce discordant trees. This project aims to identify and characterise this discordance by applying a variety of clustering methods to a matrix of tree distances derived from single-gene tree inferences.

2 Principles

2.1 Introduction

Phylogenetic inference from molecular sequences aims to provide a description of the evolutionary history of the marker sequences being analysed. The tree produced summarises two aspects of the evolutionary process: the branching order of the tree corresponds to the order of duplications of the marker sequences, and the branch lengths reflect the process of evolution along the branch – they estimate the evolutionary rate according to the model, multiplied by the time spent evolving. The tree describes the history of the marker sequences used, but if we are investigating species history we may wish to conflate the marker's branching history with the history of speciation events. However, there are many reasons why a segment of DNA can have a different history to the organism it resides in, including population genetic events such as horizontal gene transfer, recombination, hybridisation and incomplete lineage sorting, and stochastic variation in the substitution process which results in the segment evolving differently by chance. This kind of gene-tree, species-tree incongruence makes single marker phylogenies unreliable for estimating species trees.

To get a better estimate of the species history we can use several markers. An increased sample size should allow us to average out some of the stochastic variation, and we can hope that the speciation signal is present in enough of the markers that we can detect it as a common theme. Similarly, markers sharing non-speciation history can also be grouped together. Groups of markers sharing history can be found by clustering them according to their topological similarity, and by determining the number and relative sizes of the clusters found we can characterise the amount of discordance in the dataset. The process generating the discordance is not specified, but some inferences can be made, for example if markers from a particular class are found in a contiguous block in the genome, then we may be able to infer recombination breakpoints in their vicinity.

2.2 Modelling multiple classes

Suppose we have a dataset consisting of M alignments on n species. We want to estimate the number of classes, K, and for each class $k_{1 \leq i \leq K}$ estimate the underlying tree with $2n-3$ branch lengths. We can model this situation with increasing complexity:

Model	**Topologies**	**Parameters**		**Notes**
		Branch lengths	Scale factors (τ)	
Supermatrix	1	$2n-3$	0	All markers are generated by the same topology
Multiclass Supermatrix	K	$K(2n-3)$	0	Markers are generated by K class topologies; each has the same underlying branch lengths
Proportional Branch Lengths	K	$K(2n-3)$	$M-K$	Markers are generated by K class topologies; each marker within the class has a scaled set of branch lengths
General Branch Lengths	K	$M(2n-3)$	0	Markers are generated by K class topologies; each marker has independent branch lengths

Table 1: General modelling scheme

3 Methods

3.1 Tree Inference

I align sequences using some alignment method (for example Prank (Löytynoja and Goldman, 2005)), and then infer trees using either a Maximum-Likelihood approach, for which I use PhyML (Guindon et al., 2010), or a distance based approach, TreeCollection, for which distances and variances between sequences in an alignment are estimated using Darwin, with the resulting distance-variance matrices being the input for TreeCollection, which uses a Least-Squares approach to reconstruct the tree.

3.2 Estimating Tree Distances

As a way of avoiding representing trees as vectors of features observed in a defined space I use pairwise distances to construct a distance matrix as a basis for clustering. A number of distance metrics are available for comparing phylogenetic trees. I chose to use the following (not an exhaustive list of all metrics available).

- Distance Metrics:
  - Robinson-Foulds Distance (Robinson and Foulds, 1981)
  - Euclidean Distance (the Branch Length Distance of Kuhner and Felsenstein (1994))
  - Geodesic Distance (Billera et al., 2001)

The first two of these measures are calculated using the Python module DendroPy (Sukumaran and Holder, 2010). The fourth is calculated using a modified version of the GeoMeTree program (Kupczok et al., 2008). The unweighted Robinson-Foulds measure is based only on topology, the Branch Length distance is based only on branch lengths and the Geodesic distance is explicitly based on topology and branch lengths.

3.3 Clustering Methods

Clustering is a process of separating data points into groups, the members of which share common characteristics more than they share the characteristics with members of other groups. There is a large number of methods available for doing clustering.

1. Hierarchical Clustering Methods
   - Single-linkage
   - Complete-linkage
   - Average-linkage
   - Ward-linkage
2. Exemplar Methods
   - K-medoids (k-centres)
   - Affinity Propagation
3. Matrix Decomposition Methods
   - Classical Multidimensional Scaling + K-means
   - Spectral Clustering + K-means

3.3.1 Hierarchical Clustering

Possibly the simplest methods which can act directly on distance matrices are (agglomerative) hierarchical clustering. These methods iteratively combine the two closest points, representing them by a single point. How this point is taken to relate to the remaining points is defined by the linkage method used. Let u be the newly defined cluster formed by combining points s and t, and v be any other point (or cluster) in the dataset, and let $d(u,v)$ be the distance between u and v:

- Single-linkage - closest member point

$$d(u,v) = \min_{i \in u, j \in v}(d(i,j))$$

- Complete-linkage - furthest member point

$$d(u,v) = \max_{i \in u, j \in v}(d(i,j))$$

- Average-linkage - mean of member distances

$$d(u,v) = \frac{1}{|u|\,|v|}\sum_{i \in u}\sum_{j \in v} d(i,j)$$

- Ward-linkage - minimum variance

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 + \frac{|v|}{T}d(s,t)^2}$$
$$T = |v| + |s| + |t|$$

3.3.2 Exemplar Methods

Other methods acting directly on distance matrices are k-medoids (also known as k-centres), and affinity propagation (Frey and Dueck, 2007). These search for points in the data set to act as exemplars, that is cluster centres to which surrounding points are grouped. K-medoids requires that the number of clusters be set in advance, whereas with affinity propagation the number of clusters depends on the value of a tuning parameter, called the preference, which can take a value between $\min(d)$ and $\max(d)$, and is usually set as the median.

3.3.3 Matrix Decomposition Methods

Classical multidimensional scaling maps distances to points in a multidimensional space. The distance matrix is double-centred (from each element subtract the row mean, subtract the column mean, add the overall mean and divide by -2). This eigenvectors of this matrix - the Gower matrix - corresponding to the k largest eigenvalues map the data into a k-dimensional space. The points can then be clustered using k-means.
Spectral decomposition is carried out using the Ng-Jordan-Weiss (NJW) algorithm (Ng et al., 2002) with local scaling approach of Perona and Zelnik-Manor (2004).

4 Investigating the clustering procedures

References

Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

B J Frey and D Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, February 2007.

Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–321, May 2010.

M K Kuhner and J Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3):459–468, May 1994.

Anne Kupczok, Arndt Von Haeseler, and Steffen Klaere. An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees. *Journal of Computational Biology*, 15(6):577–591, July 2008.

Ari Löytynoja and Nick Goldman. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10557–10562, July 2005.

A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

P Perona and L. Zelnik-Manor. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17:1601–1608, 2004.

DF Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

J Sukumaran and M T Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26 (12):1569–1571, June 2010.