Bioinformatics – Lecture Notes

Announcements

Class 16 – March 19, 2002 -

A. Phylogenetic Trees

a) ultrameric trees

Last time we discussed the *unweighted paired group method with arithmetic mean* (UPGMA). Another similar method is the *weighted paired group method with arithmetic mean* (WPGMA). The only difference between these two methods is to weight the clusters by their sizes. Note the differences between step d) in algorithms 4.1 and 4.2.

Example – Figure 4.3 and Table 4.1
Note – these satisfy the 3-point condition (ultrametric)

Figure 4.4 – Output of UPGMA

b) additive metric

Example – Figures 4.5 and 4.6
Note – these satisfy the four point conditions (additive metric)

You cannot use UPGMA (or WPGMA) to find the clustering for this type of data. You can use the *Farris Transformed Distance Method* given in Theorem 4.5, which yields the transformed distance matrix given in Fig 4.8.

Example – Compare the incorrect (Fig 4.7) and correctly clustered (Fig 4.9) trees.

Note – If we do not know the ancestor we can improvise by using the *external reference* or *outgroup* at in place of the ancestor. The outgroup is the taxon with the farthest average distance from all other taxa.

In order for these methods to work, we must have the correct tree topology. This can be found by using various different approaches, namely, parsimony (minimize the number of substitutions), tree distance (align trees instead of sequences similar to sequence alignment distance), neighbor joining, and maximum likelihood methods (which we mentioned earlier).

For determining tree distance, there are methods that are easier than alignment. One such method involved computing the *root mean square deviation* (RMSD) between two trees.

$\sqrt{\frac{2\sum_{1<i<j<n}(d_{i,j}-e_{i,j})^2}{n(n-1)}}$ where $d_{i,j}$ and $e_{i,j}$ are the entries in the distance matrices that generated the two trees.

An alternative method involved statistical measures of deviation between two trees (See work of Fitch Margoliash and Farris for details). These methods are typically used if the true phylogenetic tree is known and we want to compute the distance of a calculated tree from this true tree.

c) estimating branch lengths

All the previous methods distance methods used assumed a constant rate of evolutionary change. If the rate of evolutionary change is not constant (which it probably is not), this introduces a large amount of stochastic error into the distance measurements. In this instance, UPGMA works well in giving the best approximation to tree topology.

To consider a model in which there is not a constant rate of evolution we turn to the work of Fitch and Margoliash who developed a methods for estimating branch lengths under this new model (Algorithm 4.3)

B. Maximum Likelihood methods for phylogenetic trees