Bioinformatics – Lecture Notes

Announcements

Next week is Spring Break (March 11-15). NO CLASS.

Additional Reference on phylogenetic trees

> *Molecular Evolution: A Phylogenetic Approach*
> Roderic D. M. Page and Edward C. Holmes
> Blackwell Science 1998.

Class 16 – March 7, 2002 -

Question: If an elephant steps on a triangle in the jungle, what happens to the musician to whom it belongs?

Evolutionary Problems – Some more information about phylogenetic trees.

Question: What do we do with phylogenetic trees?

1) measuring evolutionary change on a tree

   If the leaves of a tree each signify a sequence, the sum of the weights of the edges gives the evolutionary distance between the two sequences.

2) molecular phylogenetics

   Convert information in sequences into an evolutionary tree for those sequences.

3) Examples of trees and tree contstruction

4) Cluster methods vs search methods

   There are two basic methods for constructing trees. *Cluster methods* use an algorithm (set of steps) to generate a tree. These methods are very easy to implement and hence can be computationally efficient. They also typically produce a single tree. A big disadvantage to this method is that it depends upon the order in which we add sequences to the tree. Hence, there could be a different tree that explains the data just as well.

   *Search method*s use some sort of optimality criteria to choose among the set of all possible trees. The *optimality criteria* gives each tree a score that is based on the comparison of the tree to data. The advantage of search methods is that they use an explicit function relating the trees to the data (for example, a model of how the sequences evolve). The disadvantage is that they are computationally very expensive (NP complete problem).

5) Question: How do we compare different tree methods?

a) efficiency –How fast is the method?
b) power –How much data does the method require?
c) consistency – Will the tree converge on the right answer give enough data?
d) robustness – Will minor violations of the method's assumptions result in poor estimates of phylogeny?
e) falsifiability – Will the method tell us when its assumptions are violated?

5) How do assign weights for the edges of our trees?

a) *Distance methods* first convert aligned sequences into a pairwise distance matrix then input that matrix into a tree building method.
b) *Discrete methods* consider each nucleotide site (of some function of each site) directly.

Example of parsimony (a discrete method) vs minimum evolution (a distance method)

c) The major objections to distance methods are that summarizing a set of sequences by distance data loses information and branch lengths estimated by some distance methods might not be evolutionarily determinable
d) Two other distance methods are **neighbor joining** and **minimum evolution.**
    i) *Minimum evolution* finds the tree that minimizes the sum of the branch lengths where the lengths are calculated from the pairwise distances between the sequences. Linear programming or least squares methods can be used to do this.
    ii) *Neighbor joining* is a clustering method that is computationally fast and gives a unique result. This can use something like the *four-point condition* and clusters the closest elements.

That is given the leaves i, j, k, and l

$d(i,j) + d(k,l) \leq d(i,k) + d(j,l) = d(i,l) + d(j,k)$ (additive metric)

For an ultrameric metric the *ultrameric* or *3-point condition* holds

That is given the leaves i, j, and k

$d(i,j) \leq d(i,k) = d(j,k)$

e) The two major discrete methods are **maximum parsimony** and **maximum likelihood**. Both these are search methods.

   iii) With *maximum parsimony* we try to reconstruct the evolution at a particular site with the fewest possible evolutionary changes. The **advantages** of parsimony are that it makes relatively few assumptions about the evolutionary process, it has been studied extensively mathematically, and some very powerful software implementations are available. The major **disadvantage** to using parsimony is that under some models of evolution, it is inconsistent , that is if more data is added the wrong result might occur.

   iv) The *maximum likelihood approach* looks for the tree that makes the data the most probable evolutionary outcome. This approach requires a explicit model of evolution which is both a strength and weakness because the results depend on the model used. This method can also be very computationally expensive.

a) ultrameric trees

Clustering methods attempt to repeated cluster the data by grouping the closest elements together. They are used for phyogeny and gene expression micorarray analysis.

The *pair group method* (PGM) is a technique where the pairs are repeatedly amalgamated.

The *unweighted paired group method with arithmetic mean* (UPGMA) is used to cluster molecular data where sequence alignment distance between sequences has been determined in a distance matrix.

b) additive metric
c) estimating branch lengths