**Course: Advanced Bioinformatics**

**Module title: Phylogeny Tree Algorithm**

**Module no. : 46**

There are three methods used to calculate the tree(s) that best account for the observed variation in a set of sequences. These methods are maximum parsimony, distance, and maximum likelihood.

**Maximum Parsimony**

Maximum parsimony methods predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences. In order to construct a tree using maximum parsimony, a multiple sequence alignment must first be obtained. For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes are identified. This continues for each position in the alignment.

Those trees that produce the smallest number of changes overall for all sequence positions are identified. This is a rather time consuming algorithm that only works well if the sequences have a strong sequence similarity.

**Distance Methods**

The distance method for construction of phylogenetic trees looks at the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group. The goal of distance methods is to identify a tree that positions neighbors correctly and that also has branch lengths which reproduce the original data as closely as possible.

CLUSTALW uses the neighbor-joining method as a guide to multiple sequence alignments. The PHYLIP suite of programs employ neighbor-joining methods.

Phylip <http://evolution.genetics.washington.edu/phylip.html>

Distance analysis programs in PHYLIP

FITCH: estimates a phylogenetic tree assuming additivity of branch lengths using the Fitch-Margoliash method.

KITSH: same as FITCH, but under the assumption of a molecular clock.

NEIGHBOR: estimates phylogenies using the neighbor-joining (no molecular clock assumed) or unweighted pair group method with arithmetic mean (UPGMA) (molecular clock assumed).

For phylogenetic analysis, the distance score counted as either the number of mismatched positions in the alignment or the number of sequence positions that must be changed to generate the other sequence is used.

The success of distance methods depends on the degree to which the distances among a set of sequences can be made additive on a predicted evolutionary tree.

**Maximum likelihood**

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. The method requires a substitution model to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but because it formally requires search of all possible combinations of tree topology and branch length, it is computationally expensive to perform on more than a few sequences.

The "pruning" algorithm, a variant of dynamic programming, is often used to reduce the search space by efficiently calculating the likelihood of subtrees. The method calculates the likelihood for each site in a "linear" manner, starting at a node whose only descendants are leaves (that is, the tips of the tree) and working backwards toward the "bottom" node in nested sets. However, the trees produced by the method are only rooted if the substitution model is irreversible, which is not generally true of biological systems. The search for the maximum-likelihood tree also includes a branch length optimization component that is difficult to improve upon algorithmically; general global optimization tools such as the Newton-Raphson method are often used. Searching tree topologies defined by likelihood has not been shown to be NP-complete,[4] but remains extremely challenging because branch-and-bound search is not yet effective for trees represented in this way.