**Course: Advanced Bioinformatics**

**Module title: Likelihood**

**Module no. : 49**

Maximum likelihood is an approach that is designed to determine the tree topology and branch lengths that have the greatest likelihood of producing the observed data set. A likelihood is calculated for each residue in an alignment, including some model of the nucleotide or amino acid substitution process. It is among the most computationally intensive but most ﬂexible methods available. Maximum parsimony methods sometimes fail when there are large amounts of evolutionary change in different branches of a tree. Maximum likelihood, in contrast, provides a statistical model for evolutionary change that varies across branches. Thus, for example, maximum likelihood can be used to estimate positive and negative selection across individual branches of a tree. The relative merits of maximum parsimony and maximum likelihood continue to be explored. When sequences evolve in a heterogeneous fashion over time maximum parsimony can outperform 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, but remains extremely challenging because branch-and-bound search is not yet effective for trees represented in this way.