MOLECULAR PHYLOGENESIS

DR POONAM KUMARI
DEPT OF ZOOLOGY
M.SC SEMESTER II CC 02

Phylogenetics is the science of estimating and analyzing evolutionary relationships. Phylogenetic relationships among micro-organisms are especially difficult to discern. Molecular biology often helps in determining genetic relationships between different organisms. Nucleic acids (DNA and RNA) and proteins are 'information molecules' in that they retain a record of an organism's evolutionary history. The approach is to compare nucleic acid or protein sequences from different organisms using computer programs and estimate the evolutionary relationships based on the degree of homology between the sequences. Nucleic acids and proteins are linear molecules made of smaller units called nucleotides and amino acids, respectively. The nucleotide or amino acid differences within a gene reflect the evolutionary distance between two organisms. In other words, closely related organisms will exhibit fewer sequence differences than distantly related organisms. In particular, the sequence of the small-subunit ribosomal RNA (rRNA) is widely used in molecular phylogeny.

One advantage of the molecular approach in determining phylogenetic relationships over the more classical approaches, such as those based on morphology or life cycle traits, is that the differences are readily quantifiable. Sequences from different organisms can be compared and the number of differences can be established. These data are often expressed in the form of 'trees' in which the positions and lengths of the 'branches' depict the relatedness between organisms. Shown below is a three-domain tree of life based on small subunit rRNA sequences.

Phylogenetic Tree

A phylogenetic tree or phylogeny is a tree-like diagram used to visualize evolutionary relationships among a set of operational taxonomic units (OTUs). The OTU generally represents a species, but can also represent individual organisms in a population, a gene or protein sequence or a taxon at any taxonomic rank. The tree is composed of nodes and branches (Fig. 1). Nodes at the tips of the tree are called ‘external nodes.’ These are used to represent the OTUs. Another type of node, called ‘internal nodes,’ represents a recent common ancestor (RCA). Between these are lines, called ‘branches,’ used to connect newer and older nodes and show

the evolutionary relationships among the taxa. A branch linking two internal nodes is an ‘internal branch,’ which shows an ancient relationship. Conversely, the branch joining an internal node with an external node to show a modern relationship is called an ‘external branch.’

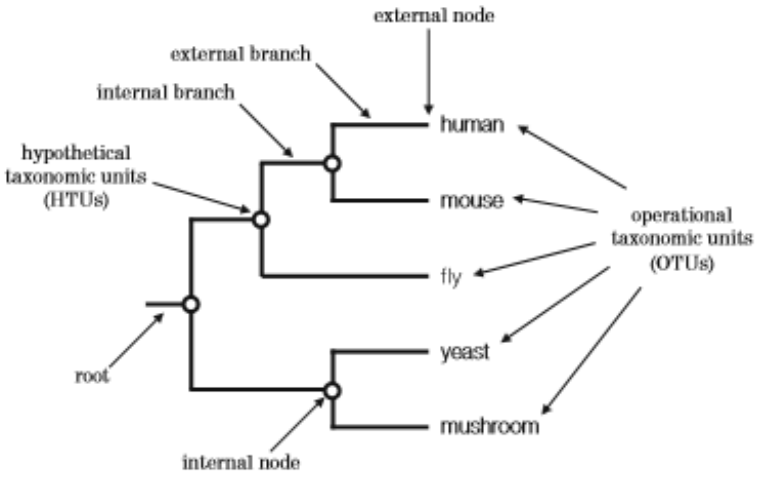

Fig. 1 Composition of a phylogenetic tree. Terminology frequently used in phylogenetic trees is labeled on the tree

Molecular Markers

DNA sequences have been accepted and widely used as molecular characters for phylogenetic tree reconstructions, surpassing the use of morphological characters. This is because the sequence states of DNA, which can be only adenine, thymine, cytosine, or guanine, are clearer than morphological states. Molecular sequences also provide a large number of characters for phylogenetic analysis. For example, a phenotype regulated by single gene or a group of genes can be recognized as one character, but almost all positions in a gene’s DNA sequence are useful characters for phylogenetic analysis.

Sequence Alignment

DNA and protein sequences are the most frequently used data types in molecular phylogenetic analysis. To study deep phylogeny, one needs ancient, universal, orthologous sequences to form the dataset. However, these sequences might be very diverse and may not align properly. To circumvent this problem, protein sequences are a better choice. This is because mutations appear to have fewer effects on protein sequences. On the other hand, the study of recent evolution or

phylogenetic analysis of OTUs within the same species needs DNA sequences, which are less conserved in their sequences than are proteins. Moreover, the analysis of non-coding sequences can be carried out on DNA sequences only.

Molecular phylogenetic analysis relies heavily on the accuracy of the sequence alignment. The programs used for the alignment of sequences are developed from several algorithmic approaches. One of the most popular algorithms is 'progressive sequence alignment,' which has been implemented in several software packages.

Phylogenetic Reconstruction

Methods for phylogenetic reconstruction can be classified into two main approaches: distance-based methods and character-based methods. The concept behind the former is the transformation of all sequence information into a distance matrix, which is then analyzed using an algorithm for clustering the taxa. Building a tree with this method is fast but all sequence information is lost in the process. The latter method is time-consuming because all the sequence information is used for the evaluation of the best phylogenetic tree.

Distance-Based Approach

The key concept behind distance matrix methods is the conversion of a pairwise sequence alignment into distant values. Because a multiple sequence alignment (MSA) must contain three or more sequences, distance values from all possible pairwise sequences generate a distance matrix. Once a matrix is developed, the alignment is no longer used for the phylogenetic reconstruction. At this point, the matrix is used as the input for the tree building. Different tree building approaches used include the unweighted pair group method with arithmetic mean (UPGMA), weighted pair group method with arithmetic mean (WPGMA), neighbor-joining (NJ), least square (LS), and minimum evolution (ME) methods.

Character-Based Approach

There are several methods that have been developed from character-based approaches, such as maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference methods. These approaches aim to reconstruct a phylogeny directly from the sequence data, without any transformation. They make extremely slow calculations but the final tree is said to be very accurate. Briefly, the

algorithm used in these begins with scoring all possible phylogenies that can be generated from the n taxa.

Conclusion

Phylogenetic analysis is one of the important techniques in the networking biologist's toolbox. It can be used to identify the evolutionary relationships among organisms, as well as gene or protein sequences. To analyze an evolutionary pathway, one needs to start with orthologous sequences and perform the analysis properly. However, single gene phylogenies generally have less evolutionary signal. As genomes are now being widely sequenced, the possibility of tree reconstruction based on entire or nearly complete genomes is emerging. This approach may replace traditional techniques in molecular evolution in the near future.