Abstract

Data Mining is a branch of knowledge discovery in the field of research and development. The biological data is available in different formats and is comparatively more complex. Knowledge discovery from these large and complex databases is the key problem of this era. Data mining and machine learning techniques are needed which can scale to the size of the problems and can be customized to the application of biology. Hierarchical Clustering is the one of the main techniques for data mining. Phylogeny is the evolutionary history for a set of evolutionary related species. One approach on determining the evolutionary histories of a dataset are scoring based methods. There are number of different distance based methods of which two are details with here: the UPGMA (Unweighted Pair Group Method using Arithmetic average) and Neighbor Joining. A method for construction of distance based phylogenetic tree using hierarchical clustering is proposed and implemented on different rice varieties. The sequences are downloaded from NCBI databank. Evolutionary distances are calculated using jukes cantor distance method. Multiple sequence alignment is applied on different datasets. Trees are constructed for different datasets from available data using both the distance based methods and pruning technique. SNAP calculates synonymous and non-synonymous substitution rates
Phylogenetic Tree Generation using Different Scoring Methods

based on a set of codon aligned nucleotide sequences. The DNA Multiple sequences to calculate the GC content of eukaryotes, molecular weight, melting temperature and tree information. Extractions of closely related varieties are performed by applying threshold condition. Then, final tree is constructed using these closely related rice varieties.

References

- Usama Fayyad et al., (1996) “From Data Mining to Knowledge Discovery in Databases”; American Association for Artificial Intelligence, Volume 17 Number 3, pp. 37-54.

Index Terms
Keywords
Data Mining   DNA   Phylogenetics