Neighbor Joining
Bioinformatics

William Willie Wells

Abstract

Neighbor Joining is an agglomerative clustering algorithm. Common uses of this algorithm is for building phylogenetic trees. Phylogenetic trees are used to display possible relations between biological elements. A data set consisting of 489 amino acid sequences was used to create a phylogenetic tree. The resulting phylogenetic tree was output in a format similar to Newick format.

Introduction

Agglomerative clustering methods begin with all n elements in separate clusters. A distance measure between elements is defined. Iteratively, the nearest elements are paired or joined into a single cluster abolishing the previous clusters while creating a new cluster. This decreases the total clusters to n-1 elements. This procedure is followed until there is only a desired amount of clusters remaining. When building a phylogenetic tree, the desired amount of clusters is one. (1)

In the case of neighbor joining an initial distance measure is defined but all successive distance updates are computed by subtracting the average distances to all other leaves from the distance between two elements. (1)

Technical Approach

The database is read into a matrix. A modified Jukes-Cantor formula is used to define the distance between elements:

distance = -(20-1)/ 20 * log(1 – elements that differ between pair / maximum sequence length) where 20 is the number of different base elements (amino acids).

The two elements with the minimum distance are selected to be joined. These elements are added to the tree with their branch lengths and a pointer to their parent node. The chosen elements are removed from the set of available leafs represented by a distance matrix and the parent node is added to the distance matrix. The number of available leaves is decremented and the process is repeated. The final pair of leaves is added to the tree as well as the final branch length.

Experiments

Due to the size of the database, test cases of uniform but randomly varied length sequences and randomly varied amount of sequences were generated to verify the correct implementation of the algorithm.

Conclusion

The subtraction of the averaged distances to all other nodes is a clever trick providing better results than other similar algorithms.

Reference

1. R. Durbin., S. Eddy., A Krogh., G. Mitchison., *Bioloigical Sequence Analysis Probablistic models of proteins and nucleic acids.*