Protein Sequence Classification Using Neighbor-Joining Method

Bo Liu

Overview

Given: A group of protein sequences, which have same function and have somewhat similarity between each other.

Input: A protein sequence with unknown function.

Output: If this query sequence belongs to the given cluster based on some rules.

Representation of Sequences Cluster

Neighbor-joining is a bottom-up clustering method used for the creation of phylogenetic trees, usually based on DNA or protein sequences. However, it requires knowledge of pairwise distance between each pair of sequences. So we can use a distance matrix to represent a cluster of sequences. There are various ways of calculating pairwise sequences distance, such as Smith-Waterman algorithm[1], BLAST[2], multiple sequence alignment and relative Lempel-Ziv Complexity[3], which is alignment free. For easy implementation and fast speed, I will use multiple sequence alignment (using CLUSTALW[4] package) to calculate the distance matrix (using Phylip[5] package).

Neighbor-Joining Tree Construction

Neighbor-Joining tree is calculated using the standard methods[6, 7].
At each step, a Q-matrix is calculated based on the distance matrix in order to find the pair of sequences with lowest Q value. The Q value for sequence i, j is calculated using equation:

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

A new internal node (u) joining these two sequences is created on the tree. Then calculate the branch length of each of the two sequences (f, g) to this newly created internal node (u) using equation:

$$d(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(r-2)}\left[\sum_{k=1}^{r} d(f,k) - \sum_{k=1}^{r} d(g,k)\right]$$

Before joining next two nearest neighbors, we have to update the distance matrix by joining the two neighbors of current step. And the distance of other sequences to this pair of neighbors is calculated as:

$$d(u,k) = \frac{1}{2}[d(f,k) - d(f,u)] + \frac{1}{2}[d(g,k) - d(g,u)]$$

Repeat this process until all sequences are joined together into a neighbor-joining tree.

Criteria for Classification

1, Sequence with longest branch length
In a neighbor-joining tree, if one leaf has the longest branch length overall, then it means this sequence evolves too fast compared with all other sequences. So if this sequence is the query sequence, then it does not belong to this group of sequences.

2, Sequence finally joined into the tree
Because at each step, neighbor-joining method is trying to merge two nearest neighbors in order to minimize the branch lengths of the whole tree. The sequence, which is the last one joined into the tree, increases the whole branch lengths of the tree most. So if this sequence is the query sequence, I consider it not belonging to this group of sequences.

Summary

The assumption behind this method is that if a group of sequences have same function and mechanism, then they are evolved from a same origin. So there is unknown evolutionary history behind these sequences. Neighbor-joining method is a good way to recover this evolutionary process based on maximum parsimony. If the query sequence evolves too fast or costs most (last joined) in a tree, there are two possibilities. This sequence has the same origin with other sequences but does not maintain the same function anymore. Or this sequence does not have the same origin with other sequences; hence they do not share same function. Both cases indicate that this query sequence does not belong to this group of sequences, and they do not share same function. So this classification method can be used for protein function annotation.

Reference

1. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7 (1981).
2. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
3. Otu, H.H. & Sayood, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* **19**, 2122-30 (2003).
4. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
5. Retief, J.D. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* **132**, 243-58 (2000).
6. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25 (1987).
7. Studier, J.A. & Keppler, K.J. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* **5**, 729-31 (1988).