Genomics and bioinformatics analysis – Phylogenetc Analysis

Aim

Based on our sequenced proteins and combined with relevant literature, we use nucleic acid (NCBI, EMBL, DDBJ) and protein (SWISSPROT) databases to identify homologous genes or proteins of the corresponding proteins and analyze their evolutionary history. In the process of database mining, we analyze the potential information in other people's data, find out the characteristic sequences of homologous proteins, and assist the experimental design.

Phylogenetic analysis is generally based on molecular clocks. Molecular clock theory is that the evolution of molecular sequence is carried out at a constant rate, so the number of accumulated mutations is proportional to the evolution time. It can be concluded that: 1) with the evolution of species, the closer the species with similar evolutionary level are, the closer their sequence is; 2) if evolved from the same species, the differentiated species will retain the mark of common ancestor, which is different from other ancestors. Based on this hypothesis, phylogenetic trees can be constructed according to the sequence or structure differences of proteins. Phylogenetic trees are usually represented by branching hierarchies or topological graphs, which can reflect

the divergent time or evolutionary distance of each species or protein molecule through the length of the tree branches.

Procedure

1. Obtain homologous sequence data. In the process of sequence similarity alignment, we use the BLAST tool of web page version.

(1) Generally speaking, we prefer the general blastp program as the retrieval algorithm. Since our query sequence and the target sequence of the database are all proteins, the protein blast function is preferred.

(2) According to our pre-designed tree-building idea, we chose the non-redundant protein sequence database as the search set. Species were limited to the landmark species in the evolutionary history, such as human, mouse, chicken, zebrafish, shark, sea urchin and etc.

(3) In the case of poor blastp comparison results, we use the PSI-BLAST program depending on PSSM matrix for iterative comparison. Normally, the statistical significance threshold is 0.05. The comparison results after each iteration conform to the significance threshold sequence, not more than 500 items, participating in the matrix construction and the next iteration. The number of iterations is set as 5 rounds.

(4) Select the genes with high similarity and download the gene sequence in Fasta format.

2. Multiple sequence alignment. ClustalW in Mega is selected for multi-sequence alignment, and the alignment results are saved in MEG format.

(1) In the Mega program, select ALIGN-Edit/Build Alignment to create a new alignment, select protein as sequence type, and open an Alignment Explorer.

(2) Import the downloaded Fasta format file and target protein sequence file through Edit-Insert Sequence From File.

(3) Select all the sequences, click Alignment-Align By ClustW or W, select Multiple Alignment to set the penalty parameters for comparison under the subdirectory, and then point OK to confirm.

(4) Click Data-Phylogenetic Analysis to proceed Phylogenetic analysis. Close the window and save it.

(5) Select Analysis-Phylogeny and confirm the analysis method to construct the evolutionary tree. Usually, NJ tree is used, verification method is set to Bootstrap, Replications is set to 1000, and other default parameters are used to calculate and generate phylogenetic tree.