Секция «Биоинформатика»

Modelling of random protein sequences' alignment from a given phylogenetic tree

Научный руководитель – Spirin Sergey

Belyaeva Julia Dmitrievna
Student (specialist)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: belyaevajd@yandex.ru

In computational biology, modeling of evolution of biological sequence is a challenging task because the evolutionary history of biological objects is usually not known. Models of amino acid substitution described as substitution-rate matrices are usually used for random generating of protein sequences. Traditional models poorly reflect the evolutionary features of real proteins; in our work, modelling is based on alignments of natural proteins by randomly selecting their alignments' elements.

Methods

During our work we used Python scripting to make the program and also bash for quick and short processing of files. The input of the developed program is a number of files with phylogenetic trees in Newick format, a directory with a database of natural sequence alignments and the other one with corresponding distance matrices. For each input tree, the program randomly selects one alignment from the database and then the root sequence is generated. The length of aligned sequences we would like to get is an input parameter. For each position of the root sequence the program randomly selects a sequence from an alignment of the database and a letter in this sequence. Then using our algorithm by traversing the tree in wide from an ancestral sequence the program generates its descendants. The program has friendly user-interface.

The database that was used to run the program was made from the latest Pfam [1] release. We selected necessary families in accordance with following criteria: the length of family's proteins must be more than 300 amino acid residues and every selected family must contain more than 100 and less than 1000 representatives. If the number of representatives less than 100, the algorithm does not work well (the effect of convergence of terminal sequences is observed) and the higher border is due to difficulties in making alignments of great number of sequences.

Results

We have modeled several random alignments using our program and then calculated distance matrices of their sequences. The correlation coefficient between the distances across the input trees and from the obtained matrices is about 0.95 for sequences with length >300. The Spearman rank correlation between the same arrays is over 0.7.

Further plans

We plan to compare the results of our tool with the results of existing analogues, such as ALF [2] and PhyloSim [3] on data presented in Dryad data repository [4], to check if the features of alignments generated by our tool are closer to features of sequence alignment of natural proteins.

References

1) ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.fasta.gz

2) Dalquen D. A. et al. ALF—a simulation framework for genome evolution //Molecular biology and evolution. – 2012. – Т. 29. – №. 4. – С. 1115-1123.

3) Sipos B. et al. PhyloSim-Monte Carlo simulation of sequence evolution in the R statistical computing environment //BMC bioinformatics. – 2011. – Т. 12. – №. 1. – С. 104.

4) https://datadryad.org/stash