Poster C-29

PhyloTree – An Integrated and Automatic Tool to Generate Phylogenetic Trees

Authors:
Ana Luiza Bessa de Paula Barros (*Computação-UECE*)
Gerardo Valdísio Rodrigues Viana (*Computação-UECE*)
Tariana Mara Fernandes Batista (*Computação-UECE*)
Gustavo Agusto de Lima Campos (*Computação-UECE*)
Raimundo Bezerra da Costa (*NUGEN-UECE*)
Rodrigo Maggioni (*NUGEN-UECE*)
Diana Magalhães de Oliveira (*NUGEN-UECE*)

Short Abstract: PhyloTree is an integrated tool to help generate phylogenetic trees in an automated way. It combines successive steps, based on well-developed algorithms including ClustalW and PHYLIP, while it is intended to assist biologists from a broad range of disciplines and is likely to be particularly helpful for non-experts.

Long Abstract:
BACKGROUND. Phylogenetic analysis in biological sequences is an important and common strategy for research into evolution and taxonomy. Therefore, biologists frequently utilize phylogenetic analysis to present or interpret sequence data. The significant quantity and constant multiplication of sequences in public databases can be a huge biological resource if researchers can efficiently extract information from a wide range of species to analyze and develop a sophisticated theory to determine phylogeny. To build phylogenetic trees, however, is sometimes a tedious task that demands several steps, reasonable ability and precious time. The main challenge is the need for diverse tools with different features. There is a number of steps from the start (when one chooses a set of organisms to analyze and a set of related proteins present in these organisms), to the end point (when the tree is really constructed). Key tasks involved in these steps are public databases search, multiple sequence alignments and distance matrix calculation. So, it is necessary one specific tool to execute each one of these steps. Tasks must be executed in a predefined order and all data generated by one tool is used as input to the next one. Unfortunately, data formats might not be compatible and the user must treat the incompatibility manually, sometimes converting one format into another one. In some cases, the user also must rename the file generated by the last tool to make it compatible to the next one. Thus, parallel calculation and an optimized database structure for parameter collection, job processing and result visualization, plus sequence preprocessor, file format converter, topology illustrator and job controller are all needed. By that we can conclude than that the whole process is complicated and handful and that it is a very good idea to have a tool that makes all tasks and data format treatment automatically. Computational tool integration is a very difficult challenge. Problems may arise at different abstraction levels and from several sources such as heterogeneity of manipulated data and incompatible interfaces. Implementing a development environment composed of many integrated tools requires the management of several aspects, such as synchronization of operations, data transformations, performance tuning, and management of permissions. To develop a tool, PhyloTree, capable of

generating phylogenetic trees automatically using information obtained from public databases, we applied the following METHODS. PhyloTree was developed using Java programming language inside Eclipse platform. Some advantages are Java portability and the existence of many libraries with different features. Two of them were very important to this development: java.net.HttpURLConnection and java.net.URL These libraries made possible the connection with public databases. To execute specific tasks, PhyloTree requires classic programs, such as CLUSTAL W, TREEV32 and some programs from PHYLIP (PHYLogeny Inference Package) packet. The public database used is NCBI. At the beginning, the tool executes a database search at NCBI to obtain protein sequences for the organisms from which the phylogenetic tree will be generated. After, programs are executed at the following order: 1) CLUSTALW, to realize multiple protein sequences alignment; 2) SEQBOOT, to generate many instances of obtained data from CLUSTAL W; 3) PROTDIST, to generate a distance matrix from sequences alignment executed by CLUSTAL W and replicated by SEQBOOT; 4) NEIGHBOR, to generate phylogenetic tree using distance matrix calculated by PROTDIST and 5) TREEV32, to graphically visualize the tree generated by NEIGHBOR. Graphically, the tool has two bars. The first one has the following elements: text field where the user can enter the word that will be searched; a field with options of registers quantity that the tool will show as result; a “Search” button to realize the search in databases and a “FASTA Format” button that generates the file in FASTA format. The second bar is located at the bottom. It contains the buttons to execute each program, disposed in the correct execution order. There is also a button “Clean”, to clean the screen in case of user wish to start a new search. The central part of the screen is used to exhibit the results of the search and of programs execution. The user must enter the name of a gene or part of one and the number of registers he/she wants in return. After the search is executed using “Search” button, the tool exhibits the results in the screen. They are composed of a gene description and a list of proteins. User must choose which proteins will be used to generate the phylogenetic tree. RESULTS. PhyloTree is an integrated tool to help generate phylogenetic trees for different sets of protein of different organisms in an automated way. Users can easily manipulate the parameters using online help and query logs for iterative jobs. PhyloTree involved constructing a simple and friendly user interface with default parameters for quick analysis and the option to define specific parameter settings. The platform covers almost all combinations of ClustalW and PHYLIPS parameter settings and provides valuable help for each step to satisfy most needs to analyze molecular evolution. Using the graphical tool, users can easily interpret a phylogenetic tree without improper crossing. Tree building iteration is a common process among biologists trying to find a proper tree topology. Tracking changes in a tree resulting from repeated parameter tuning and addition or deletion of sequences is difficult. PhyloTree overcomes this problem by providing detailed logs of user-defined parameters, all output files and process history. The tool returns some output files, such as fileFASTA.txt (contains all proteins selected by user in FASTA format), resultCLUSTALW.phy (contains multiple protein alignment executed by CLUSTALW), resultSEQBOOT.dat (multiple series of multiple alignment present in resultCLUSTALW.phy), resultPROTDIST.dat (contains distance matrix generated by PROTDIST) and Phylogenetictree.tre (file with the phylogenetic tree generated by NEIGHBOR). PhyloTree seamlessly and flexibly combines successive steps of phylogenetic analysis, all based on well-developed algorithms including ClustalW and PHYLIP, while it is intended to assist biologists from a broad range of disciplines and is likely to be particularly helpful for non-experts.