Phylojava : a Generic Client-Server Tool for Phylogenetic Tree Reconstruction - Application to Grid Computing

T. Silvestre, E. Nugues, G. Perrière, M. Gouy & L. Duret

Laboratoire de Biométrie et Biologie Evolutive (BBE), CNRS-UMR 5857
Université Claude Bernard, Lyon I, 43, Bd du 11 novembre 1918,
69622 VILLEURBANNE cedex – France
{silvestr,duret}@biomserv.univ-lyon1.fr

Abstract : Phylojava is a graphic user interface devoted to edit and calculate phylogenetic trees. It is programmed in Java to allow portage on a wide-range systems and is based on a client-server architecture. Phylogenetic trees are computed on the remote server, and are sent via internet to the client that allows the user to handle alignments and phylogenetic trees. New phylogenetic methods can be easily added to the server and will be automatically displayed by the client. Besides phylojava server is a node of the european DATAGRID project that has allowed us to parallelize bootstrap process with fastDNAml method. A five fold acceleration factor has been raised using the grid compared to a standalone computer and demonstrates the future potential of the grid for bioinformatics.

Keywords : phylogenetic trees, client-server, grid computing, DATAGRID, parallelization, fastDNAml, bootstrap

1. Introduction

Many different methods exist for phylogenetic tree reconstruction. All phylogenetic programs implement a fixed set of methods which are difficult to change or modify unless by recoding the application. However new algorithms are frequently published. Therefore we have developed a client-server application that allows integration of new methods without any coding effort for the user. The client is a Graphic User Interface (GUI) devoted to edit and calculate phylogenetic trees whereas the server contains all the phylogenetic methods. The client will automatically display new methods as soon as it is installed on server. Thus the user dispose of a complete and up-to-date set of methods in one single application.

Besides, phylojava offers the possibility to parallelize jobs using a grid environment. Compared to High Performance Computing (HPC) systems like PC clusters or mainframe systems, computational grids have the advantage to directly provide available ressources. While many users are competing for the same ressource and are waiting in a batch system, a *ressource broker* allocates the jobs in the grid where it best fits in terms of user requirements (operating system, memory, storage, etc.). Our phylojava server is a node of the European grid project DATAGRID[1] that allows us to accelerate highly time consumming phylogenetic methods. The parallelization of phylogenetic trees calculations on the grid can be useful to perform the *bootstrapping* procedure with an accurate but slow method such as fastDNAML[2]. The parallelization consists of distributing the calculation of each individual tree on the nodes of the grid.

2. Phylojava Graphic User Interface

Phylojava GUI is programmed in Java to allow portage on a wide-range systems (UNIX, LINUX, DOS & MacOSX). The user has just to install the GUI on his computer to be able to submit tree jobs on the server. Phylojava works with DNA or protein alignment files in MASE format. It is possible to import alignments from various other formats (PHYLIP, FASTA, NEXUS, MSF, CLUSTAL). Taxonomic species, groups and sets of conserved regions can be defined by clicking in the alignment and stored into the sequence file, thus avoiding multiple data files. It is possible to modify trees (select the root, swap nodes, add comments, etc.) and to save them within the sequence file or as separate file.

After selection of a method and of a set of sites and species, jobs are sent to the server. Short jobs are automatically processed and the result is immediatly displayed by the client. Longer jobs are directed into a batch queue. The user is informed by e-mail when his job is finished. It is also possible to check at any time the status of a job and to kill it if necessary. Once a job is finished, results can be retrieved through the phylojava client graphical interface. Moreover, the client will be able connect to different servers that will enhance the overall choice of methods.

3. Phylojava server

Phylojava server is based on Common Gateway Interface (CGI) scripts that launch the appropriate phylogenetic methods. Communications with the client are made by a Web server through the HTTP protocole. One could install his own phylojava server and set his methods of choice. Presently, there is only one phylojava server : *pbil.univ-lyon1.fr*. In the future, we hope other mirror servers should be made available. The client will then provide an access to all the different servers and their methods. The Phylojava server will be distributed for UNIX/LINUX only with a default set of phylogenetic methods.

The list of methods available on the server and their associated parameters are automatically sent to the client. It is therefore possible for the administrator to update the server (add new methods or remove obsolete ones), without having to modify the client. Server administrators just have to write an HTML form and a shell script to interface the form with the installed command-line programs. Indeed HTML constitutes a particularly simple manner to display the different parameters of a given method.

3. Grid computing

DATAGRID is an European research project that endeavour to set up an international grid between widely distributed scientific communities. A production grid has been set up between 5 countries (Netherland, Italy, United Kingdom, Germany and France) regrouping up to 11 sites. Each site is composed of PC farms that may vary from tens of processors (FZK[3], Heidelberg) to several hundreds (IN2P3[4], Lyon).

In the parallelization of 200 jobs on the grid, a five fold acceleration factor has been raised compared to a standalone computer (Fig.1).

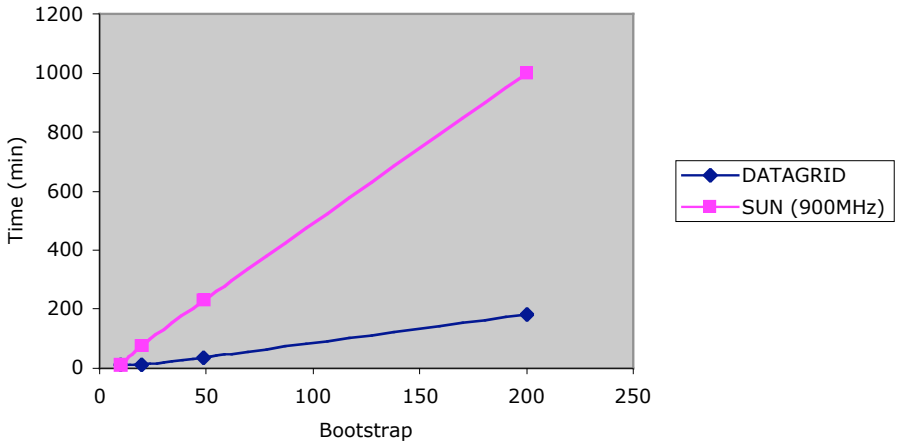

Fig.1 : CPU time estimation of bootstrap calculation in DATAGRID environment compared to a standalone computer (fastDNAml algorithm)

These results seem not to be so drastic in regards to what we could obtain in theory with a highly parallel mainframe machine. However, one have to consider that these datas are reflecting real use conditions, with a lot

of user submitting jobs at the same time. Moreover ressources availability is far much better compared to a batch system. In an average out of 100 job submissions, 10 jobs were directly running accross the different sites.

We have developed a generic algorithm to submit jobs to the grid with several parameters : the phylogenetic program, the alignemnt file, the number of jobs to parallelize, the delay between jobs submissions and job time-out. When time-out is reached, the job will be cancelled and relaunched. The user has the possibility to to stop all running jobs and to recuperate intermediate results. An interesting factor we can modify is the granularity of job submissions. Indeed we can test for instance whether it is more efficient to send 10 packets of 10 jobs compared to 100 jobs individually.

4. Conclusions

An alpha version of phylojava is now used by a small community of testers. A public beta version should be available for download at the end of october and the complete client-server package is scheduled for december, 2003.

The grid techology has proven to be successfully applicable to phylogeny and that DATAGRID middleware provides a working grid infrastructure. Although it is too early to envisage grids as a directly productive tool in bioinformatic due to instability problems, we believe that this technology will play a major role in the next decade. Next steps will focus on how to efficiently implement parallel algorithms on the grid. Parallel versions of fastDNAML based on MPI[5] or PVM[6] are available[7][8], however its implementation may be hampered on the grid by network latency and communications.

Dedicated grids will be perhaps the most appropriate solution in regards to the different fields specific needs. Grids of individual PC may be more suitable for an intranet use in a company or for small jobs over the internet whereas grids of computing center are probably more efficient for highly intensive computational jobs and fullfill stability and security requirements.

Acknowledgements

This work is a part of the DATAGRID project, funded by European Union. We thank all the people among the biomedical work package group (WP10) for their helpful discussions and the Laboratoire de Physique Corpusculaire de Clermont-Ferrand for his technical support.

References

[1] EDG : European Datagrid Project (http://www.eu-datagrid.org).
[2] Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10: 41-48.
[3] FZK : ForschungZentrum Karlsruhe
[4] IN2P3 : Insitut National de Physique Nucléaire et de Physique des Particules
[5] MPI : Message Passing Interface
[6] PVM : Parallel Virtual Machine
[7] C. Ceron , J. Dopazo , E. L. Zapata , J. M. Carazo , O. Trelles, Parallel implementation of DNAm1 program on message-passing architectures, Parallel Computing, v.24 n.5-6, p.701-716, June 1998
[8] Stewart, C.A., D. Hart, D. K. Berry, G. J. Olsen, E. Wernert, W. Fischer. 2001. "Parallel implementation and performance of fastDNAml - a program for maximum likelihood phylogenetic inference." Proceedings of SC2001, Denver, CO, November 2001.