Demonstrations / success stories

On the Performance of Large Scale Bayesian Phylogenetic Analyses with Grid Portals and Robot Certificates

Roberto Barbera[1,2,*], Giacinto Donvito[3], Alberto Falzone[4], Giuseppe La Rocca[1,*], Giorgio Pietro Maggi[3], Saverio Vicario[6], Luciano Milanesi[5,*]

[1]Italian National Institute of Nuclear Physics, Division of Catania, Via S. Sofia 64, I-95123 Catania, Italy
[2]Department of Physics and Astronomy of the University of Catania, Viale A. Doria 6, I-95125 Catania, Italy
[3] Italian National Institute of Nuclear Physics, Division of Bari, Via E. Orabona 4, I-70126 Bari, Italy
[4]NICE Srl, Via Milliavacca 9, I-14100 Asti, Italy
[5] Institute for Biomedical Technologies – CNR, Via Fratelli Cervi 93, I-20090 Segrate (MI), Italy
[6]CNR – ITB Bari, Via Amendola 122D, I-70126 Bari, Italy

Associate Editor: Sandra Gesing and Jano van Hermet

ABSTRACT
Motivation:

ICT based Infrastructures today play a crucial role in the development of collaborations among scientists to address new global scientific challenges, particularly those that have high societal and economic impact. In the Life Science domain, for instance, the massive computational resources exposed by Grid infrastructures is indispensable when dealing both with the complexity of models and the enormous quantity of data to be processed, for example, to perform genome scale analysis or when carrying out docking simulations for the study of new drugs. At present, Grid technology is presented to end-users as a collection of virtual services and complex protocols and this makes its full exploitation very complicated. A notable step forward to foster the adoption of this technology in e-Science has recently been achieved with the adoption of portals and robot certificates. Robot certificates have been conceived and introduced to allow non expert users to access Grid Infrastructures and reduce the initial barriers. Each robot certificate is associated with a function which identifies the specific application the user wants to share with all the members of the same community. In this manuscript the solution proposed by the Italian National Institute of Nuclear Physics to allow bioinformaticians to access the Grid via a portal enabled by a robot certificate and perform large scale Bayesian Phylogenetic analyses is presented. The solution described in this manuscript strongly simplifies the exploitation and the utilization of Grid Infrastructures and represents a valuable step forward towards the adoption of this computing paradigm in Life Sciences.

1 INTRODUCTION

The study of molecular evolution is based on the principle hypothesis that the level of similarity between genes displays the degree of evolutionary relationship between them. The reconstruction of the evolutionary history of a group of organisms (a phylogeny) is used throughout Life Sciences, as it offers a mean to organize the knowledge and data accumulated by researchers. It is commonly known that the unraveling of phylogenetic relationships between organisms' gene sequences is an important first step towards the understanding of their evolution. Molecular evolution is generally studied only by inference: a pattern (in this case a set of biological sequences) is observed and different possible processes are evaluated to infer what process produced this pattern. The MrBayes program (http://mrbayes.csit.fsu.edu/) *(Ronquist et al. 2003)* uses a Markovian integration to obtain samples from the posterior distribution of the parameters to calculate the inference. The program generates a Bayesian phylogenetic inference among different aligned bio-sequences. The inference allows to identify the distribution of the most likely genetic relationship among the set of chosen bio-sequences and, at the same time, the best set of values for the parameters of the postulated model of evolution of the bio-sequences. MrBayes has a great richness of models of evolution for DNA (both as nucleotide and codon), RNA (model for evolution of doublet of nucleotide to model the secondary structure of an RNA molecule), protein, and even arbitrary hereditary discrete characters. Another peculiarity of the application is that it allows the usage of "mixed" models such as using different models for different parts of each bio-sequence with the possibility to share parameters among the different models. The program uses a Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC) to perform the Markovian integration necessary to solve numerically the Bayesian equation *(Altekar et al. 2004)*. Due to the nature of Bayesian inference, in order to achieve the better estimation, the MrBayes program has to run for millions of iterations (generations) which require a large amount of computation time. The input required is a single text file, nexus formatted *(Maddison et al. 1997)*, subdivided in a data block and a MrBayes block in which the models and the parameters of Markovian integration are defined and declared. The output consists of three kinds of large files (typically in the order of several hundreds of Megabytes each) that describe, respectively, the posterior distribution of numerical and topological parameters and several diagnostic measures related to the mixing of Markov chains and the converging of the algorithm

*To whom correspondence should be addressed.

as a whole. The use of a distributed version of MrBayes is more problematic given the nature of the Markovian integration. Typically, each MrBayes run, although starting from a fairly small input data, has a quite long execution time. This is a typical analysis which can be tackled by a "high-throughput" approach made feasible with the use of Grid infrastructures. This is why in this work we decided to restrict the executions, each made of a single complete analysis, to high performance nodes of the EGEE grid (http://www.eu-egee.org/) configured to accept MPI jobs. This ensured that jobs would run in sufficiently efficient manner and they would arrive at completion before the maximum run time allowed by the chosen nodes of the EGEE grid. In addition, to allow all the bioinformaticians of the LIBI project (http://www.libi.it/) to access the computing resources of the EGEE Grid Infrastructure without owning a personal X.509 certificate and run the parallel version of MrBayes, the credentials of a robot certificate have been made available with the GENIUS Grid portal according to the architectural schema reported in the next section. Thanks to the introduction of Grid portals and robot certificates now it is possible to reduce the initial barriers and extend the benefits of the Grid paradigm to a wider community of users. In section 2 the details of the distributed grid environment designed and deployed in the context of the LIBI project to perform phylogenetic analyses on a large scale is presented. Results are summarized in section 3.

2 METHODS

The EnginFrame (http://www.enginframe.com/) Java framework (ver. 4.1), on which the GENIUS Grid portal *(Andronico et al. 2003, Barbera et al. 2007)* is built, has been enhanced in order to provide a transparent support to robot certificates and allow non-expert Grid users to access the distributed computational resources of a Grid using a conventional web browser. The additional features introduced in GENIUS Grid portal are sketched in fig. 1

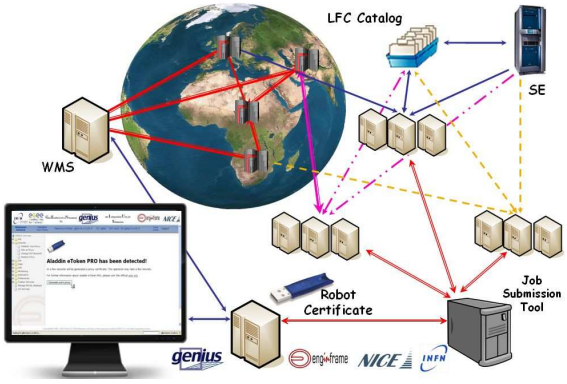

Fig. 1 – The workflow set up to perform large scale phylogenetic analyses.

The distributed environment for the application has been built on top of the EGEE stack (http://public.eu-egee.org/) using many gLite (http://www.glite.org/) services such as: the LFC File Catalog and Data Management System, Storage Elements, Workload Management System and X.509 certificates. The client side is represented by a user's workstation running a web browser (bottom left in the figure). The server side is represented by a gLite User Interface (UI) machine, equipped with the latest stable release of middleware services to submit jobs and manage data on Grid, the Apache Web Server, the Java/XML portal framework EnginFrame (http://www.enginframe.com/docum/) developed by NICE Srl (http://www.nice-italy.com/) and the GENIUS Grid Portal itself (bottom center in the figure). After user's login, a proxy certificate is requested by the portal to access the distributed resources of a Grid Infrastructure according to the Grid Security Infrastructure (GSI) standard. If no proxy is available, the credentials of the robot certificate, if any, will be read by the GENIUS Grid portal to generate in a few seconds the needed proxy. This operation is completely transparent to end-users. Once the proxy certificate has successfully been created, users are automatically redirected to the home page of the application related to the robot certificate. In this context the robot certificate has been requested to run the phylogenetic application, thus after login he/she is redirected to the MrBayes' home page. For this purpose some dedicated services have been designed and implemented into the GENIUS portal to allow users to specify input settings before sending parallel instances of MrBayes jobs to the Grid. Besides, in order to enhance the reliability and the performance of the architecture, the support for the submission/re-submission of a large number of jobs in an almost unattended way has also been introduced (bottom right in the figure). This tool is based on the concept of "task" to be executed *(De Sario et al. 2009)*. The entire problem is first subdivided into elementary tasks, then all the tasks are inserted into a DB server. In the submission phase all the jobs are completely identical. Only when the jobs lands and starts executing on a worker node, it requests to the central DB a task to execute. Information on the execution of each task is logged into the central DB. Only if all steps are correctly executed by the job, the status of that particular task on the central DB is updated to "Done". In this way the central DB provides a monitoring of the task execution and no manual intervention is required to manage the re-submission of the failed tasks. Tasks which are found in a "running" state after a given time interval are considered failed and automatically reassigned to new jobs. Figure 2 shows the service introduced into the portal to query the central database of the Job Submission Tool and monitor users' tasks. This service, based on HTML, XML, JavaScript and PHP refreshes data every 5 minutes.

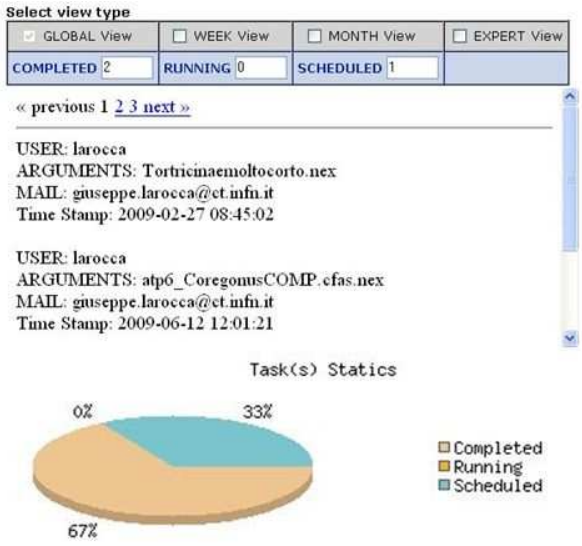

Fig. 2 – View of the task monitoring system.

The completed tasks obtained at the end of the analysis are stored in a Grid Storage Element (SE) and then downloaded locally in the user's home directory by means of a pop-up service. The results of the computation can then be used to display the phylogenetic tree with third party software like, for example TreeViewX (http://darwin.zoology.gla.ac.uk/~rpage/treeviewx/).

Fig. 3 – Display of the phylogenetic tree with TreeViewX.

3 RESULTS

The adoption of a personal certificate to access Grid resources has represented so far a limiting factor for the real spreading of the Grid paradigm in many types of researches. Many scientists are interested in using Grid as a tool to solve their computing problems and speed up the production of scientific results but the basis of the Grid Security Infrastructure risks to discourage many of them. The benefits introduced by the GENIUS portal and robot certificates in Life Sciences are far reaching because they can contribute both to effectively reduce the scientific gap requested to access Grid infrastructures and make the adoption of this technology transparent not only to biologists but also to many other scientific communities. The solution presented in this paper has been successfully evaluated and adopted by different EU co-founded projects. In the context of the e-NMR project (http://ww.enmr.eu/) the HADDOCK web portal (http://www.haddocking.eu/) makes use of robot certificates issued by the Dutch CA according with the VO Portal Policy draft documented by the Joint Security Policy Group (JSPG) of EGEE. The portal gives access to information-driven docking at various levels of expertise, from an easy to a guru interface providing full control on the docking parameters. The GridSPM is another web portal *(Corradi et al.)*, that allows the statistical analysis of SPECT and PET cerebral images through the Statistical Parameter Mapping (SPM) system (http://www.neuroinf.it/medico/Analisi/). Finally, in the context of the EU co-founded GRIDCC (http://www.gridcc.org/) and DORII (http://www.dorii.eu/) projects, ELETTRA (http://www.elettra.trieste.it/) has developed the Virtual Room (VCR) (http://www.dorii.eu/middleware) , a grid portal which allows users to interactively control remote Instruments Elements and supports both user and robot certificates.

ACKNOWLEDGEMENTS

We gratefully acknowledge all the people who supported this work contributing with ideas, comments and feedback and the e-Science Institute in Edinburgh.

Funding: This work was supported by the MUR FIRB LIBI "Italian Laboratory for Bioinformatics", LITBIO (http://www.litbio.org/, RBLA0332RH), and ITALBIONET (RBPR05ZK2Z_001) Italian projects and by the EGEE-III (contract number: 222667) and BIOINFOGRID (http://www.bioinfogrid.eu/, contract number: 026808) European projects.

REFERENCES

Altekar G, Dwarkadas S , Huelsenbeck JP and Ronquist F.(2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20(3): pp. 407–415

Andronico G, Barbera R, Falzone A, Lo Re G, Pulvirenti A, Rodolico A – GENIUS: a web portal for the grid. Nucl. Instrument and Methods in Phy. Res. A 2003.

Barbera, R *et al.* (2007) The GENIUS Grid Portal: Its Architecture, Improvements of Features and New Implementations about Authentication and Authorization. 16th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2007), pp.279-283

Corradi, L *et al.* (2009) XTENS - an eXTensible Environment for NeuroScience, HealthGrid 2009, Berlin

De Sario G *et al.* (2009) – The Job Submission Tool, "High-throughput GRID computing for Life Sciences, in Mario Cannataro (Ed.), Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, IGI Global", Vol. 1, pp. 198-203

Maddison, DR *et al.* (1997) NEXUS: an extensible file format for systematic information. Syst. Biol., pp. 590-621

Ronquist F and Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12): pp. 1572–1574