BioGrIP – An Italian Bioinformatics Grid Partnership

R. Calogero[2], M. Canonico[5], M. Belluardo[1], G. Franceschinis[4], C. Anglano[4], M. Botta[3], R. Boraso[5], G. Ballocca[5], A. Rodolico[6], E. Rossi[7], A. Emerson[7] and E. Medico[1]
[1]Institute for Cancer Research and Treatment, University of Torino Medical School; Departments of [2]Clinical and Biological Sciences and [3]Informatics, University of Torino; [4]Department of Informatics, University of Piemonte Orientale, Alessandria; [5]CSP-Innovazione nelle ICT, Torino; [6]Nice srl.; [7]High Performance Systems, CINECA.

Address for correspondence:
Enzo Medico, Institute for Cancer Research and Treatment, University of Torino Medical School, S.P. 142, km 3.95, 10060 Candiolo (TO) Italy
Tel (+39) 011 9933 -234 (office) / -200 (lab)
Fax (+39) 011 9933 -225
enzo.medico@ircc.it

Summary

One of the main challenges facing bioinformatics is to overcome the dispersal and fragmentation of data and of the necessary tools for their analysis. Although there are a few large centres providing data and software repositories, many other important resources are scattered around the globe. The main aim of this work is to use the Grid concepts and methodology to link together resources to facilitate their use for those who are not experts in information technology. This is particularly important in a country like Italy, which has no large bioinformatics centres. The Italian Grid infrastructure described uses standard Grid enabling technologies such as the Globus toolkit (www.globus.org), EnginFrame for the web portal (www.enginframe.com) and OpenCA software (www.openCA.org) for public-key exchange (i.e. encryption). Job scheduling and management has been performed using standard software such as Condor (www.cs.wisc.edu/condor) and OpenPBS (www.openPBS.org). The Grid has been running for two years now on various hardware installations in the Turin area of Italy. A first pilot operation, consisting of two web services to submit BLAST[1] and MultiBLAST calculations, has already been implemented. A second pilot, currently under advanced development (Lazzarato et al. Bioinformatics 2003, accepted for publication), consists of a database allowing the extraction of non-coding regions surrounding a coding sequence from sequenced genomes. Though limited in size, the current testbed implementations already indicate that Grid philosophy will be of great benefit to bioinformatics, particularly in Italy. Recent joining of CINECA and other major Italian bioinformatics centers, is further increasing BioGrIP size and potential.

Keywords: Grid, bioinformatics.

Introduction

Over the last few years, research on Grid Computing has been carried out in Italy mainly through pilot studies in the IST field, aimed at building internal grid systems as test-beds for demonstration and optimization purposes. However, a research partnership aimed at supporting Bioinformatics applications has been running for almost two years now in Italy. This work resulted in the establishment of a working grid system connecting computers and clusters of multiple independent centers, and in the development of Bioinformatics tools designed to work in – and benefit from – a Grid environment.

More recently, the partnership was extended to other researchers and organizations across Italy, and now encompasses:

- University of Torino, Departments of (i) Oncological Sciences, (ii) Informatics, (iii) Clinical and Biological Sciences, (iv) Genetics, Biology and Biochemistry
- Department of Informatics, University of Piemonte Orientale, Alessandria
- Fondo Edo Tempia per la Lotta Contro i Tumori, Biella
- CSP - Innovazione nelle ICT, Torino
- Politecnico di Torino.
- IFOM, Milano
- IST, Genova
- CNR, Bari
- CINECA, Bologna
- Department of Biomolecular Sciences and Biotechnologies, University of Milano
- Nice srl., Camerano Casasco

Objectives

The main objectives of the partnership are:

- To support the development and use of algorithms, software and analytical methods to solve well-defined biological problems;
- To encourage the interaction between biological sciences and other disciplines like computer science, mathematics, statistics, physics and chemistry;
- To facilitate the wide dissemination and sharing of tools and data collections;
- To encourage the adoption of common standards and curation to allow the widest possible sharing of data.

The key biological priorities include:

- User-friendly, versatile and transparent systems to meet researcher's needs: computing portals to access computing resources;
- Support for data analysis, prediction and modelling: implementation of statistical methods, software and algorithms for data mining and knowledge management;
- Tools for data management: quality-assured methods of acquiring and managing large volumes of data; ability to integrate different data types; analysis methods over multiple categories of data; accurate tracking of data;
- Standardization of data: to facilitate reliable integration of data from different sources and experimental results publications;
- High throughput processing and networking: using Grid technology and more efficient algorithms.

The BioGrIP is an open infrastructure to support the activities of research groups across different organizational and administrative domains. It allows the sharing of computing resources and data according to policies to be defined by the owners.

Of course, this requires a dedicated and coordinated effort: specific software components are needed to enable seamless access to resources across organizational boundaries; specific administration operations are required to enable single users to use the shared resources.

First of all we need a group of components taking care of user identification, delegation of rights for remote job execution, data access and storage, resource scheduling. Other issues to deal with are network configuration to allow and control trans-organization access, digital certificates to provide authentication for machines and users, a portal to enable easy access to the available resources also for untrained users. Last, but not least, applicative software has to be deployed and data have to be replicated and distributed over the grid.

Fundamental activities can be summarized as follows:

Setup of new nodes. When a new organization joins the grid, the middleware has to be deployed on every single node and certificates have to be installed.

Infrastructure maintenance. Middleware and software upgrading, patching, data distribution on repository nodes, applicative software installation and configuration, portal integration.

Initial training, general troubleshooting and tuning.

An important aspect of the partnership agreement is also the sharing of computing resources, which has to be defined and possibly quantified (number and types of computers, computing power, storage amount and available software). BioGrIP should not bind members to excessive resource sharing, however a minimum amount should be fixed for the sake of grid efficiency.

The Italian Grid infrastructure is expected to offer the following services:

- transparent access to the computing resources
- management of replicated copies of data files
- support for applicative services
- usage of the most popular bioinformatics applications (eg BLAST, PATSER, Rosetta Resolver) on a high throughput computing infrastructure
- creation and management of a shared coherent relational database to resolve incoherencies and inconsistencies in the actual databases and to provide the infrastructure to gather data coming from microarrays experiments
- security: there are three relevant points of interest: database security (in particular all aspects concerning data confidentiality), data transfer channel encryption and, last but not least, user authentication and authorization.

Current BioGrIP state

A web portal prototype was developed based on EnginFrame (by NICE-ITALY) in addition to a resource broker to coordinate the access to hardware and software resources.

We have implemented two pilot applications, enabling parallel BLAST analysis[1] and genomic sequence extraction, respectively. Two web services have been implemented to submit BLAST and MULTI-BLAST to the grid. In order to operate, each user must obtain a X509 certificate from our Certification Authority, then he can submit his job from the portal home page using any web browser. The user can then monitor the status of his job (pending, running, done or fail) and get the output files using the web interface. Obviously the user can see the files only if the corresponding job status is “done”. It must be emphasised that each job has been configured to interface with local job schedulers (PBS, CONDOR, LSF…) of the resource where the job will be run. A database allowing the extraction of non-coding regions surrounding a coding sequence from sequenced genomes is under advanced development (Lazzarato et al. Bioinformatics 2003, accepted for publication). A Spitfire server (http://spitfire.web.cern.ch) provides a grid-enabled middleware service for access to the relational database The db is replicated on 16 machines and a “broker” application distributes queries on the basis of the machines load.

In the following, the first testbed features are reported:

Hardware:

Interconnecting network: the academic and research network GARR.

Nodes: (i) a 32-node ix86 cluster running Condor (Computer Science Department at Università del Piemonte Orientale); (ii) a 4-node ix86 cluster running OpenPBS (CSP); (iii) a 8-node ix86 cluster - to be configured (Department of Clinical and Biological Sciences at Università di Torino); (iv) a 2-node Linux computer with OpenPBS at CINECA; (v) 8 further nodes to be configured as a geographically distributed Condor cluster (Politecnico di Torino, Department of Electronics, Università di Torino, Department of Computer Sciences and IRCC).

Software

- The middleware is Globus Toolkit v. 2.2.
- The web portal is being developed using EnginFrame.
- PKI (Public Key Infrastracture) is based on the OpenCA software release run under EuroPKI Certification Authority and is managed by CSP.
- Job scheduling and resource management on the clusters will be performed using OpenPBS and Condor.

Future work

In our scenario, the Resource Broker (the decision process that assigns application ``tasks" to resources both in time and space) must choose the “best” resource on the Grid for a given job, considering two aspects, respectively, the resources with enough computational power and the resources where the Databases are stored

In many grid projects (like DataGrid) the Resource Broker considers only the computational power and moves the input files to the target resource. This because DataGrid has been implement for intensive computational applications with small input. This is not the case for biological applications, where the input file could be very large (i.e., a Blast execution can require an input database of 2 Gbytes). In this kind of applications, if the chosen resource does not host the input database, the time needed to transfer input data could be larger than the execution time.

Future work will focus on:

- Resource Broker for **scheduling** and the **deployment** of distributed scientific applications. We are taking into account different implementations (G.R.A.I.L. by San Diego Super Computer Center and DataGrid) to find out how to improve them, also considering the locations of input files.
- Database management on the Grid for periodic upgrading of databases, replica catalogue and Directory Service to know location of each database on the Grid, and further incorporation of CINECA resources in the Grid. CINECA is Italy’s largest computer centre and as well as hosting the most powerful supercomputers in the country it also able to provide state-of-the-art storage facilities and backup, high-speed networking and a wide range of computational software and databases for bioinformatics. The addition of further CINECA’s facilities would thus greatly enhance and extend BioGriP.

References

1. Altschul SF, Gish W, Miller W, Myers, EW, Lipman, D.J. Basic local alignment search tool. 1990; J. Mol. Biol. 215:403-410
2. Breton V, Medina R, Montagnat J. DataGrid, prototype of a biomedical grid. Methods Inf Med. 2003;42(2):143-7.