Challenges and Opportunities of HealthGrids
V. Hernández et al. (Eds.)
IOS Press, 2006

GPS@ Bioinformatics Portal: from Network to EGEE Grid

Christophe Blanchet [a,1] , Vincent Lefort [a], Christophe Combet [a] and Gilbert Deléage [a]
[a] *Institut de Biologie et Chimie des Protéines (IBCP UMR 5086); CNRS; Univ. Lyon 1;* IFR128 BioSciences Lyon-Gerland; 7, passage du Vercors, 69007 Lyon, France

Abstract: Bioinformatics analysis of data produced by complete genome sequencing projects is one of the major challenges of the current years. Integrating up-to-date databanks and relevant algorithms is a clear requirement of such analysis. Grid computing would be a viable solution to distribute data, algorithms, computing and storage resources for Genomics. Providing bioinformaticians with a good interface to grid infrastructure, such as the one provided by the EGEE European project, is also a challenge to take up. The GPS@ web portal, "Grid Protein Sequence Analysis", aims to provide such a user-friendly interface for these grid genomic resources on the EGEE grid.

Keywords: Bioinformatics, Grid computing, Tool integration, Web portal

Introduction

Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects [1], is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks (of sequences, patterns, 3D structures, *etc.*) and relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, *etc.*) [2]. Since 1998, we are developing the Web server NPS@ ([3], Network Protein Sequence Analysis), that provides the biologist with many of the most common resources for protein sequence analysis, integrated into a common workflow. These methods and data can be accessed through a HTTP connexion with a web browser, or bioinformatics program like MPSA [4] or AntheProt [5].

Today, the computing resources available behind the NPS@ Web portal limit the capabilities of our server as well as other genomics/post-genomics web portals. Indeed some methods are very computing-time and memory consuming. All these web portals have to face to an increasing demand of CPU and disk resources and to the management of the bioinformatics resources (algorithms, databanks). Most of the time, the portal administrators restrict user queries following different levels of access rights on the available methods and databanks.

Grid computing concept [6], as deployed in the European EGEE project [7], may be a viable solution to foresee these resources limitations [8] [9]. EGEE's goals are to build a European grid infrastructure, providing today users with more than 15,000

[1] Corresponding author: Institut de Biologie et Chimie des Protéines (IBCP UMR 5086), 7 passage du Vercors, 69007 Lyon, France; Christophe.Blanchet@ibcp.fr

CPUs. These resources are usable for grid users through specific components of this middleware: the user interface (UI), the job description language (JDL) and job workload management commands. Nevertheless EGEE user interface and usage are still raw and hardly accessible to non-computer scientist.

1. EGEE: European grid infrastructure

1.1. European project EGEE

The Enabling Grids for E-sciencE (EGEE [7]) project is funded by the European Commission and aims to build on recent advances in grid technology and develop a service grid infrastructure. EGEE aims to integrate current national, regional and thematic computing and data Grids to create a European Grid-empowered infrastructure for the support of the European Research Area, exploiting unique expertise generated by previous EU projects (DataGrid, CrossGrid, DataTAG, etc.) and national Grid initiatives (UK e-Science, INFN Grid, Nordugrid, GridIreland, etc.). The EGEE consortium involves 70 leading institutions in 27 countries, federated in regional Grids, with a combined capacity of over 15,000 CPUs, the largest international Grid infrastructure ever assembled.

1.2. EGEE infrastructure

The project EGEE is building a grid computing platform as it usually defined [6]: “a grid is a set of information resources (computers, databases, networks, instruments, etc.) that are integrated to provide users with tools and applications that treat those resources as components within a « virtual » system”.

EGEE middleware provides the underlying mechanisms necessary to create such systems, including authentication and authorization, resource discovery, network connections, and other kind of components. The platform is built on the LCG-2 middleware (Large Collisioner Grid release 2), which has been inherited from the EDG middleware developed by the European DataGrid Project ([10], FP5 2001-2003), initially based on the Globus toolkit [11].

The EGEE middleware permits grid users to launch a job on the EDG grid through an User Interface (UI). Then the job is processed by the workload management system in the Resource Broker (RB). This component, RB, determines where and when the submitted job have to be computed: on which given computing element and according to the needed storage element in case of simple jobs, using several of them in case of large jobs. A computing element (CE) is a cluster of several computing servers, the worker nodes (WN) managed by a scheduler system using batch mechanisms, such as PBS/Torque. A storage element (SE) is a server providing a storage space usable for distributing the application data around the grid. The resource broker knows the current state of the grid by querying the information system that centralizes all parameters raised by the grid components (cluster, storage, network,). When available resources have been chosen, the job is transferred to these components and launched. Once executed, the resource broker is informed and gets it back to the user interface.

1.3. EGEE usage: workload and data management

The usage of the EGEE middleware is still raw and hardly accessible to non-computer scientist. Firstly, the user has to connect to an user interface (UI) machine: getting an account on an existing UI or ,harder way , installing one in its laboratory. The UI needs a dedicated Linux machine, and the installation of the LCG-2 middleware is manual and needs some skills in system administration. Secondly, when the UI is up and ready, a grid user has to deal with the middleware command line interface (CLI) providing different sets of programs to manage job or data. Moreover, submitting a job means to write a valid JDL (Job Description Language) file describing completely the job to be run on the grid. The principal actions needed to run a job are the following: job submission (edg-job-submit) getting status (edg-job-status) and downloading results (edg-job-get-output). For data management the main programs are: data registration (lcg-cr), replication (lcg-rep) and data suppression (lcg-del). All these command are not integrated and need to be executed manually by the user.

2. Bioinformatics portal on the grid

As seen, the current job submission process on the EGEE platform is relatively complex, as well as non-automated, for non-computer scientist. Indeed, biologists who are using the grid have to submit their jobs manually, and have to check periodically the resource broker for the status of his job. A job goes through different steps during the workload management process: “Submitted", "Ready", “Scheduled”, “Running”, *etc.* until the “Done” status. And finally, they have to get the results with a raw file transfer from the remote storage area to the local file system of their user interface.

2.1. Gridification of Bioinformatics data and programs.

One major problem with a grid-computing infrastructure is the distribution of files and binaries, as BLAST [12] or ClustalW [13] through the job submission process. Sending a binary of the algorithm to a node on the grid is quite simple because of its size, few kilobytes, and can be done at each execution, although it isn’t the best efficient way to do it. But putting on the grid a databank, from tens of megabytes (as Swiss-Prot [14]) to gigabytes (as EMBL [15]), consumes a large part of network bandwidth if done at each job submission, and greatly enlarge the execution time if done each time a BLAST is submitted to the grid.

One simple solution can be to split databanks into subsets sent in parallel to several node of the grid, in order to run the same query on each subset. Another solution is to maintain used databanks on several storage elements (SE) of the grid and to launch the algorithm on computing resources (worker nodes) closer to these SEs.

According to these two solutions, the submission process is different. The algorithm submission processes implemented in our GPS@ portal have been adapted to the EGEE grid context. The algorithms and short datasets are sent at submission time through the grid sandbox process. While the other ones, algorithms analyzing large dataset are executed on grid nodes close to the related databanks, that have been replicated earlier or on demand through the replica management system.

2.2. GPS@ - Grid Protein Sequence Analysis.

The EGEE job submission could be boring for scientists that are not aware of advanced computing techniques. Thus, we decide to provide biologists with a user-friendly interface for the EGEE computing and storage resources, by adapting our NPS@ web portal [3]. The grid portal GPS@ ("Grid Protein Sequence Analysis) simplify and automated the EGEE grid job submission and data management mechanisms with XML descriptions of available Bioinformatics resources: algorithms and databanks (see Figure 1).

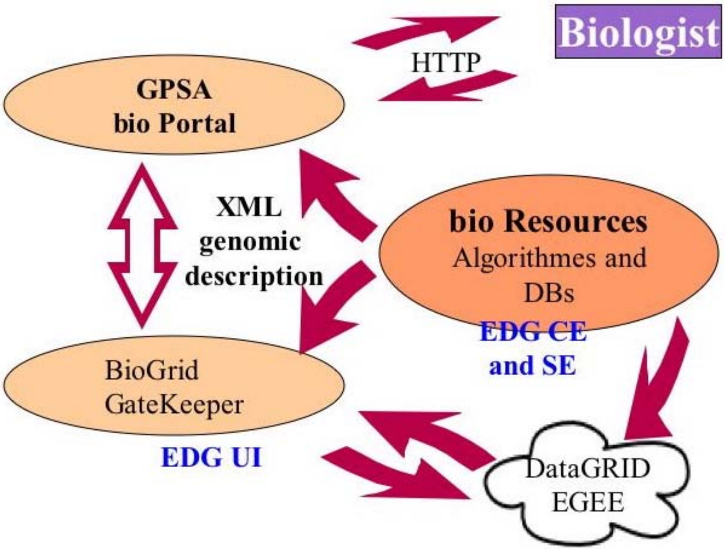


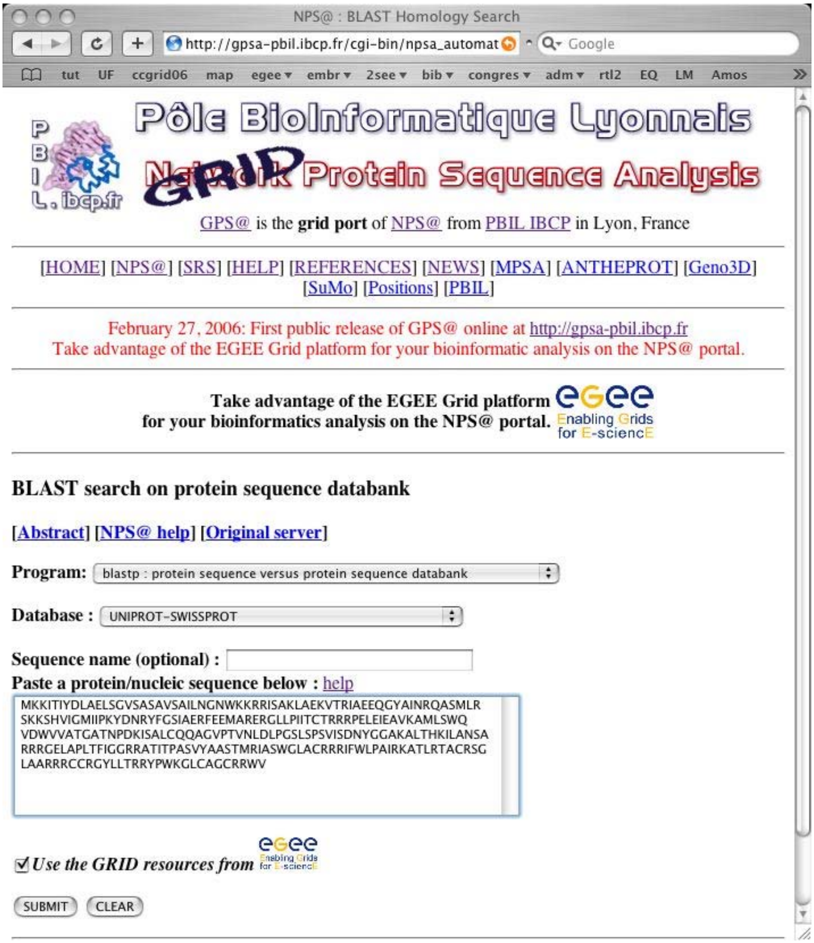
Figure 1. GPSA architecture and interface to the EGEE grid.

In GPS@, we simplify the grid analysis query: GPS@ Web portal runs its own EGEE low-level interface and provides biologists with the same interface that they are using daily in NPS@. They only have to paste their protein sequences or patterns into the corresponding field of the submission web page. Then simply pressing the "submit" button launches the execution of these jobs on the EGEE platform. All the EGEE job submission is encapsulated into the GPS@ back office: scheduling and status of the submitted jobs. And finally the result of the bioinformatics jobs are displayed into a new Web page (see Figure 2), ready for other analyses or for results download in the appropriate data format.

3. Example of use: submitting BLAST analyses to the EGEE grid

NPS@ [3] is providing biologist with a Web form to input their data (protein sequences) in order to run a BLAST analysis against a given protein sequence database. As in Figure 2, the user simply paste is sequence of protein in the corresponding field. Then he chooses the database that will be scan with the query sequence. All the available protein databanks can be selected through a multi-valued list of the form.

Selecting the “EGEE” check-box will schedule the submission of the BLAST on the EGEE grid when clicking on the “submit” button.


NPS@ : BLAST Homology Search
http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat
Google
tut
UF
ccgrid06
map
egee ▾
embr ▾
2see ▾
bib ▾
congres ▾
adm ▾
rtl2
EQ
LM
Amos
PBIL.ibcp.fr
Pôle BioInformatique Lyonnais
GRID
Network Protein Sequence Analysis
eGee
Enabling Grids for E-sciencE

GPS@ is the **grid port** of NPS@ from PBIL IBCP in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.ibcp.fr
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the NPS@ portal.

Take advantage of the EGEE Grid platform
for your bioinformatics analysis on the NPS@ portal.

BLAST search on protein sequence databank

[Abstract] [NPS@ help] [Original server]

Program: blastp : protein sequence versus protein sequence databank

Database : UNIPROT-SWISSPROT

Sequence name (optional) :

Paste a protein/nucleic sequence below : help

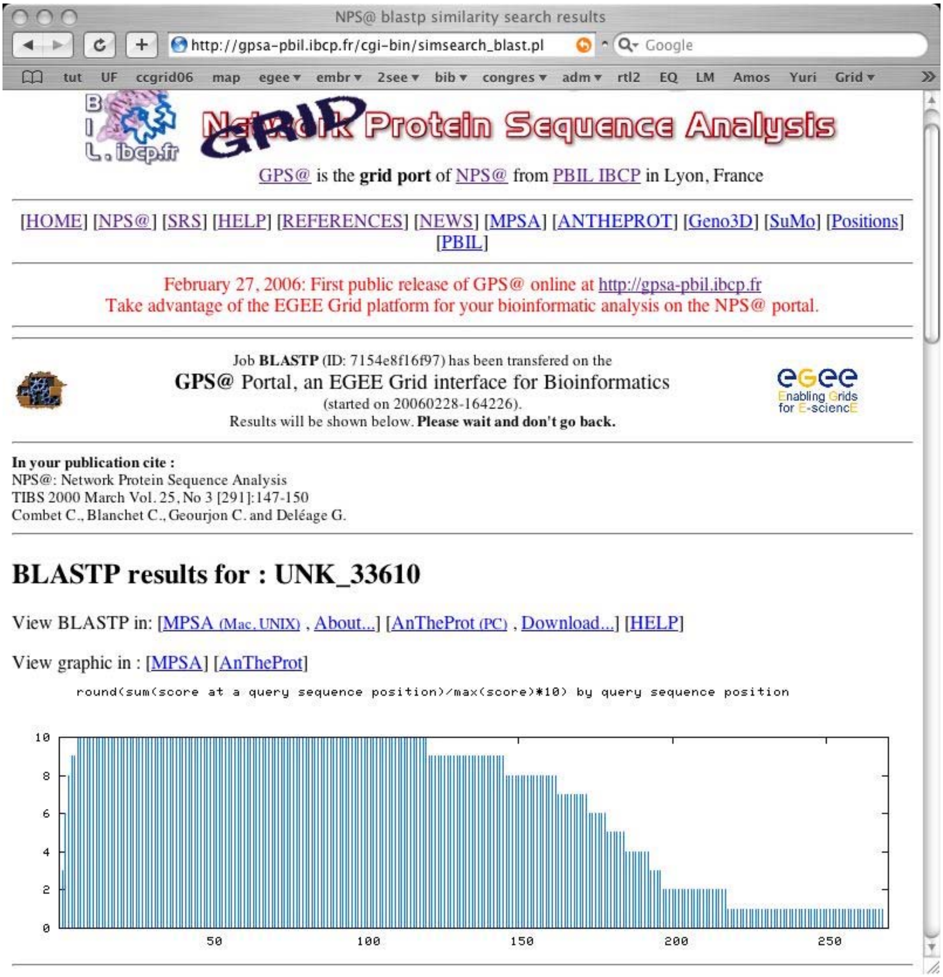
```
MKKITIYDLAELSGVSASAVSAILNGNWKKRRISAKLAEKVTRIAEEQGYAINRQASMLR
SKKSHVIGMIIPKYDNRYFGSIAERFEEMARERGLLPIITCTRRRPELEIEAVKAMLSWQ
VDWVVATGATNPDKISALCQQAGVPTVNLDLPGSLSPSVISDNYGGAKALTHKILANSA
RRRGELAPLTFIGGRRATITPASVYAASTMRIASWGLACRRRIFWLPAIRKATLRTACRSG
LAARRRCCRGYLLTRRYPWKGLCAGCRRWV
```

☑ *Use the GRID resources from*

SUBMIT CLEAR

Figure 2. GPS@ web portal: submission form of a bioinformatics analysis with BLAST.

As the GPS@ portal integrated is own EGEE user interface (see Figure 1), an automatized process then submits the BLAST job on the grid. First, the job description in the Web form is converted into a JDL file, that can then be submitted to the workload management system of EGEE. The GPS@ sub-process that have submitted the job, is also checking periodically the status of this job by querying the resource broker with the good commands. All steps are notified to the user through the Web page of the submission, indicating the time and the duration of the current step. When achieved, *i.e.* reaching the “Done” step, the GPS@ automat is downloading the result file from BLAST. Then this raw result file in BLAST format is processed and converted into a HTML page showing, in a colored and graphical way, the list of similar protein sequences, and also graph and pairwise alignments of them (as in Figure 3). This formatting process is directly inherited from the original NPS@ portal, providing biologists with a well-known interface and way of displaying results.


NPS@ blastp similarity search results
http://gpsa-pbil.ibcp.fr/cgi-bin/simsearch_blast.pl
Google
tut
UF
ccgrid06
map
egee
embr
2see
bib
congres
adm
rtl2
EQ
LM
Amos
Yuri
Grid
BIL.ibcp.fr
GRID
Network Protein Sequence Analysis
egee
Enabling Grids for E-sciencE
10
8
6
4
2
0
50
100
150
200
250

GPS@ is the **grid port** of NPS@ from PBIL IBCP in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.ibcp.fr
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the NPS@ portal.

Job **BLASTP** (ID: 7154e8f16f97) has been transfered on the
GPS@ Portal, an EGEE Grid interface for Bioinformatics
(started on 20060228-164226).
Results will be shown below. **Please wait and don't go back.**

In your publication cite :
NPS@: Network Protein Sequence Analysis
TIBS 2000 March Vol. 25, No 3 [291]:147-150
Combet C., Blanchet C., Geourjon C. and Deléage G.

BLASTP results for : UNK_33610

View BLASTP in: [MPSA (Mac, UNIX) , About...] [AnTheProt (PC) , Download...] [HELP]

View graphic in : [MPSA] [AnTheProt]

round(sum(score at a query sequence position)/max(score)*10) by query sequence position

Figure 3. GPS@ web portal: results of a BLAST scan for protein sequence similarity.

4. Conclusion

GPS@ grid web portal (Grid Protein Sequence Analysis, http://gpsa-pbil.ibcp.fr) is a Bioinformatics integrated portal such as the current NPS@ protein portal, and would provide the biologist with a user-friendly interface for the GRID resources (computing and storage) made available by the project EGEE (2004-2005).

This genomic grid user interface hides the mechanisms involved for the execution of Bioinformatics analyses on the grid infrastructure. The bioinformatics algorithms and databanks have been distributed and registered on the EGEE grid and GPS@ runs its own EGEE interface to the grid. In this way, GPS@ portal simplify the Bioinformatics grid submission, and provide biologist with the benefit of the EGEE grid infrastructure to analyze large biological dataset: e.g. including several protein secondary structure predictions into a multiple alignment, or clustering a sequence set by analyzing, with BLAST or SSEARCH, each sequence against the others, …

In the future, main efforts should be focused on taking bioinformatics specific constraints and requirements into account on the EGEE grid. That means, for example, including ontology and semantic parameters into the gridified data with the replica manager system. An other effort should concern the security of the bioinformatics data and methods on the grid: encryption of data, network isolation and algorithm execution sandboxing, fine grain access to data, monitoring private data transfer and replication, *etc.*

Acknowledgements

This work has been funded by GriPPS project (ACI GRID PPL02-05), EGEE project (EU FP6, ref. INFSO-508833) and EMBRACE Network of Excellence (EU FP6, LHSG-CT-2004-512092).

References

[1] Bernal, A., Ear, U., Kyrpides, N. : Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. NAR 29 (2001) 126-127

[2] G. Perrière, C. Combet, S. Penel, C. Blanchet, J. Thioulouse, C. Geourjon, J. Grassot, C. Charavay, M. Gouy, L. Duret and G. Deléage, Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res., 31:3393-3399, 2003.

[3] Combet, C., Blanchet, C., Geourjon, C. et Deléage, G. : NPS@: Network Protein Sequence Analysis. Tibs, 25 (2000) 147-150.

[4] Blanchet, C., Combet, C., Geourjon, C. et Deléage, G. : MPSA: Integrated System for Multiple Protein Sequence Analysis with client/server capabilities. Bioinformatics, 16 (2000) 286-287.

[5] Deleage, G, Combet, C, Blanchet, C, Geourjon, C. : ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. Comput Biol Med., 31 (2001) 259-267

[6] Foster, I. And Kesselman, C. (eds.) : The Grid: Blueprint for a New Computing Infrastructure, (1998).

[7] EGEE – Enabling Grid for E-science in Europe; http://www.eu-egee.org

[8] Vicat-Blanc Primet, P., d'Anfray, P., Blanchet, C., Chanussot, F. : e-Toile : High Performance Grid Middleware. Proceedings of Cluster'2003 (2003).

[9] Jacq, N., Blanchet, C., Combet, C., Cornillot, E., Duret, L., Kurata, K., Nakamura, H., Silvestre, T., Breton, V. : Grid as a Bioinformaticstool. , Parallel Computing, special issue: High-performance parallel bio-computing, Vol. 30, (2004).

[10] EDG - European DataGrid project, http://www.eu-datagrid.org

[11] GLOBUS Project, http://www.globus.org/

[12] Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ : Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403-410

[13] Thompson, JD, Higgins, DG, Gibson, TJ : CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (1994) 4673-4680.

[14] Bairoch, A, Apweiler, R : The SWISS–PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 27 (1999) 49-54

[15] Stoesser, G, Tuli, MA, Lopez, R, Sterk, P : the EMBL nucleotide sequence database. Nucleic Acids Res. 27 (1999) 18-24.