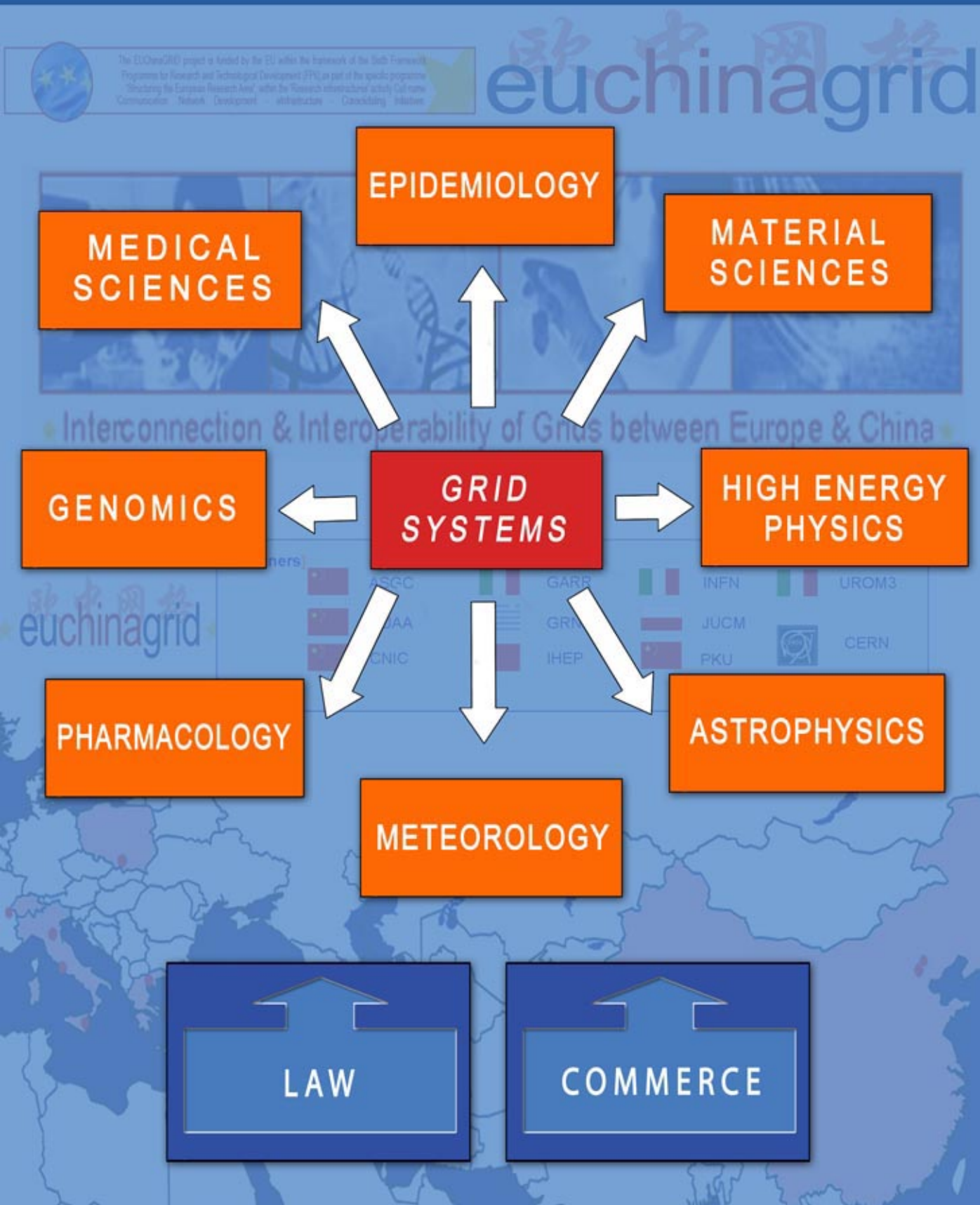
BIO-ALGORITHMS AND MED-SYSTEMS

Vol. 3, No. 5
2007



JOURNAL EDITED BY MEDICAL COLLEGE - JAGIELLONIAN UNIVERSITY – KRAKOW – POLAND
http://www.bams.cm-uj.krakow.pl

EDITORIAL BOARD

Polish Ministry of Science and Higher Education journal rating: 3.000
Punktacja KBN: 3.000

ISSN 1895-9091 (print version)
ISSN 1896-530X (electronic version)

http://www.bams.cm-uj.krakow.pl

Contents

OPENING ARTICLE

3 Grid Projects at Academic Computer Center CYFRONET AGH, Krakow
M. Kwaśniewski

GRIDS IN SCIENCE

7 Grid Computing in Peking University
S. Zhu, S. Qian

17 GRID: from HEP to e-Infrastructures
F. Ruggieri

23 Grid Infrastructures as Catalysts for Development on e-Science: Experiences in the Mediterranean
G. Andronico, R. Barbera, K. Koumantaros, F. Ruggieri, F. Tanlongo, K. Vella

27 RandomBlast a tool to generate random “never born protein” sequences
G. Evangelista, G. Minervini, P.L. Luisi, F. Polticelli

33 A solution for data transfer and processing using a grid approach
A. Budano, P. Celio, S. Cellini, R. Gargana, F. Galeazzi, C. Stanescu, F. Ruggieri, Y.Q. Guo, L. Wang, X.M. Zhang

39 High throughput protein structure prediction in a grid environment
G. Minervini, G. La Rocca, P.L. Luisi, F. Polticelli

45 An approach to protein folding on the grid – EUChinaGrid experience
M. Malawski, T. Szepieniec, M. Kochanczyk, M. Piwowar, I. Roterman

51 Massive identification of similarities in DNA materials organized in Grid environment
M. Piwowar, T. Szepieniec, I. Roterman

53 Computers in medicine
J.K. Loster, A. Garlicki, M. Bociąga, P. Skwara, A. Kalinowska-Nowak

SHORT COMMUNICATION

57 Grids and their role in supporting worldwide development
F. Tanlongo

59 Grids at 4300 meters over the sea level: argo on EUChinaGrid
C. Stanescu, F. Ruggieri, Y.Q. Guo, L. Wang, X.M. Zhang

61 Euchinagrid: a high-tech bridge across Europe and China
F. Tanlongo

63 Radiology on Grid
A. Urbanik

65 Grid monitoring in EUChinaGrid infrastructure
Lanxin Ma

67 SELVITA

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 3-5

OPENING ARTICLE

GRID PROJECTS at ACADEMIC COMPUTER CENTER CYFRONET AGH, KRAKOW

MAREK KWAŚNIEWSKI

Academic Computer Center CYFRONET AGH, Nawojki 11, 30-950 Krakow, Poland.

Academic Computer Centre CYFRONET AGH, established over 30 years ago, is an autonomous organizational and financial entity of the AGH University of Science and Technology. The Centre belongs to the largest computer centers in Poland oriented on supercomputing and net systems. The organization system of the center: High-Performance Computing Department, Software Department, Computer Networks Department, Storage & Security Data Department, Technical Department, Administration Department, Financial and Accounting Department and the Operators Section ensure the exploitation as well as development of academic computer network as well as large scale computing service.

CYFRONET is responsible for:

1. Provision of computing power and other computing-related services to the scientific community acting in research & education;
2. Development, maintenance and extension of computing infrastructure;
3. Participation in programs supported by the Polish government in the area of application of new information technologies for science, education, management and business;
4. Scientific research (individually and in collaboration with other academic communities) in the field of high-performance computers application and computer network systems services;
5. Research, analysis and implementations of new technologies applicable to the design, creation and maintenance of computer infrastructures;
6. Consultations, services and training courses in the field of information technology, computer networks and high-performance computing;
7. Promotion of new solutions for science, education, management and business to make them more innovative;

CYFRONET has been participating in many projects of EU IST: to FP5 and FP6.

Ambient Networks - strategic objective of "Mobile and Wireless Systems Beyond 3G".

GREDIA – creation of a reliable Grid application development platform with high-level support for the design, implementation and operational deployment of secure Grid business applications.

ViroLab - virtual laboratory for studying infectious diseases including HIV virus resistance to drugs in particular.

int.eu.grid - Interactive European Grid Project's objective - advanced Grid empowered infrastructure in the European Research Area for application in: medicine, environment, astronomy and physics.

EGEE - Enabling Grids for e-Science in Europe – integration of current national, regional and thematic Grid efforts in order to create a seamless European Grid infrastructure for the support of the European Research Area (ERA).

K-WfGrid - Knowledge-based Workflow System for Grid Applications - addresses the need for a better infrastructure for the future Grid environment.

CoreGRID - the CoreGRID Network of Excellence (NoE) - strengthening and advancing scientific and technological excellence in the area of Grid and Peer-to-Peer technologies.

CrossGrid - international project focusing on applications whoich requirefrequent interaction with the user and real time responses from a system: distributed data analyzis uin High Energy Physics, surgery decision support application, weather forecasting, flood crisis tean decisions support system.

GridStart - clustering all of the 5FP IST-funded Grid research projects with the intention to stimulate wide deployment of appropriate technologies and to support early adoption of best practices.

Pellucid - Platform for Organizationally Mobile Public Employees (EU 5FP).

Pro-Access - The ImPROving ACCESS of Associated States To Advanced Concepts In Medical Informatics (PRO-ACCESS) - creation of a platform for promotion, dissemination and transfer of advanced health, telematics and experiences from development and deployment of telemedicine solutions to NAS.

And national projects:

CLUSTERIX - this project contains the concept of building the National Cluster of Linux Systems

PROGRESS - Polish Research On GRid Environment for Sun Servers.

The Cracow Telemedicine Centre - collaboration with hospitals and health care centers to upgrade the medical services introducing new technologies implementing IT solutions for scientific projects sponsored by the Polish Ministry of Science, as well as by EU 5FP and practical medicine.

PIONIER programme bases on the document PIONIER: Polish Optical Internet - Advanced Applications, Services and Technologies for Information Society". Introducing Poland to international communities and allowing the partnership in collaboration with other countries.

High Performance Computing and Visualization with the SGI Grid for Virtual Laboratory Applications - the SGI Grid project aims to design and implement the innovative activities and technologies.

Since 2000 CYFRONET has been organizing the yearly Cracow Grid Worshops.

The last one, planed to be organized together with the project EuChinaGRID:

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 7-15

GRID COMPUTING IN PEKING UNIVERSITY

SHULEI ZHU, SIJIN QIAN

Peking University, Beijing, China

Abstract: Grid computing enables the massive computer resource sharing, so that many applications (e.g. experimental high energy physics (HEP) and biology researches, etc.) can be greatly benefited from this new technology to proceed to the level which was unthinkable or unreachable before. Peking University is one of 10 partners in the EUChinaGRID project funded by European Commission. In this paper, the BEIJING-PKU site (based on the middleware gLite of European grid project EGEE) in the EUChinaGRID infrastructure is described. Some result of grid application in Peking University and our future plans (on HEP and biology applications as well as on the grid technology development itself) are outlined.

Key Words: EGEE, LCG, gLite3, EUChinaGRID, Interoperability

1. Introduction

Grid computing, a newly developing technology after the internet and WWW, harnesses the distributed computer resources to facilitate with collaboration, data sharing and management of all resources involved. In fact, all resources in the computing grid environment are virtualized to create a pool of assets for authorized users to retrieve seamlessly. With the grid computing, it becomes possible to solve many problems too intensive for any stand-alone computers or computer clusters. For end-users, by accessing the computing grid they seem to hold vast IT capabilities [1]; similar as the electric power grid, the users would not need to care where the resources (e.g. the power station or electric generator and power line, etc.) are located. Currently, in some scientific organizations and communities, researchers may use the computing grid infrastructures shared in the Virtual Organization (VO, see the next Chapter) as long as they join the VO, even with free of charge; but this situation will evolve to be similar as the electronic power grid once the accounting services in the middleware of computing grid shall be more mature.

Peking University (PKU) is one of 10 partners of the EUChinaGRID project funded by European Commission under the 6th Framework Programme (FP6) for Research and Technological Development. PKU group is consists of two subgroups, one is the biology group led by Prof. Bin XIA, another is the High Energy Physics (HEP) group led by Prof. Sijin QIAN. Among 5 Working Packages (WPs) of EUChinaGRID project, PKU group participated in WP3 (pilot infrastructure operational support), WP4 (grid application) and WP5 (dissemination). Within the scope of WP3, a grid site of BEIJING-PKU has been built since the beginning of 2007. PKU group's activities in WP4 include the biology and HEP applications. We have heavily engaged in the dissemination work in WP5, including to host a tutorial at PKU in November of 2006.

In this paper, Chapter 2 is to further elaborate the grid computing and the virtual organization, as well as the two major projects (LCG and EGEE) for the HEP and other scientific applications, and the brief of EUChinaGRID project; Chapter 3 is to describe the middleware "gLite" of EGEE system which is installed at PKU; Chapter 4 explains the status of grid site BEIJING-PKU and some result from the HEP application obtained by PKU group; Chapter 5 is to outline the future plan in PKU group on the biology application and on the computing grid technology; the summary is given in Chapter 6.

2. Grid computing and 3 relevant Grid projects (LCG, EGEE and EUChinaGRID)

Grid computing is an evolution of related development in information technology, such as p2p (Peer to Peer), distributed computing and so on. It shares many common grounds with these technologies and works as a combination to climb to a level which the individual precedent technology could not reach. Grid computing has many features such as distributed, dynamical, diversity, self-comparability, autonomic and multiple management, etc. Therefore, Ian Foster "defined" the grid computing as "Flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources (i.e. "Virtual Organizations", VO, see the next Section) [2]. Here the resource includes computers, data storages, databases, sensors, networks and software, etc. A "VO" can be conceived as a group of people (and resources) belonging to the same or different real organizations that want to share common resources in order to achieve the goals which are unreachable by each individual alone.

From view point of application, grid computing may be classified into data grid, computational grid, collaboration grid, information grid, knowledge grid and semantic grid, etc. In

reality, many grid systems can be a combination of some above types.

At present, some stable computing grids have been being tested in scientific fields. They play (or are going to play) important roles in solving some complex and important problems encountered by researchers. On the other hand, people believe that the computing grid also could be used in the enterprises to increase the productivity and efficiency in the organizations and may help to solve the security problems too. IBM, Microsoft, Oracle and other global IT enterprises response to this growing technology actively and inject increasingly more efforts to its development.

2.1. Virtualization of grid computing

By virtualization, the grid computing enables across network heterogeneous IT systems to work together to form a large virtual computing system offering a variety of virtual resources [3]; and the concept of Virtual Organization (VO) contribute the essence in the development and the application of grid computing.

VOs are some dynamical virtual entities which correspond to real organizations or projects, such as IT department of global enterprises, the four experiments (ATLAS, ALICE, CMS and LHCb) on Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research, in Geneva, Switzerland), the community of biomedical researchers and so on. VOs strictly enforce security rules to their members which regulate the privileges and priorities between users and resources. In VOs, members share all kinds of resources including equipments, software, hardware, licenses and others. Of course, these resources are virtualized and dynamically assembled.

Figure 1 describes the relation between a VO and some real organizations [4]. Some resources (including personnel) in the real organizations are contributed to a big virtual world which collects the contributions from all real organizations to form a big pool so that all resources in the pool can be shared by the members in this VO under some agreed rules and strict security measures.

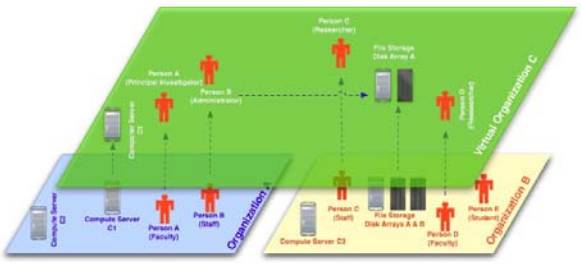


Fig. 1. An illustration of the VO with respect to real organizations

2.2. LCG, EGEE projects and the grid application in high-energy physics

Currently being built and soon-to-be one of the largest scientific instrument in the world, the Large Hadron Collider (LHC) will hopefully be completed and be operational at the beginning of 2008; it will produce roughly 12-14 Petabytes (1Petabytes = 1 million Gigabytes, if being stored in normal CDs, the accumulation of CDs for 1 PB of data will be piled up to several kilometers tall) of data annually, which will be distributed around the globe and analyzed by thousands of scientists in some 500 research institutes and universities worldwide that are participating in the LHC experiments. About 100 000 CPUs at 2004 measures of processing power are required to simulate and analyze these data. No any single computer or supercomputer center in the world can satisfy the requirement to analyses and store the data.

LCG (LHC Computing Grid) project emerged in 2002, as Prof. Les Robertson (CERN's LCG project manager) said "The LCG will provide a vital test-bed for the new Grid computing technologies that are set to revolutionize the way scientists use the world's computing resources in areas ranging from fundamental research to medical diagnosis" [5]. The data from the LHC experiments will be distributed around the globe according to a four-tiered model. The Tier-0 centre of LCG is located at CERN; those data which arrive at Tier-0 will be quickly distributed to a series of Tier-1 centers after initial processing, then continuously to the Tier-2s and Tier-3s. BEIJING-PKU site [6] will act as a part of Tier-3s, which can consist of local clusters in a Department of University or even of individual PCs, and which may be contributed to LCG on a regular basis [7]

The core task of implementing LCG project is the development of grid middleware. Nowadays, the heterogeneous IT systems are not compatible with the model of computing grid; therefore we need an extensible system, called as grid middleware, to enable the interaction of grid and existing network. The “grid middleware” refers to the security, resource management, data access, instrumentation, policy, accounting, and other services provided for applications, users, and resource providers to operate effectively in a Grid environment. Middleware acts as a sort of 'glue' which binds these services together [8]. LCG project had studied and deployed the grid middleware packages which come from some components developed by other projects and organizations, such as EDG (European DataGrid), Globus, Condor, PPDG, GriPhyN and others. The middleware widely distributed at CERN and the LHC community latter gradually has been replaced by the “gLite” middleware that is maintained and developed by EGEE (Enabling Grids for E-Science in Europe) project.

EGEE is another important European project which was started in April 2004 and aims to establish a Grid infrastructure for e-science (in European first, then later beyond Europe), and its goal is to provide researchers with access to a geographically distributed computing grid infrastructure, available around clock. LCG contributed to the initial environment for EGEE: the gLite3 middleware of EGEE comes out as the fruit of convergence of LCG 2.7.0 and gLite 1.5.0 in the spring of 2006. One major difference between two middleware is that LCG middleware focused on data handling but gLite3 does on data analysis.

The site of BEIJING-PKU has been upgraded to gLite3 by following the general trends. So we will focus on gLite3 middleware because it includes the complete components inherited from LCG-2.

2.3. EUChinaGRID project and Peking University

EUChinaGRID project focuses on extending the European GRID infrastructure for e-Science to China and strengthening the collaboration between China and Europe in computing grid field [9]. Interoperability between two middleware, i.e. gLite3 of EGEE and GOS (Grid Operation System) of CNGrid (China National Grid) is one of the key goals of the project which will be introduced in Chapter 5.

As introduced in Chapter 1, Peking University group has been mainly engaged in 3 among 5 Working Packages (WPs) of EUChinaGRID project. Within the scope of WP3 (pilot infrastructure operational support), we have set up a fully functional grid site BEIJING-PKU which is going to be described in rather details in Chapter 4.

Two subgroups in PKU are participating in WP4 (grid application) of EUChinaGRID pertaining to different disciplines of sciences: Biology and Physics. The Beijing Nuclear Magnetic Resonance Center (BNMRC) is a national center for bio-molecular structural studies in China located at PKU; this group will make use of new grid technology to enhance the quality of Never-Born-Protein (NBP) applications. The PKU high energy physics (HEP) group has participated in the CMS experiment on LHC at CERN since 11 years ago; it will use the computing grid on the huge amount of Monte-Carlo event generation and data analysis. Some results obtained by HEP group will be shown in Chapter 4.

In WP5 (dissemination) of EUChinaGRID, we have taken part in organizing the training and other activities (e.g. to briefing the journalists and medias for their participation in the project conference, to making the presentations at various international grid conferences, etc.). In November of 2006, PKU has hosted a Grid tutorial taught by all Chinese tutors (in its first time) and it got one of the highest feedback scores evaluated by the trainees.

EUChinaGRID project is preparing to apply for the extension under the 7th framework programme (FP7) of EC. Hopefully, more partners would be able to join the second term of project; also we would be able to continue our activities and some new foreseen programs as outlined in Chapter 5.

3. gLite Grid middleware

Some architectures of middleware were designed after the proposition of computing grid concept, such as Five-Level Sandglass Architecture designed by Ian Foster, OGSA (Open Grid Services Architecture) and WSRF (Web Service Resource Framework), etc. Of them, Five-Level Sandglass Architecture is the most significant one, which leads to the definition of grid protocol architecture. This model focus on the protocol, but it also emphasizes the services, e.g. API (Application Programming Interfaces) and SDK (Software Development Kits) are two aspects considered much by this model.

Just as its name implies, five components layers are included in the Five-Level Sandglass Architecture [2]. Starting from the bottom of the stack and moving upward, they are fabric layer, connectivity layer, resource layer, collective layer and application layer. The “fabric layer” defines the interface to local shared resources; the “connectivity layer” defines the basic communication and authentication protocols required for grid-specific networking-service transactions; the “resource layer” uses the communication and security protocols (defined by the connectivity layer) to control secure negotiation, initiation, monitoring, accounting, and payment for the sharing of functions of individual resources; the “collective layer” is responsible for all global resource management and interaction with collections of resources; and the “application layer” enables the use of resources in a grid environment through various collaboration and resource access protocols. Thus it can be seen that there are some evident differences between grid protocol and internet TCP/IP protocol (Fig. 2) [10]

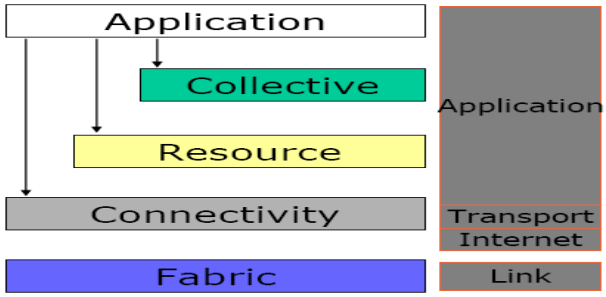


Fig. 2. Differences between the grid protocol (left) and the internet TCP/IP protocol (right)

Another important grid architecture “OGSA” is likely to become the standard of grid protocol. OGSA is a kind of Service Oriented Architecture (SOA), which concerns with the description of the services that have a network-addressable interface and that communicate with protocols and data formats. OGSA receives the strong help from the Globus project which provides a collection of Grid services that follow OGSA architectural principles and a development environment for producing new Grid services that follow OGSA principles.

From Five-Level Sandglass Architecture to OGSA, the essential change is from the models of function-based to of service-oriented. The gLite middleware is developed with this background and is the representative of second generation grid middleware.

The gLite3 Middleware [11] developed by EGEE project follows SOA architecture, which share many standards and services with OGSA. Therefore, it is compatible with the OGSA and this would be important if OGSA would become the standard of grid protocol. The services work together in a coherent way as an integrity component but they can also be deployed independently, this allows their development in different contexts. The architecture [12] of gLite3 middleware is shown in Fig.3 and is described in more details for each system in next sub-sections.

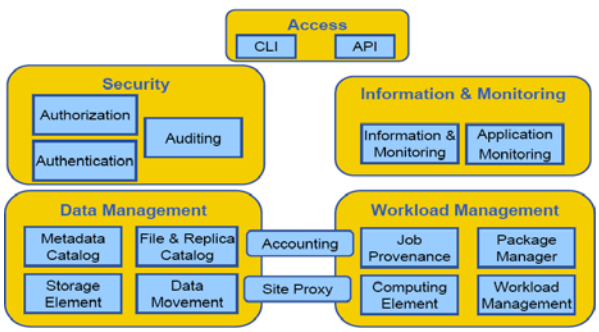


Fig. 3. The gLite3 architecture

3.1. Security Service

To ensure the security of grid system, there must be some forceful security rules so that only users with privileges and authorization are allowed to access it.

In gLite middleware, authentication is based on X.509 PKI infrastructure which is issued by Certificate Authorities (CA). The certificates will work like a passport to identify individuals. A user or host holding the certificate has a private key protected by password to prove the identity. Submitting jobs to remote hosts with private key may not be very safe. In order to reduce vulnerability, a proxy is used to connect to the remote hosts on behalf of the user. Proxies and private keys are vital to users or hosts because persons who steal them can impersonate the owner.

As explained above, the user management in gLite middleware is realized by VOs. A user must read and agree to the usage rules and any further rules for the VO he (or she) wishes to join and register some personal data with a Registration Service in order to use resources of the VO. VOMS (VO Management Service) is responsible to manage information about the roles and privileges of users within a VO.

Though certificate is not a short lived authentication, it has the expiration date, after which the certificate is no longer valid and users have to renew the certificate from CAs. However, proxies usually have a lifetime of only a few hours. To manage a large job, the user must extend the lifetime of proxy first.

3.2. CE (Computer Element) and Workload Manage System (WMS)

The Computing Element (CE), including Grid Gate (GG), BLASH, Local Resource Management System (LRMS), Work Nodes (WNs) and other components, mount computing resources, therefore represents the power, of a grid site. Here, GG is the generic interface to the computer cluster and BLASH is the interface passing the job to a layer that interacts with the local resource manager; the executable jobs submitted to CE will queue in LRMS to wait to be dealt with by WNs. There should be some VO-specific application software pre-installed at the grid sites in a dedicated area which WNs can access.

Jobs assigned to CE are firstly selected by RB (Resource Broker) that is the machine where WMS (Workload Management System) services run. RB chooses CE according to the information of Job Description Language (JDL) file provided by the job submitter, and the Logging and Bookkeeping service (LB) tracks history and status of jobs managed by the WMS.

3.3. SE (Storage Element) and Data Management Service (DMS)

SE (Storage Element) provides the interface to allow a user or an application to store data. The Storage Resource Manager (SRM) has been designed to be the single interface (through the corresponding SRM protocol) for the management of disk and tape storage resources which can be the single disk server or disk array or MSS (Massive Storage System). Any type of Storage Element offers an SRM interface except for the Classic SE, which is becoming obsolete by being phased out; now in gLite3 [11], SRM has been migrated to v2.2 which can hide the storage system implementation from users, and it can check the access rights to the storage system and the files.

Table 1. Types of SE in gLite3

Type	Resources	File transfer	File I/O	SRM
Classic SE	Disk server	GSIFTP	insecure RFIO	No
DPM	Disk pool	GSIFTP	secure RFIO	Yes
dCache	Disk pool/MSS	GSIFTP	gsidcap	Yes
CASTOR	MSS	GSIFTP	insecure RFIO	Yes

In gLite SE, GSIFTP (a GSI-secure FTP) is the protocol for whole-file transfers, while RFIO (Remote File Input/Output) or gsidcap is for local and remote file. In addition, normally a monitoring service of “MON Box” is installed on the computer where SE is installed to be responsible for the monitoring of whole system.

3.4. Information System (IS)

A Grid site publishes and monitors grid resources and their status with Information System (IS). For users, IS help them to find the best place to submit jobs; while for administrators, more intuitionist information (e.g. to trace the execution status of CE and to check the available storage space in SE, etc.) can be found in IS.

IS publish much of the data conforming to the GLUE（Grid Laboratory for a Uniform Environment） Schema（which defines a common conceptual data model to be used for Grid resource monitoring and information finding. There are two types of IS in gLite3: “Monitoring and Discovery Service (MDS)” and “Relational Grid Monitoring Architecture (R-GMA)”, more details can be found in [11].

4. BEIJING-PKU site and Grid application on HEP in Peking University

Along with the development of grid computing technology, the grid computing team of Peking University mainly considers itself as a grid user. Our aim is to run a stable site, to exploit more computing and data storage resources when needed, to offer our spare resources (whenever available) to other users and to make full use of the grid for the tasks in the high energy physics and biology researches. This quite coincides to the objectives of EUChinaGRID project.

4.1. BEIJING-PKU grid computing site

The construction of BEIJING-PKU site was started in the middle of 2006, and become almost fully functional in the Spring of 2007 after the bottleneck problem of international network connection has been solved. It should be emphasized that the construction of this site would not be successful if without the help from experts of EUChinaGRID project. Fig.4 shows the layout of the site. The assignment of computer hosts is listed in Table 2. The site now can be constantly detected by and shown at the GridICE monitoring system (Fig.5)

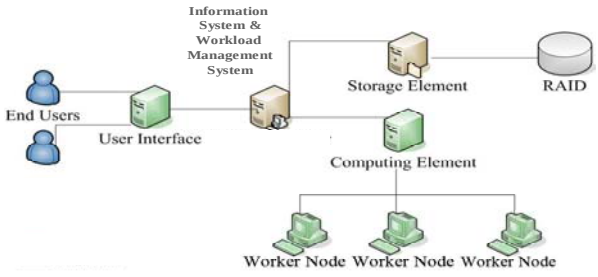


Fig. 4. Topological layout of BEIJING-PKU site

Table 2. The assignment of hosts in BEIJING-PKU site

Host	Compo-nents	Middleware version	system	Remark
grid.$MYDOAIN	UI	gLite3_0_0	SLC308	
grid01.$MYDOMAIN	SE+MON	gLite3_0_0	SLC308	
grid03.$MYDOMAIN	WN1	gLite3_0_0	SLC308	no host certificate
grid04.$MYDOMAIN	CE+SB	gLite3_0_0	SLC308	
grid06.$MYDOMAIN	WN	gLite3_0_0	SLC308	no host certificate
grid07.$MYDOMAIN	RB	gLite3_0_0	SLC308	

Where $MYDOMAIN=phy.pku.edu.cn SLC=Scientific Linux CERN

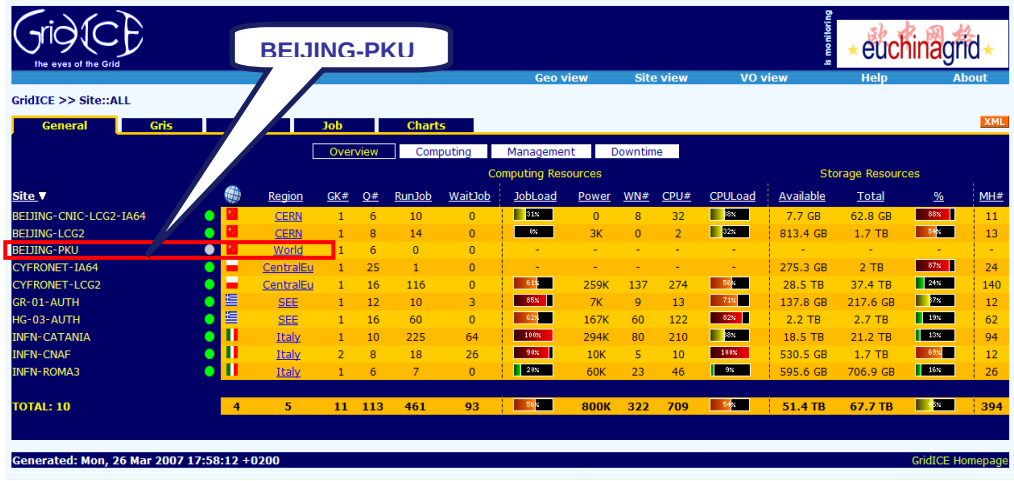


Fig. 5. BEIJING-PKU site is detected by GridICE monitoring system

The site has been tested repeatedly. As a small-scale site, at this stage we have not installed all components of gLite3 yet, but only some key components which will be helpful for the robustness and stableness.

4.2. Grid application on HEP in Peking University and our physics goal

Due to the huge amount of data going to be collected from LHC which is scheduled to collide the proton beams with the highest energy in the world in less than 6 months from now, the PKU physics group must be ready for analysing these data, not only the real data collected by CMS detector from the middle of 2008, but also the Monte-Carlo (MC) data (with the similar amount as the real experimental data) from now on. The PKU physics group has worked on this application in following aspects:

- established the BEIJING-PKU site for getting access to the LCG system;
- used the above system to have analysed a large MC dataset stored at CNAF in Italy, and have produced some result;
- provided a configuration file for CMS collaboration in order to generate at least 1 million prompt J/ψ events.
- has estimated the computer and storage resources needed to handle these 1 million events.

The physics goal of PKU-CMS group is to use the heavy Quarkonia (J/ψ or ϒ) for verifying the Non-Relativistic Quantum ChromoDynamics (NRQCD). In the past, normally the p-p colliding beam experimental data can be explained approximately by the Color Singlet Model (CSM) of NRQCD, but CSM has large discrepancy (Fig. 6) on the high transverse momenta J/ψ production rate from the CDF experimental data on Tevatron (a proton-antiproton collider) at Fermilab.

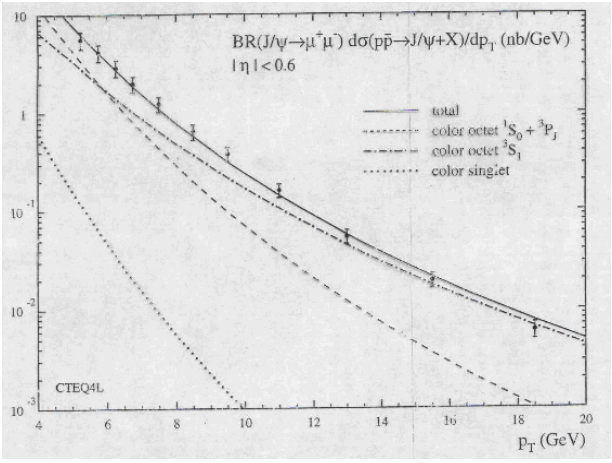


Fig. 6. J/ψ Production rates & NRQCD

In contrast, if a Color Octet Mechanism (COM) is introduced, CSM + COM together can fit the experimental data much better. However, when use the COM to predict the J/polarization, the COM is still not coincide the data from CDF experiment (Fig. 7)

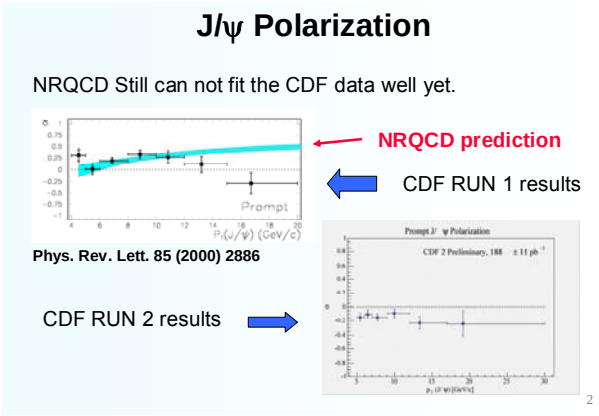


Fig. 7. J/ψ Polarization

With LHC's high luminosity (100 times higher than Tevatron) and high energy (7 times higher than Tevatron), the larger statistics of data are hopefully to help to solve the J/ψ polarization puzzle.

4.3. Result of analysing the large Bs event data set by using Grid tools

The huge amount (expected in the order of several PetaBytes per year) of CMS data have been (and are going to be) distributed at many places around world, We have used the BEIJING-PKU grid site to submit the jobs for analysing a large data set stored in Italy (as shown in Fig.4 below)

After analyzing nearly 20,000 events in a Bs → J/ψ + φ event data set (stored in Italy), some results have been obtained, an example is shown in Fig. 9 below.

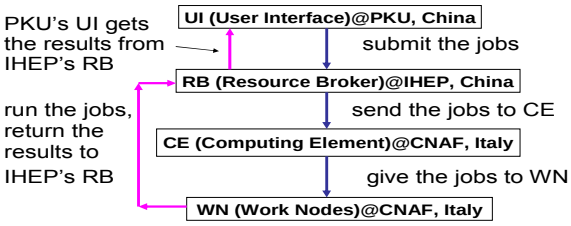


Fig. 8. The latest procedure via the IHEP LCG

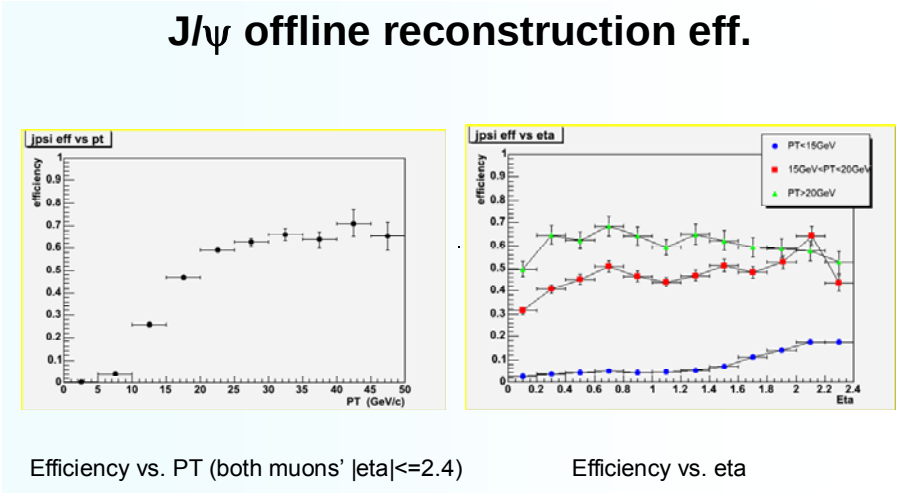


Fig. 9. The sample result from the physics analysis with the grid tool.

There results have been summarized into a CMS Analysis Notes [13] which has been approved by CMS at the end of 2006.

4.4. Ongoing work and an estimate of required resource

The next steps for us are to generate 1 million prompt J/ and 1 million prompt ϒ events, then to put them through the CMS full simulation and reconstruction software chain (CMSSW). We have estimated that,

- for each million events, it needs about 24,000 hours (or 1000 days) of CPU time (for one P4 Xeon 1.5GHz computer), and about 1.1 TB of storage space;
- in result, we would need ~2800 days (i.e. ~ 9 years) of CPU time and ~3.1 TB of storage space for such 2 million J/ψ and ϒ events plus 40% of background events

5. Future plan on grid computing in Peking University

5.1. Application of grid computing on biology research

The Peking University biology subgroup in EUChinaGRID is located in the Beijing Nuclear Magnetic Resonance (NMR) Center which is sponsored by Ministry of Science and Technology and Ministry of Education of Chinese government, also by Chinese Academy of Science and Chinese Academy of Military Medical Sciences. Beijing NMR Center is managed

by Peking University and is a national NMR facility established on Nov. 4th, 2002. The center is for research and training in bio-molecular NMR studies. We need to use computer for processing and analyzing NMR data, for solution structure calculation, and for molecular dynamic simulation.

The NMR Spectroscopy is a key method for obtaining high resolution structure in addition to X-ray structure. It is operated at the physiological temperature and condition which are closer to native functional state. The structure calculation is very time consuming for multiple structures and multiple rounds. Fig. 10 is the procedures for calculation of 3D structure of protein molecules. Fig. 11 is a sketch to show how the structures are formed from constrains.

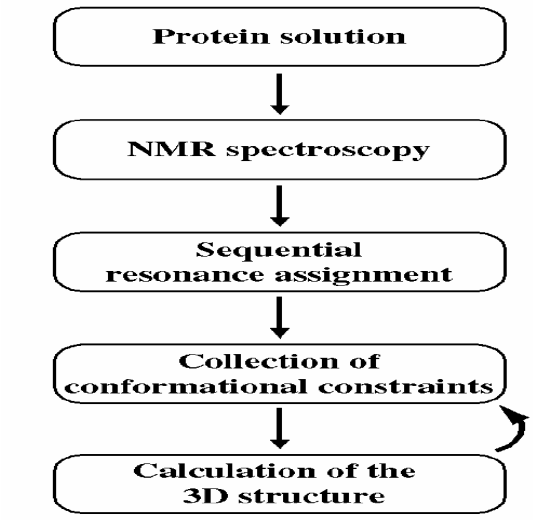


Fig. 10. NMR structure determination

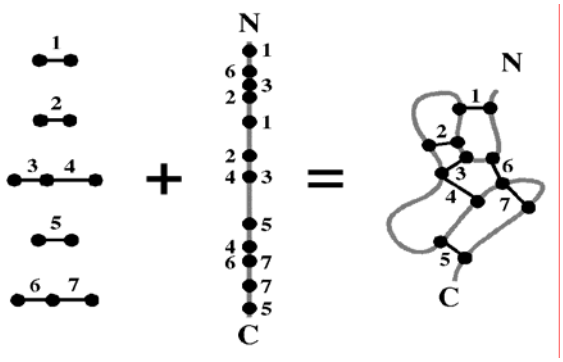


Fig. 11. Restrained molecular dynamics and simulated annealing

The structure calculation includes the energy minimization. The empirical energy (which is from experimental data) contains all information about the primary structure of the protein and also data about topology and bonds in proteins in general. Fig.12 is an example of structure calculation and refinement, each round of calculation involves many structures, normally 200 structures per round, and each protein may need 10-30 (or more) rounds of calculations. Some structures calculated recently are shown in Fig.13.

The analysis software for protein structures is “Amber” which is a commercial software and the licenses need to be granted on all computers involved. .University Rome III has procured the license and is testing it, hopefully it can be available for us to use in near future

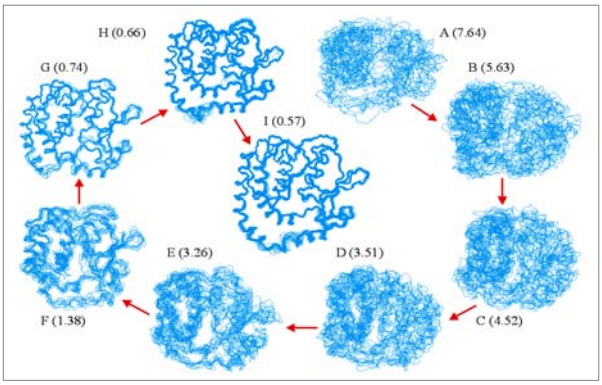


Fig. 12. Structure calculation and refinement

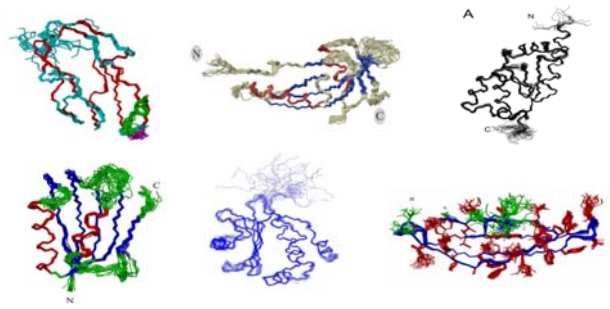


Fig. 13. Examples of recent structures being calculated

Similar as the PKU-Physics group, we also estimated the computing resources needed by PKU-Biology group:

- By using the Intel 2.4 GHz Xeon CPU
- Each structure needs 4 hours, each round to compute 200 structures
- Each protein needs to be computed for 10 rounds
- Totally if 10 proteins to be analyzed

→ ~ 80,000 hours (> 9 years) CPU time and > 1TB storage space

5.2. Interoperability between middleware GOS (of CNGrid) and gLite3 (of EGEE)

At present, the future possible standard of grid protocol OGSA is just a big frame, without much concrete content yet; users and designers also have many conflicts. Namely nowadays there is no any mature grid standard yet. On other hand, this would be an opportunity for the grid researcher to make contribution on standardization of grid computing. However, from a practical application point of view, currently different grid systems can not easily share the resource yet due to the different middleware, this is directly contrary to the purpose of the grid computing, i.e. resource sharing.

CNGrid (China National Grid) has been supported by Ministry of Science and Technology of Chinese government. Its objective is to build a Chinese national grid system and to promote the grid application. On another hand, the gLite middleware of EGEE is becoming more and more popular is physics, biology and other scientific applications, with the

increasingly demand on the computing and storage resources, while the CNGrid seems have some idling resources. Therefore, the one goal of EUChinaGRID project is to study the interoperability between two grid middleware (i.e. the GOS of CNGrid and the gLite of EGEE), wishing that jobs can be submitted each other between two Grid systems stably. Next part will give an overview of GOS middleware, and compare it with gLite.

The GOS middleware is divided into three levels: the lower one is the Device level; the middle one is the Bus level which can manage the resource information; the upper one is the VOE (VEGA Operation Environment) level, which provides the user support environment, including the basic API and the management client for grid batch jobs [14].

Generally, there are following evident differences between GOS and gLite systems.

A) Information system: GOS information services use resources relying on resource Routers and reveal resource organizations, as well as information retrieval. But the information system of gLite is implemented through globus MDS and RMMA packages, which conform to GLUE Schema and publish information with LDAP according to hierarchical structure
B) Security mechanism: gLite manages users with VOs and ensures the security of grid through CA certificates, proxies, SSL (Secure Shell). However, GOS grants users privileges and roles to access grid system with community rather than VOs.
C) Data management: GOS organizes data with grid catalog system, while gLite manages data complexly and stably.
D) Workload management system: Interoperability will focus on this part. gLite uses GRAM protocol to interact with LCG-CE, and choose Condor-G as GRAM client to submit batch job to LCG-CE. In contrast, GOS can simply implement jobs broker.

Interoperability is an important objective of EUChinaGrid project, a team in Beijing University of Aeronautics and Astronautics collaborating with a team in INFN/Catania has got some progress [15]. A special gateway has been designed with SEDA model and IoC model to change destination of job submitting; in gLite middleware, for example, jobs will selected to extend job management system rather than PBS queues. There are also some breakthroughs in data transfer. The simple transfer between two systems has been tested with the sandbox model.

The core of all the designs is the gateway, with which developers now can implement simple interoperability. However, there are still many problems to be solved (e.g. the large jobs still can not be submitted each other yet, the management of different security systems are badly needed, etc.) and more collaborators are welcome to take part in the task.

5.3. Grid portal and promotion of grid application

From gLite middleware installment and usage of UI we can easily notice that users have to face complex commands and the inconvenient operational interface, which the general users should not waste time to learn. If users can work conveniently with just buttons or intuitionist orders without complex operations, the grid computing technology will be earlier to be promoted. Therefore, the research on grid portal is significant for grid computing from this sense.

One of goals of the EGEE project is to construct a good development platform where users can design various application programs through some interfaces. With these interfaces, we could provide web operations which the user are familiar with, also through these interfaces we could implement authentication, submission of jobs and querying information, etc. For the clients, it will be more convenient to visit grid resources without considering the issues like differences in the operating system, etc. For the administrators, to manage and to test the grid system may be visualized by using these interfaces.

Grid portal generally consists of a three-tier structure that supports (1) the SSL client browser, (2) the Web Application Server (where the web application is running) and (3) the grid service layer which includes some services such as file transfer, job submitting and so on. With this network portal it is expected to providing secure access, user management, execution of operation, information publishing and monitoring, etc.

The GENIUS (Grid Enabled web eNvironment for site Independent User job Submission) developed by Italian INFN (Istituto Nazionale di Fisica Nucleare) is a typical Grid portal with rather rich functionalities. It is a web operational interface developed based on the kernel components and services of Globus' base layer, and it is very suitable to be operated by the non-professional grid users. The Supercomputing Center of Chinese Academy of Science also has some successful experience on the development of this kind of application program. However, along with the new problems emerged in the interoperability between different middleware, these existing portals face some new problems on the aspects of authorization and authentication, job submission and information inquiring etc. We wish to develop a more suitable grid portal with the solutions for the problems from the interoperability.

6. Conclusion

We have briefly introduced the concepts and great potentials of Grid computing, which will have attractive vast prospectives on the applications in biological and medical science, HEP, geo-science, astronomy and many other fields. Some middleware of various grid computing projects have entered the practical application stages, the gLite3 middleware explained in this paper is a typical one.

Peking University group has accumulated some experience on the grid computing in last few years, But much more work are needed to be done, for example,

- to start the biology application after the software license issue is solved;
- to gear up the readiness of HEP application for the huge amount of MC and real data to pour in when LHC to start operation in less than a year;
- to participate in the interoperability study for different grid middleware, etc.

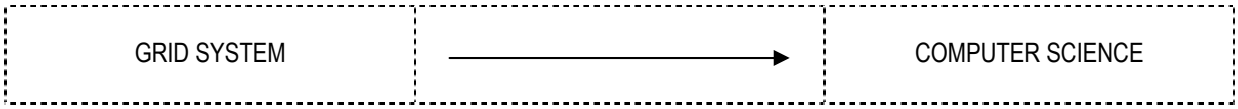
We strongly believe that, with the collaborative effort from all colleagues in the grid computing field, this promising new technology will be more mature and will produce more great application results which were unreachable in the past.

Acknowledgment

We are very grateful to EUChinaGRID project, the helps and supports from all partners are essential for our achievement in last two years. The construction of BEIJING-PKU site has been a collective effort from all members of PKU group; we particularly thank the contribution from Ms. K. Kang, Mr. L. Zhao, D. Mu, Z. Yang, S. Guo and L. Liu. We are indebted to Prof. B.Xia who provided all materials related to the biological study. Finally, we appreciate the great help from Polish colleagues in Jagiellonian University, Medical College – Cracow on publishing this article.

References

1. http://www-03.ibm.com/grid/about_grid/what_is.shtml
2. I. Foster, C. Kesselman (editors). The Grid: Blueprint for a New Computing Infrastructure, 2nd edition. Morgan Kaufmann (2004)
3. IBM RedBooks: Introduction to grid computing with globus http://www.redbooks.ibm.com/Redbooks
4. https://documents.euchinagrid.org/getfile.py?docid=50&name=EUChinaGrid-Del3.1v1.7&format=pdf&version=1
5. http://press.web.cern.ch/Press/PressReleases/Releases2003/PR13.03ELCG-1.html
6. http://euchina-gridice.cnaf.infn.it:50080/gridice/host/host_summary.php?siteName=BEIJING-PKU
7. http://lcg.web.cern.ch/lcg/overview.html
8. http://www.ncess.ac.uk/learning/start/faq/
9. http://www.euchinagrid.org/
10. http://www.nesc.ac.uk/talks/talks/RobAlan+SteveBoothPresentation/Globus_Part2_22-10-01.ppt
11. https://edms.cern.ch/file/722398//gLite-3-UserGuide.html
12. http://osg-docdb.opensciencegrid.org/0004/000458/001/gLite-Architecture-4Bob-OSG.ppt
13. Z. Yang, S. Qian, "/Psi → mu+ mu- reconstruction in CMS", CMS Analysis Notes 2006/094 (2006).
14. http://vega.ict.ac.cn/gos/gos11/vega_gos_manual.pdf
15. Yongjian WANG, State-of-the-art of Interoperability Work in EUChinaGrid Project, Beijing University of Aeronautics and Astronautics



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 17-21

GRID: FROM HEP TO E-INFRASTRUCTURES

FEDERICO RUGGIERI – INFN

Abstract: GRID technology has been applied to several scientific applications. High Energy Physics has been one of the earliest adopters of the GRID approach due to problematic treatment of the huge quantity of data that the Large Hadron Collider (LHC) at CERN will produce in the next years. GRID Infrastructures, initially set-up by those early users, are now deployed in large number of countries and Europe is one of the big investors in the field. Several scientific applications are now available on the GRID which is now recognised as one of the enabling e-Infrastructures technologies. Development of new e-Infrastructures, especially in new emerging countries, could be relevant as an acceleration factor for the growth of scientific communities in those countries.

Introduction

GRID is not an acronym and GRID technology is basically an evolution of concepts like meta-computing and distributed computing.
The GRID Bible is the famous book: "The GRID: Blueprint for a new computing infrastructure" [1] edited by Ian Foster and Carl Kesselman where the first (as far as I know) official definition of GRID can be found: `A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities'.
They also started the first GRID project, Globus [2], which developed the first "Middleware": the Globus Tool Kit.
Then the GRID was intended as:
a dependable infrastructure that can facilitate the usage of distributed resources by many groups of distributed persons or Virtual Organizations;
an extension of the WEB concept, which was originally limited to distributed access to distributed information and documents.
The classical example is the Power GRID where you plug in and receive power; you don't know (and you don't care) where it comes from.

Ian Foster in 2002 suggested [3] that GRID is a system that:

- *"coordinates resources that are not subject to centralized control … (A Grid integrates and coordinates resources and users that live within different control domains–for example, the user's desktop vs. central computing; different administrative units of the same company; or different companies; and addresses the issues of security, policy, payment, membership, and so forth that arise in these settings. Otherwise, we are dealing with a local management system.)"*
- *"… using standard, open, general-purpose protocols and interfaces… A Grid is built from multi-purpose protocols and interfaces that address such fundamental issues as authentication, authorization, resource discovery, and resource access … omissis… it is important that these protocols and interfaces be standard and open, otherwise, we are dealing with an application specific system."*
- *"… to deliver nontrivial qualities of service. A Grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, relating for example to response time, throughput, availability, and security, and/or co-allocation of multiple resource types to meet complex user demands, so that the utility of the combined system is significantly greater than that of the sum of its parts."*

This new and more extensive definition clarifies the main differences between a GRID and a cluster or a farm of computers.

My short history in Grids

In the 80' and early 90' the accent was on client-server and meta-computing; many computing centres were trying to overcome the limitations (and costs) of single mainframes using clusters of servers and workstations.
In 1998 I. Foster and C. Kesselman edited their famous book [1] and I knew about GRID by the first GRID presentation at CHEP'98 (Computing in High Energy Physics) conference in Chicago (USA).

My interest was also renovated by my colleague Giovanni Aloisio who came to Bologna to present in a seminar the possible use of the Globus Toolkit. It was 1999 and we started the INFN-GRID Project based on Globus, and in November of that year in the HEP-CCC Meeting at CERN there was a discussion with F. Gagliardi (CERN), Georges Metakides and Thierry Van der Pyl (senior officers from the EC IT programme), on our major computing challenges related to the data analysis of the experiments at the Large Hadron Collider (LHC) [4] and possible new initiatives. My suggestion to present a proposal to the European Commission (EC) based on GRID technology was favourably accepted by the European HEP Community and CERN agreed to lead it.

In 2000 the UK Particle Physics Grid (GridPP) [5] was started, and at the CHEP2000 conference in Padova (Italy) the ideas were already defined. The proposal was accepted by EC and in the same year started as the first European GRID Project: DataGRID [4].

At the same time our HEP colleagues in US proposed two GRID projects PPDG [6] and GriPhyN [7].
DataGRID lasted until 2003 and then a new Grid Infrastructure activity was approved by the EC: EGEE Project [8], while in US OSG (Open Science GRID) [9] was started.
Nowadays many other projects in many countries (Japan, China, etc.) have been started and GRID is now considered an enabling technology for the emerging e-Infrastructures.

GRID for LHC and HEP

We, as HEP community, got involved in Grids in the late nineties to solve the huge LHC computational problem which was starting to be investigated (after an initial under-evaluation).

At that time, client-server and meta-computing were the frontier and first implementations of Computer Farms were appearing (Beowulf [10]). The largest problem anyway was the huge amount of data expected to be produced and analyzed (tens of PetaBytes).

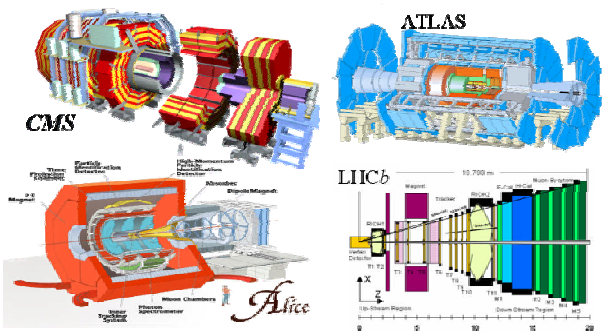


Fig. 1. The four Experiments of LHC

The "social" challenge was to allow thousands of physicists to access those data easily from tens of countries in different continents. It was also clear that even taking into account the Moore's Law for the Computing power evolution, the CERN budget traditionally dedicated to Computing resources was largely insufficient. There was no obvious solution on the market and such a worldwide enterprise requested new approaches.

Several "new" technologies were proposed like Object Oriented (OO) Programming and OO DataBases, and several Research and Development projects proposed to solve it.

HEP computing is, on the other hand, a typical High Throughput Computing that allows a very simple or "natural" parallelization based on the replica of the application program and the Event based data structure (Figure 2).

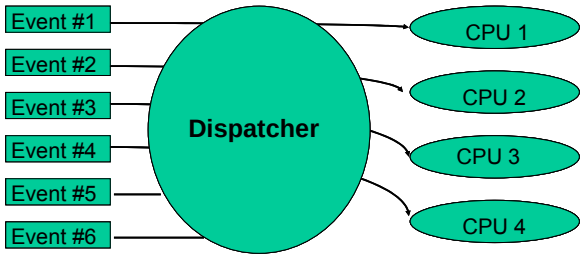


Fig. 2. Event based HEP High Throughput Computing

The basic approach proposed was to distribute the load of LHC computing in the various laboratories with CERN being the Data Source and the main repository of such data. The model, proposed by the MONARC Project [12], defined a few levels or "tiers", with CERN as a Tier0 and the other Regional Centres as Tier1 with Tier2 underneath (Figure 3). The GRID appeared as a natural answer to those requirements.

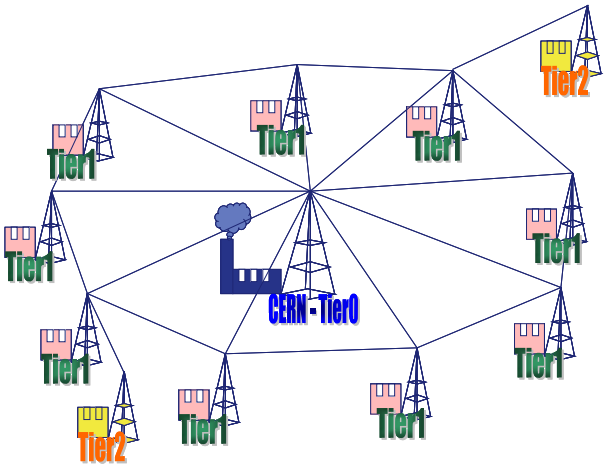


Fig. 3. Tier Structure of LHC Grid

GRID Architecture

DataGRID suggested a layered Architecture for the GRID and the four foreseen layers were half application related and half Grid Hardware and Software related.

The basic GRID services implemented in software are normally referred as middleware. The hardware and software configuration is still very similar to that first architectural view (Figure 4).

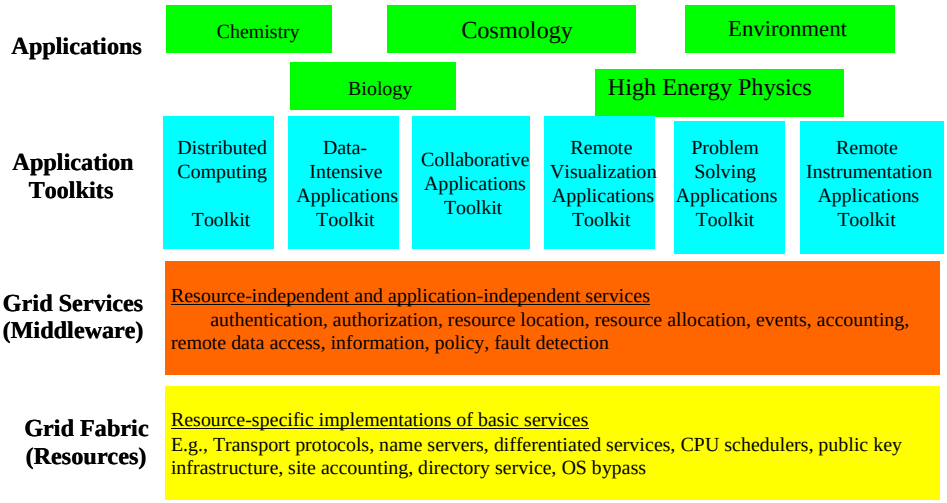


Fig. 4. EU-DataGRID Architecture (2001)

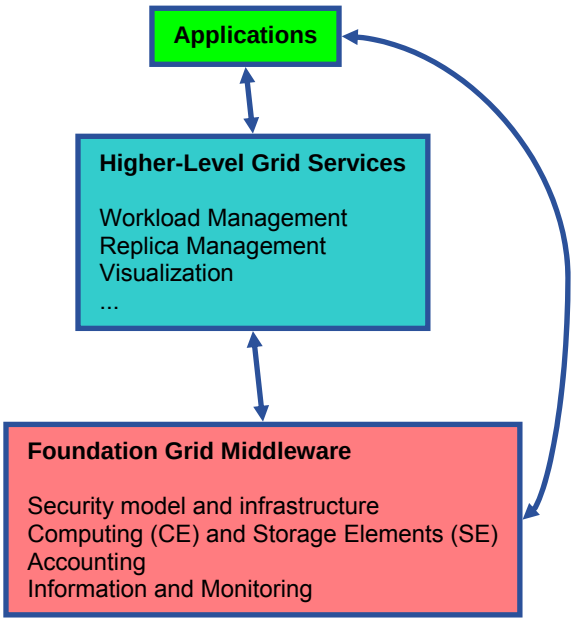


Fig. 5. EGEE Middleware Layered structure

Computing GRID basic components

A very simple description of the Computing GRID hardware building blocks can be schematically presented as follows:

- computing resource or Computing Element (CE);
- storage resource or Storage Element (SE).

Those components shown in Figure 5 will be described in the following paragraphs.

Computing Element

The Computing Element (CE) is the basic component of the Computing Resources, it essentially corresponds to a Batch Queue that processes the jobs submitted by the users. Behind a CE there can be tens or hundreds or even thousands of real computing machines or CPU (Central Processing Unit). Those servers are organized in a cluster or a farm of computers and the batch scheduler assigns to them the jobs to be executed, just like in a farm cows in a row produce milk or chickens make eggs.

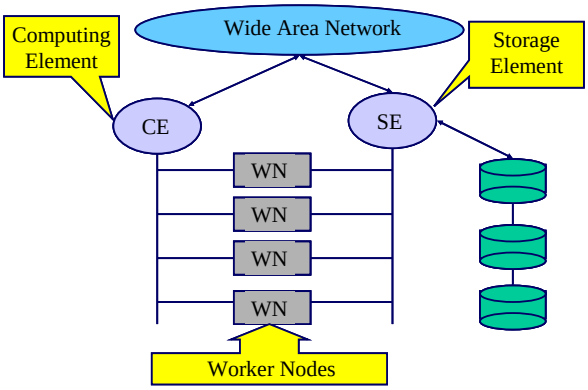


Fig. 6. A CE and a SE made by a Computer Farm and a set of Disks

Storage Element

The Storage Element (SE) is a system that allows the storage of data and programs in the Grid. The hardware architecture of such a storage is not relevant, provided that the service is accessible via Grid tools, like GridFTP [13], for data transfer, which allow storage and retrieval of data in the Grid.

Grid Services

The list of Grid services developed on top of the basic components is quite long. The main services currently used are the following:

- Workload Management System (Resource Broker) [14] chooses the best resources matching the user requirements.
- Virtual Organization Management System (VOMS) [15] allows to map User Certificates with Virtual Organizations (VO) [16] describing rights and roles of the users.

- Data Oriented Services: Data & Meta-data Catalogs, Data Mover, Replica Manager, etc.
- Information & Monitoring Services which allow to know which resources and services are available and where: GridICE [17].
- Accounting services to extract resource usage level related to users or group of users and VOs.

Social impact of Grid Infrastructures

The Grids are considered as part of a more general specie called e-Infrastructures which include also communication networks. They support wide geographically distributed communities and then enhance international collaboration of scientists. The deployment and usage of such resources is also promoting collaboration in other fields where we speak of e-Business, e-Government and industrial take up.

One of the ways the Research and Education Grids and networks make an impact on the society is that they allow the access of many researchers to scientific resources, laboratories and data, distributed around the world. Researchers from developing countries will have less need to travel and leave their home countries to participate into big science and frontier scientific activities and then the so called brain drain can be reduced.

Another important aspect is that the e-Infrastructures promote the usage of network connectivity, computing resources and open source software stimulating not only the scientific activity, but also the technical development of communities in the countries contributing to fight the digital divide.

Grid Infrastructures around the world

A large number of projects around the world are currently deploying Grid infrastructures or have already reached production quality level. Large Grid Infrastructures are already used in China (CNGrid [18] , ChinaGrid [19]), Europe (EGEE [9]), Japan (NAREGI) and United States (OSG, Teragrid) and many National Grid Initiatives (NGI) were created to support Grid Infrastructures at national level.

The European Commission has largely invested in Grids trough the projects funded in the past Framework Programs (FP5 and FP6) and is currently planning to invest even more in FP7 (2007-2013).

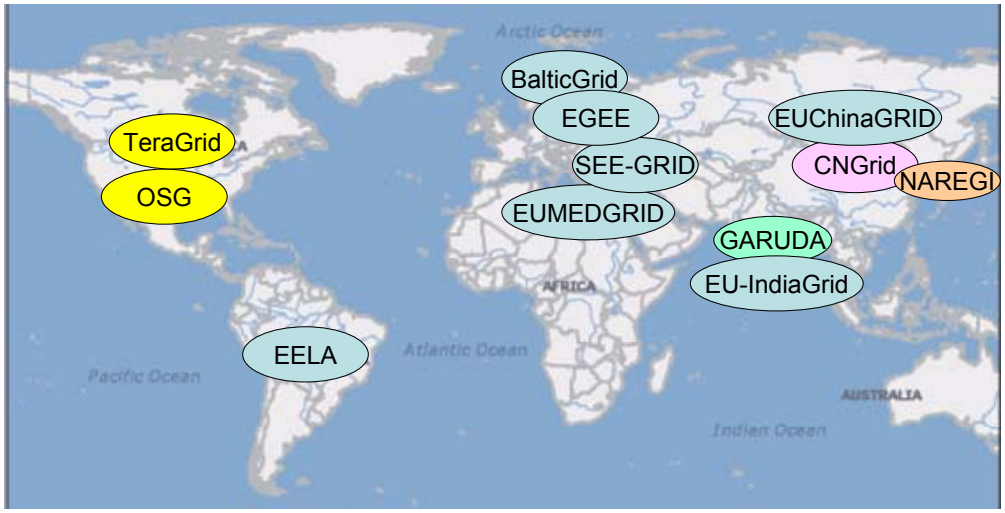


Fig. 7. Grid infrastructures around the world

Conclusions

As it was discussed in the previous paragraphs, Grids are part of the concept of e-Infrastructures, together with communication networks they provide the necessary layers of communication and collaboration tools needed by modern scientists.

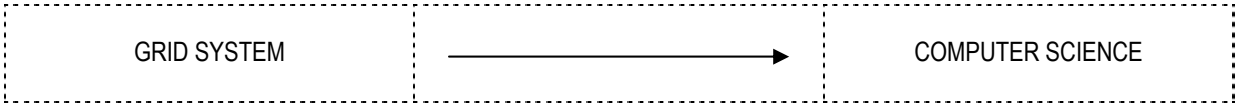
Grids can not only optimize the usage of resources, but increase their usability and accessibility being a valid instrument for cooperation in Science and Education fostering the creation of a Human Network among scientists and researchers.

e-Infrastructures are fundamental for long term development and can play a role to mitigate phenomena like: the Digital Divide and the Brain Drain.

References

1. "The GRID: Blueprint for a new computing infrastructure" edited by Ian Foster and Carl Kesselman, Morgan Kaufman 1998.
2. Globus Project: http://www.globus.org/
3. What is the Grid? A Three Point Checklist. I. Foster, GRIDToday, July 20, 2002.
4. Large Hadron Collider: http://lhc.web.cern.ch/lhc/
5. EU DataGRID Project: http://eu-datagrid.web.cern.ch/eu-datagrid/
6. GridPP project: http://www.gridpp.ac.uk/
7. Particle Physics Data Grid: http://www.ppdg.net/
8. GRIPHYN: http://www.griphyn.org/
9. Enabling Grids for E-sciencE: http://www.eu-egee.org/
10. Open Science Grid: http://www.opensciencegrid.org/
11. Beowulf Project: http://www.beowulf.org/overview/index.html
12. MONARC Project: http://monarc.web.cern.ch/MONARC/
13. GridFtp: http://www.globus.org/grid_software/data/gridftp.php

14. WMS: http://egee-jra1-wm.mi.infn.it/egee-jra1-wm/wms.shtml
15. VOMS: http://infnforge.cnaf.infn.it/voms/
16. Virtual Organization: http://en.wikipedia.org/wiki/Virtual_organization
17. GridIce: http://gridice.forge.cnaf.infn.it/
18. CNGrid: http://www.cngrid.org/en_introduce.htm
19. ChinaGrid: www.chinagrid.edu.cn/
20. NAREGI: http://www.naregi.org/index_e.html
21. OSG: http://www.opensciencegrid.org/
22. Teragrid: http://www.teragrid.org/
23. FP7 – Cordis Web site: http://cordis.europa.eu/fp7/home_en.html



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 23-25

GRID INFRASTRUCTURES AS CATALYSTS FOR DEVELOPMENT ON ESCIENCE: EXPERIENCES IN THE MEDITERRANEAN

GIUSEPPE ANDRONICO[*], ROBERTO BARBERA[**], KOSTAS KOUMANTAROS[***], FEDERICO RUGGIERI[****], FEDERICA TANLONGO[*****], KEVIN VELLA[******]

[*]*INFN Sezione di Catania, Via S. Sofia, Catania, I-95123, Italy; giuseppe.andronico@ct.infn.it*
[**]*University of Catania and INFN Sezione di Catania, Via S. Sofia, Catania, I-95123, Italy; roberto.barbera@ct.infn.it*
[***]*GRNET, Mesogion Avenue 56, Athens, 11527, Greece; kkoum@grnet.gr*
[****]*INFN Sezione di Roma Tre, Via della Vasca Navale 84, Roma, I-00146, Italy; federico.ruggieri@roma3.infn.it*
[*****]*GARR, Via dei Tizii 6, Roma, I-00185, Italy; federica.tanlongo@garr.it*
[******]*University of Malta, Msida Campus, Msida, MSD06, Malta; kevin.vella@um.edu.mt*

Abstract: The digital gap prevents today in many parts of the world the diffusion of e-Science which is considered as one of the key enablers of progress and development in the 21st Century. On the other hand, investing in e-Infrastructures is the key for a long-term growth and changes in the Society in developing Countries. The paper discusses this topic and provides some details about the EUMEDGRID Project experience in the Mediterranean area.

Keywords: Digital Divide, e-Infrastructure, e-Science, Grids, Information Technology, Mediterranean

In the last few years, the scenario of international collaboration in Research and beyond has swiftly evolved with the gradual but impressive deployment of large bandwidth networks. A number of advanced services and applications have been using these networks, enabling new ways of remote collaboration. The environment resulting from the integration of networking and other resources, such as computing, storage, instruments and related systems is also known as e-Infrastructure. In the most advanced economies, knowledge is nowadays one of the major elements of progress and economic welfare and e-Infrastructures are, in turn, one of the major enablers of development in a knowledge economy.

On the other hand, this menaces to widen the digital gap between developing economies and the most advanced ones, where knowledge is a commodity and an important share of the budget of companies and governments is allocated on R&D and on Education: the latter gets, as a return of their conspicuous investments, more and more advanced infrastructures and techniques that enable in turn new developments, while the former, taken off late and with less resources and urged by more fundamental needs, seem incapable to reduce the gap.

At a first glance, to invest the limited budget of a developing country in building e-Infrastructures could seem unnatural, foolish as they have much more basic and compelling needs. Nevertheless, it is important to understand the role of e-Infrastructures in breaking this loop. In a saying: "if you give a fish to a hungry man you feed him for a while, but if teach him how to fish, you feed him for a life."

Although needs such as food, water, medical services are fundamental in the short term, a long-term solution cannot build just upon them: other activities are necessary to create favourable conditions for a sustainable growth. Agriculture and industry developments are needed to produce food and employment depending on the specific local situation, start social innovation and improve the quality of life, and science is at the basis of long-term innovation in both of them. Digital infrastructures are necessary to allow researches to participate to frontier scientific activities, to share competences and experiences with their counterparts all around the world, thus being up with the most recent tools and methods.

This kind of investment should be therefore understood and evaluated on several (tens of) years and should have a "figure of merit" with respect to the obtained results and the sustainability of future activities.

One of the most significant news in the outline of global e-Infrastructures is the so-called "grid paradigm", a revolutionary distributed environment for sharing computing and storage resources, allowing new methods of global collaborative research - often referred to as e-Science. This new paradigm, although still under development, is foreseen to have a large

impact well beyond the field of mere research: the national and international initiatives developed to date are making the "World Wide Grid" and its applications one of the major global R&D topics of the century.

Grids are a set of services over the Internet, allowing geographycally dispersed users to share computer power, data storage capacity and remote instrumentation. The basic concept of this new technology as well as its revolutionary potential are in the very world “grid”, usually meaning the electric distribution system in English: electric power is indeed distributed to final users who are not aware how and where it was produced, nor they need to use it: with grid computing, it is just the same for remote resources.

Grid computing is in fact a particular example of distributed computing based on the idea to share resources on a global scale. Several elements are needed for a grid infrastructure to work:

- An Authentication and Authorization system, providing secure access to resources, to guarantee data privacy and integrity (a critical factor in several application fields such as biomedicine);
- A mechanism (the so-called middleware) able to manage and allocate resources in an optimal way to all users and applications who need them, just like the Operative System does with programs running on your PC;
- A reliable, high-performance network connection amongst resources, ensuring that the time taken for data transfer is negligible in comparison with the benefit of quicker processing obtained thanks to distributed computing.

First Grids were developed in the framework of the so-called e-Science, an innovative approach to research, thanks to the use of advanced technologies of communication and regardless to geographical location of instruments, resources and last but not least, brains.

The expectation that Grids will become very soon a commodity service, thus producing deep changes not only in Science, but industry and the Society at large, is a common belief amongst ICT experts. Accordingly, the European Commission, several national programmes and large private companies are investing in R&D projects since 2001, thus funding the creation of pilot Grid implementations and collaborative models for the usage of computing and data resources across technological, administrative and national domains.

Although experts believe that, within the next two decades, Grids will have an impact comparable to that of the WWW, at the present time (further analogy with the WWW) the development of Grids is, for the most part, in the hands of the worldwide scientific community. Scientists are exploiting the new technology to solve ever-more-difficult computational and data management problems across a wide range of domains.

OECD (Organisation for Economic Co-operation and Development) recognized the importance of Grids since 2004, when the GSF approved a proposal to convene a workshop on Grids and Basic Research Programmes. This workshop, held in Sidney on 25-27 September 2005, highlighted the potential benefits for developing countries:

> *“Grids can provide access to vast scientific and computing resources with only a modest investment in a local infrastructure (a minimal useful installation would consist of an Internet-linked high-performance workstation). The potential benefits to developing countries are considerable, since scientists would be able to join international collaborations based on their potential intellectual contributions alone. Thus, for example, it is already foreseen that elementary particle physicists in developing countries will be able to fully participate in the operation and exploitation of the Large Hadron Collider experiments at CERN (which is scheduled to begin operations in 2007). After it is completed in 2007, the LHC will generate 15 petabytes of data per annum 8, servicing 5000 research scientists in 500 research organisations or universities around the world. Such global-scale collaboration among researchers will be enabled by the Grid. Similar collaboration opportunities are emerging in other data-intensive domains such as astronomy, bioinformatics, the earth sciences and the social sciences.”*

One of the recommendations from the workshop focused on facilitating the access to this technology to scientists from emerging countries:

> *“Consideration should be given to the creation of new mechanisms (or the strengthening of existing ones) to facilitate access to Grids for researchers and research organisations in developing countries, plus other appropriate measures to broaden international participation in Grid projects. Telecommunications policies and regulations could be reviewed and, if appropriate, modified to facilitate access to high-speed computer networks in developing countries.”*

In line with this vision, in the context of the last EU Framework Programme for Research and Technological development (FP6), several projects aiming to extend the European flagship Grid infrastructure EGEE [1], outside the boundaries of EU were funded and launched, such as SEEGRID (addressing South-East Europe) [2], EELA (Latin America) [3], BALTIC-GRID (Baltic Region) [4] and EUMEDGRID (Mediterranean countries) [5], while others focused on interoperating such infrastructures with the ones existing in other regions of the world, such as EUChinaGRID (addressing the interoperability of Grids between Europe and China) and EU-IndiaGrid (addressing the same issue between Europe and India).

These experiences have proved to be very useful in order to speed-up the process of adoption of this new technology within the scientific communities of the beneficiary Regions. But from experiences such as the SEE-GRID and EUMEDGRID ones, we are learning that there is something more, and perhaps more important, than the mere adoption of a new technology developed by someone else. Indeed, it appears that the grid paradigm is especially useful for those countries that have scarce, often scattered in a wide territory, IT resources at their disposal. The implementation and

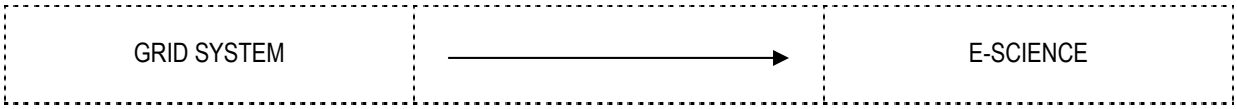
coordination of a grid infrastructure at a national (or larger) level can be regarded, especially in developing countries, as an opportunity to optimize the usage of existing, limited storage and computing resources and to enhance their accessibility for all research groups.

Many research fields have indeed very demanding needs in term of computing power and storage capacity, which normally are provided by large computing systems or supercomputing centres. Furthermore, sophisticated instruments may be needed to perform specific studies. Such resources pose different problems to developing economies: they are expensive, they need to be geographically situated in a specific place and – this is the case especially for those countries where the larger part of researcher is forced to emigrate in richer countries to continue their work in research – they could not reach a critical mass of users, because, for example, they are very specific and interest only a small community of researchers, or small communities scattered across the country. Thanks to the creation of a virtual distributed environment, all these drawbacks can be overcome. Through an appropriate access policy, different user groups can use resources wherever disperse according to their availability. Furthermore, geographically dispersed communities working at the same problem can collaborate in real time on the same study or experiment, thus optimizing not only hardware and software resources, but also human effort and “brains”.

This paper aims to present the experience of the EUMEDGRID project and the achievements reached during the first year of activity towards bringing the Mediterranean Countries to adopt the Grid paradigm for the benefits of their

References

1. http://www.eu-egee.org
2. http://www.see-grid.org
3. htp://www.eu-eela.org
4. http://www.balticgrid.org
5. http://www.eumedgrid.eu



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 27-31

RANDOMBLAST A TOOL TO GENERATE RANDOM "NEVER BORN PROTEIN" SEQUENCES

GIUSEPPE EVANGELISTA, GIOVANNI MINERVINI, PIER LUIGI LUISI, FABIO POLTICELLI*

*Department of Biology, University Roma Tre, 00146 Rome, Italy;** polticel@uniroma3.it

Running title: RANDOM NBP SEQUENCES GENERATION

Abstract: In an accompanying paper by Minervini et al., we deal with the scientific problem of studying the sequence to structure relationships in "never born proteins" (NBPs), *i.e.* protein sequences which have never been observed in nature. The study of the structural and functional properties of "never born proteins" requires the generation of a large library of protein sequences characterized by the absence of any significant similarity with all the known protein sequences. In this paper we describe the implementation of a simple command-line software utility used to generate random amino acid sequences and to filter them against the NCBI non redundant protein database, using as a threshold the value of the Evalue parameter returned by the well known sequence comparison software Blast. This utility, named RandomBlast, has been written using C programming language for Windows operating systems. The structural implications of NBPs random amino acid composition are discussed as compared to natural proteins of comparable length.

Introduction

The number of proteins that can be obtained combining the 20 natural amino acids is astronomically large (100^{20} for proteins just 100 residues long) and thus natural proteins represent only an infinitesimal fraction of the protein sequences space. From this simple consideration arises the concept of "never born proteins" (NBPs), *i.e*. protein sequences which have never been exploited by nature [1]. In the accompanying paper by Minervini et al., we describe a computational approach undertaken for the study of the sequence/structure relationships in NBPs using a grid implementation of the well known Rosetta protein structure prediction software [2]. The final aim of this study is that of answering the question if natural protein sequences were selected during molecular evolution because they have unique physico-chemical properties or else they just represent a contingent subset of all the possible proteins with a stable and well defined fold [1]. If the latter hypothesis was true, this would mean that the protein realm could be exploited to search for novel folds and functions of potential biotechnological and/or biomedical interest. To be able to approach this problem and to obtain statistically significant results it is essential to analyse a large library of protein sequences (at least 10^5 to 10^7) which do not display any significant homology with natural proteins. In other words, it is necessary to sample the protein sequences space in different points far away enough from the ensemble of natural proteins. In this context, a reasonable approach is that of generating random amino acid sequences and to compare them to the known natural proteins in order to eliminate from the sample under study those protein sequences which display statistically significant similarity with natural proteins. A collection of all the known natural protein sequences is represented by the National Center for Biotechnological Information non redundant protein sequence database [3], hereafter named NR database. Among the available tools to determine if a given query amino acid sequence displays significant similarity with any known natural protein, the Blast software [4], which stands for Basic Local Alignment Search Tool, is one of the most used in the computational biology community. Blast finds regions of local similarity between sequences and can be used to compare nucleotide or protein sequences to sequence databases, calculating the statistical significance of matches [5]. A parameter used by Blast to evaluate the statistical significance of a match is the Expect value (or *Evalue*). This is a parameter that describes the number of hits one can "expect" to find by chance for a database of a given size. The *Evalue* is related to the score S that is assigned to a match between two sequences, to the lengths *m* and *n* of the sequences and to the parameters *K* and λ (natural scales for the search space size and the scoring system respectively [5]) by the equation:

$$Evalue = Kmn\, e^{-\lambda S}$$

The *Evalue* parameter can be used as a threshold to distinguish significant from non-significant matches. If the statistical significance of a match is greater than the *Evalue* threshold, the match is not considered significant, or the query sequence is not considered to display significant similarity to any protein present in the database.

In this paper we describe the implementation of a software utility used to generate NBP sequences, or in other words, random amino acid sequences with no significant similarity with known natural proteins present in the NR database, as

evaluated by the *Evalue* parameter returned by Blast. Average amino acid composition of a restricted NBP database (2×10^4 sequences) is analysed in comparison to that of natural proteins and discussed in terms of its possible influence on the structural properties of NBPs.

Results

Software description

RandomBlast consists of two main modules: a pseudo random sequence generation module and a Blast software interface module. A high level description of RandomBlast workflow is shown in figure 1 using an activity diagram. The first module uses the Mersenne Twister 1973 pseudo-random number generation algorithm [6] to generate pseudo-random numbers between 0 and 19. A free implementation of this algorithm, available in C programming language from Matsumoto and Nishimura [7], was used in RandomBlast. Random numbers are then translated in single character amino acid code using the conversion table shown in Table 1. Single amino acids are then concatenated to reach the sequence length specified by the user in the input parameters.

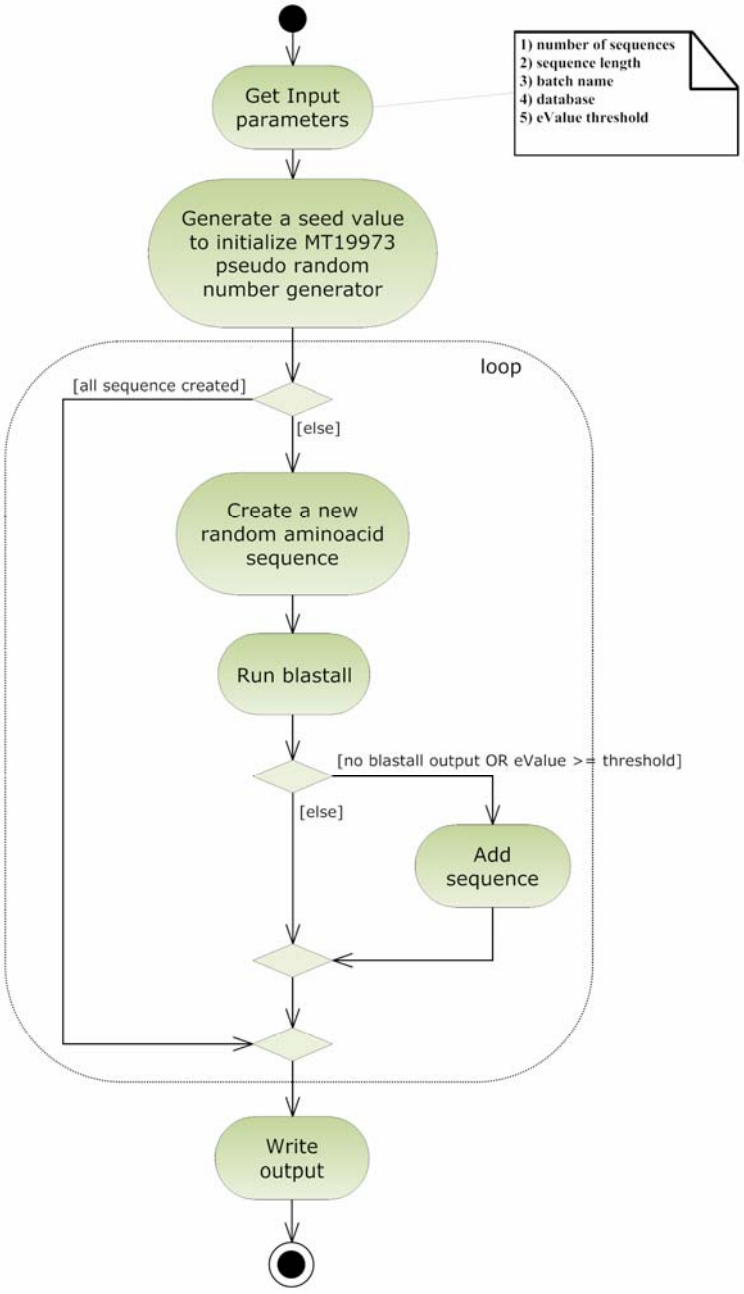


Fig. 1. Activity diagram showing the RandomBlast workflow. The inset details the RandomBlast input parameters.

Table 1. RandomBlast random number to amino acid type conversion table

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	A	V	L	I	C	M	F	W	P	S	T	Y	N	Q	D	E	K	R	H

Each generated sequence is then given in input to the second RandomBlast module, an interface to the Blast blastall program which invokes the following command:

blastall -m 8 -p blastp -d database -b 1;

where database in our case stands for the NR database, and the parameters –m 8 and –b 1 indicate the alignment format (tabular form) and the number of sequences to be returned (just the first hit), respectively. Blastall output is then retrieved by RandomBlast and the *Evalue* extracted from it. If the *Evalue* is greater than or equals the threshold chosen by the user, the sequence is valid and is added to the output log file. Note that in our case we regard as valid only the sequences that do not display significant similarity to any protein sequence present in the database, so that, contrary to the normal Blast usage, valid sequences are those displaying an *Evalue* higher than the threshold. When the number of valid sequences is equal to the number of sequences to be generated, specified by the user, the program execution is terminated and a log file, containing all the information about input parameters and some information concerning the sequences, is created. An example of this output log file is shown in figure 2.

```
# Starting evaluation...
SEED_VALUE=1180003591
THRESHOLD=1.100000
DB=ecoli.aa
SEQ_SIZE=70
SEQ_NUMBER=20
BATCH_NAME=batch02
SEQUENCE=WRCQYCGRIHCDQPFMCYEDLGFTFWAALEAKACWQAAPHDYQCVTAFSEGEQKSAEVKMQISRCDATKF
EVALUE=1.700000
SEQ_CODE=batch02_1
OUTPUT=tmpseq_0     gi|1789055|gb|AAC75745.1|       33.33   27      18      0       42      68      48      74      1.7     24.3

SEQUENCE=WKCPDPRSWESKTWAQCVVNCQHGSHQCQASIGQTFFAQFENAEFCNRCPHWYCMDHCEVKMVGRMQHYD
EVALUE=2.200000
SEQ_CODE=batch02_2
OUTPUT=tmpseq_0     gi|1789221|gb|AAC75896.1|       34.78   23      15      0       36      58      81      103     2.2     23.9

SEQUENCE=GMTWPDRAVSCPMECAFFFVCNAWHTDFAMGFAYTYLMDWHQILSANREDPSVDRHPMMRMCDCERGRGN
EVALUE=2.900000
SEQ_CODE=batch02_3
OUTPUT=tmpseq_0     gi|1788611|gb|AAC75335.1|       23.53   51      34      1       1       51      43      88      2.9     23.5

SEQUENCE=WWAYPITREIAPLPHSIEGIRKFERLPLVTQFLLQPRILEESLVGKPHLVLDEPHAGRIEIHCSNTVHGM
EVALUE=2.900000
SEQ_CODE=batch02_4
OUTPUT=tmpseq_0     gi|1786589|gb|AAC73493.1|       43.48   23      13      0       24      46      17      39      2.9     23.5

SEQUENCE=EETLCGCPHWAGTNPEQDAGCIYFNQILRCVCDTGHLLACVFEMTKDIIIAQPYIQYYAILTFEMVYQSL
EVALUE=-2.000000
SEQ_CODE=batch02_5
OUTPUT=

SEQUENCE=YPQTNPNEERALYHSGNKGITNWYESISDTHYHYCARQTWVYEPTLIMEQRPILDVRAGQMHAMATSMKK
EVALUE=1.700000
SEQ_CODE=batch02_6
OUTPUT=tmpseq_0     gi|1788934|gb|AAC75633.1|       25.00   76      44      3       2       66      65      138     1.7     24.3
```

Fig. 2. RandomBlast sample output log file

The program is invoked using the following command:

randomBlast <numberOfSequences> <sequenceSize> <batchName> <dbName> <threshold>;

so that the user can decide the total number of sequences that will form a single batch, the size of each sequence (in our case 70 amino acids), the name of the batch (that will unequivocally identify the sequences), the name of the database against which to execute the Blast search (in our case the NR database) and the Evalue threshold (as already mentioned, 1 in our case).

The RandomBlast utility has been written in C programming language and it's available, upon request to the authors, for Windows operating systems.

Analysis of a restricted NBP database generated using RandomBlast

Comparison between the average amino acid composition of natural proteins (NPs) and that of a restricted database of NBPs generated using RandomBlast (2×10^4 amino acid sequences) reveals several interesting differences (Figure 3). In fact, as expected for random sequences, in NBPs all the twenty amino acids are almost equally represented. On the contrary, in NPs some amino acids classes are largely overepresented (Table 2). In NPs aliphatic amino acids account for almost 42% of the total (as compared to 30% in NBPs), while aromatic amino acids make up just 8% of the total (as compared to 15% in NBPs). This can have important implications for the ability of NBPs to fold in a stable and well defined three-dimensional structure. For instance, the nearly 10% relative abundance of Leu in NPs as opposed to the 1% abundance of Trp (Figure 3) can be connected to the ability of branched aliphatic sidechains to easily "adapt" within a protein hydrophobic core as compared to the bulky and rigid Trp aromatic sidechain. Along the same line of considerations, the 1.5% abundance of Cys residues in NPs is likely connected to the high reactivity of this amino acid which can result in structure stabilization by disulphides formation but also in uncorrect Cys pairing and misfolding.

Table 2. Percentage amino acid composition of NPs and NBPs by amino acid classes*

	NPs	NBPs
Hydrophobic	49.99	44.93
Aliphatic	41.90	29.94
Aromatic	8.09	14.99
Polar	24.24	29.96
Basic	13.64	15.01
Acid	12.01	9.99

* Cys and Met have been included in the amino acids class "polar" for simplicity

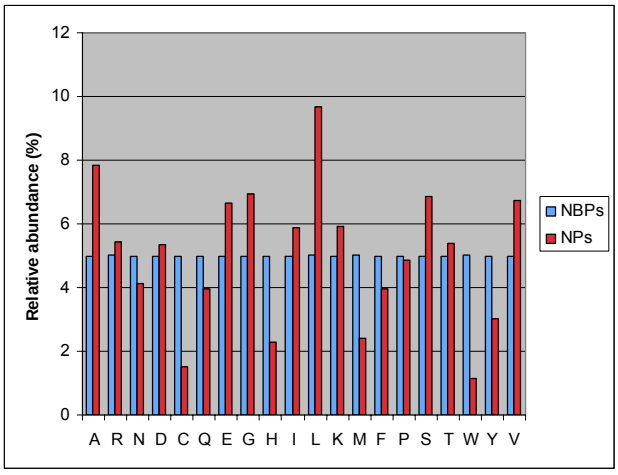


Fig. 3. Average amino acid composition of NPs (SwissProt) database [8] and of a restricted NBPs database. Amino acid composition has been calculated using the perl script freqaa.pl [9], available at the URL: http://www-alt.pasteur.fr/~tekaia/HYG/scripts.html.

On the other hand, the amino acid composition of an individual protein can be quite different from the average amino acid composition of an entire database, also taking into account the relatively short length of NBPs (70 amino acids). To dissect this aspect we also compared the amino acid composition of four randomly chosen NBPs with that of four natural proteins of the same length (Table 3). For each of the four NPs analysed, the same considerations made for the complete database remain valid. In fact, for three of them the relative abundance of Trp is 0%, while for one is under 1,5%, the total aromatic amino acids relative abundance being under 10% in all four cases. Also the Leu, and more generally the aliphatic amino acids relative abundance is in line with that observed for the complete database (Table 3), reinforcing the idea that a high proportion of aliphatic residues may be an intrinsic property of proteins which display a stable fold. Regarding the four NBPs analysed the observed amino acid composition does not dramatically deviate from the average database composition. In particular the aliphatic/aromatic amino acids ratio is significantly higher than that observed for NPs (Table 3). It is suggestive to speculate that this could be one of the physico-chemical factors that guided molecular evolution and shaped the ensemble of NPs. However, the statistical significance and thus the relevance of these considerations for protein structure studies will be assessed only once the structural characteristics will be analysed for a large library of NBPs, which will be the object of our future studies.

Table 3. Percentage amino acid composition of selected NPs and NBPs*

	NPs				NBPs			
	Taut	Trasp	L35	Peps	1	3000	6000	9000
A	8.57	7.14	7.14	7.14	5.71	4.28	5.71	1.42
C	1.42	2.85	0.00	1.42	4.28	1.42	5.71	1.42
D	1.42	2.85	1.42	7.14	4.28	2.85	10.00	7.14
E	12.8	4.28	1.42	5.71	5.71	7.14	4.28	1.42
F	1.42	1.42	4.28	2.85	1.42	5.71	12.8	4.28
G	8.57	10.00	7.14	7.14	8.57	10.00	7.14	4.28
H	4.28	4.28	7.14	2.85	8.57	5.71	1.42	2.85
I	2.85	4.28	2.85	7.14	8.57	1.42	1.42	4.28
K	1.42	2.85	10.00	7.14	5.71	7.14	1.42	8.57
L	11.40	7.14	8.57	7.14	2.85	7.14	8.57	11.4
M	4.28	2.85	5.71	8.57	1.42	2.85	7.14	2.85
N	0.00	1.42	0.00	4.28	2.85	2.85	2.85	5.71
P	2.85	7.14	2.85	4.28	7.14	2.85	2.85	2.85
Q	2.85	0.00	2.85	4.28	7.14	4.28	2.85	4.28
R	12.80	25.70	21.4	2.85	2.85	5.71	2.85	8.57
S	4.28	7.14	5.71	1.42	2.85	5.71	8.57	5.71
T	7.14	1.42	5.71	2.85	5.71	5.71	1.42	0.00
V	8.57	4.28	4.28	10.00	2.85	7.14	2.85	8.57
W	1.42	0.00	0.00	0.00	5.71	7.14	5.71	2.85
Y	1.42	2.85	1.42	5.71	5.71	2.85	4.28	11.4

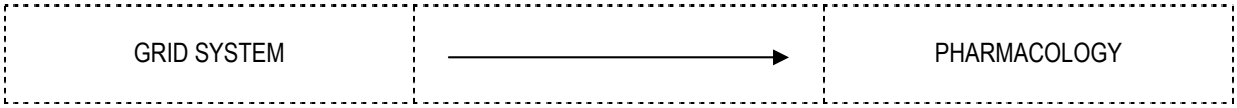
*The Abbreviations and NCBI gi codes for the NPs are the following: Taut, 4-oxalocrotonate tautomerase, gi:148568806; Trasp, Trasposase, gi:148521558; L35, ribosomal protein L35, gi:148567297; Peps, pyrrolidone-carboxylate peptidase, gi:147930188.

Acknowledgements

This work has been supported by a European Commission grant to the project "EUChinaGrid: Interconnection and Interoperability of grids between Europe and China" (contract number: 026634).

References

1. Chiarabelli C., Vrijbloed J.W., De Lucrezia D., Thomas R.M., Stano P., Polticelli F., Ottone T., Papa E., Luisi P.L.: Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display, Chem. Biodivers., 3, 840-859, 2006
2. Rohl C.A., Strauss C.E., Misura K.M., Baker D.: Protein structure prediction using Rosetta, Methods Enzymol., 383, 66-93, 2004
3. Wheeler D.L., Barrett T., Benson D.A., Bryant, S.H., Canese K., Church D.M., et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 33, D39-D45, 2005
4. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J. Mol. Biol., 215, 403-410, 1990
5. Karlin S., Altschul S.F., Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl. Acad. Sci. USA 87, 2264-2268, 1990
6. Matsumoto M., Nishimura T., Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Transactions on Modeling and Computer Simulation, 8, 3-30, 1998
7. Source code for MT19937 available at the URL: http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html
8. Bairoch A., Boeckmann B., Ferro S., Gasteiger E.: Swiss-Prot: Juggling between evolution and stability, Brief. Bioinform. 5, 39-55, 2004
9. Tekaia F., Yeramian E., Dujon B., Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis, Gene, 297, 51-60, 2002



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 33-37

A SOLUTION FOR DATA TRANSFER AND PROCESSING USING A GRID APPROACH

A. BUDANO[*], P. CELIO[**], S. CELLINI[*], R. GARGANA[*], F. GALEAZZI[*], C. STANESCU[*], F. RUGGIERI[*], Y.Q. GUO[***], L. WANG[***], X.M. ZHANG[***]

[*]*INFN Roma Tre, Roma, Italy.*
[**]*Dipartimento di Fisica, Università Roma Tre and INFN Roma Tre, Roma, Italy.*
[***]*IHEP, Beijing, China.*

Abstract: An important aspect in a lot of physics and biology experiments is the access and the processing of the data from a different places and in the shortest possible time. We implemented a data files moving system, based on GRID tools and services, to automatically transfer the files from a site that generated the data to some other site in the same collaboration group for processing and analysis. We also describe the GRID approach to a unified job submission and processing system and the mirroring of data files using the catalogues. This approach allows GRID communities to cooperate more efficiently in data analysis, to share the available resources and to backup the data at the same time.

Introduction

As part of our activity in the EUChinaGRID 1 project we focused on the problem of data and resource sharing within collaborations distributed in two different continents. Within the activities supported in the project we chose to study the problem of a physics experiment led by a Chinese-Italian collaboration. The experimental site is located in Tibet region in China and the data need to be transported and processed at the two main computing centres of the Collaboration, located at IHEP-Beijing in China and CNAF-Bologna in Italy.
In the following we will refer to the specific problem, although the solution we propose could be applied equally well to other cases.

Data taking is organized in RUNs, a period of data taking during which conditions are kept reasonably constant. Each RUN is made of several files (about 1 Gbyte each, in our specific case). The experiment aims at a very high duty cycle, and the expected amount of data collected in one day is of the order of 300 GB. The computing resources available at the experimental site allow only for some limited data processing and data storage, whereas it is estimated that the computing activities related to data processing, simulation and analysis require of the order of 500 kSPECint2000 2.

This amount of resources is not available at a single computing centre so we developed a GRID approach. This environment provides us with the needed CPU power and storage space, and furthermore provides more features like redundancy of the services, security access and enforces the definition of a common environment accessible from any site.

Data Moving

We started to develop the so called "Data Mover" application to transfer data from the experimental site to the collaboration computing centres using gLite 3 Grid services. The "Data Mover" application is based on four Grid services: the Storage Element (SE) 4, the File Transfer Service (FTS) 5, the Logical File Catalog (LFC) and the User Interface (UI) 7.
The FTS is the component that permits to move in a controlled way the data from a SE to another, provided that they both support the Storage Resource Manager (SRM) 8 interface. The FTS service works with "channels" that connect the SEs. A channel is a named uni-directional logical connection from one SE to another and it is configurable in terms of bandwidth, number of streams, access policies, etc. Transfer of one file or of a group of files is called a "job". FTS jobs are processed asynchronously: upon submission a job identifier is returned, which can be used at any time to query the status of the transfer.

The LFC permits to the GRID users to assign a logical name to a physical file present on a SE. The association is one-to-many, a logical name can point to several physical copies of the same file ("replicas").

The UI is the gateway to the Grid, where users are authenticated and authorized to use the gLite Grid services.
The graphical sketch of our system follows: the arrows indicate the FTS channels and the arrow type (continuous or dashed) distinguishes channels owned by different FTS servers.

The system was built with a certain degree of redundancy, using more FTS servers and defining many channels for the same destination. To keep the system as simple as possible, no FTS server needs to be installed at the experiment site.
At the experiment site, data from the data acquisition (DAQ) system are sent to the local storage system and routinely migrated to the SE by a program running in crontab 9. Since, in our case, the SE and the DAQ machines share the same disk the migration involves no data copying, rather only the metadata information stored in the SE database is updated.

As soon as a run has been successfully migrated, a flag is set in the local DAQ database.

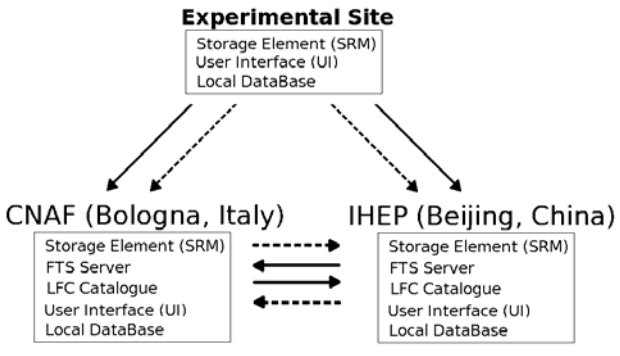


Fig. 2. Schematic view of "Data Mover" Application.

The Data Mover application is written in Perl 10 that is one the most powerful languages for scripting: moreover, APIs written in Perl are available for the FTS. The monitoring application is written in Java language 11.
The application consists of three modules.

The first module takes care of initiating the transfer of runs from the experiment site to one of the computing centers. The status of the transfer is mapped to the Data Mover database. For each run, the start time of the transfer is recorded into the database. The starttime field is used to identify new runs which still need to be transfered and it is also used to decide to retry the transfer after a timeout in case of some malfunctioning of the system. When a new run is scheduled for transfer we register the identifier (ID) of this job into the database, for later checking of the status of the transfer.
The second module has an instance running at each computing center (China and Italy) and takes care of synchronizing the two LFC catalogues. This job compares entries in the two catalogues: if a file that is registered in the remote catalog is not present in the local catalog, a transfer is started from the remote to the local site. Upon successful completion of the transfer the local file is registered in the local catalogue and the remote copy is registered as a replica. This module also takes care of updating the Data Mover database.

The third module is the garbage collector, which is responsible for cleaning up the buffer disk at the experiment site. All files that have been correctly transferred and which have two distinct replicas in the remote LFC catalogues are removed from the buffer disk. If there is still need for disk space, the garbage collector starts to copy the files to tape. Upon successful copy to tape, the files are deleted. The tapes will then be sent to one of the main computing centres, and there files will be stored in the SE and registered in the LFC catalogue.

To perform a monitoring and check on the data transfer we have also implemented a very simple DB in which we store useful information and that is described in the following Fig. 7.

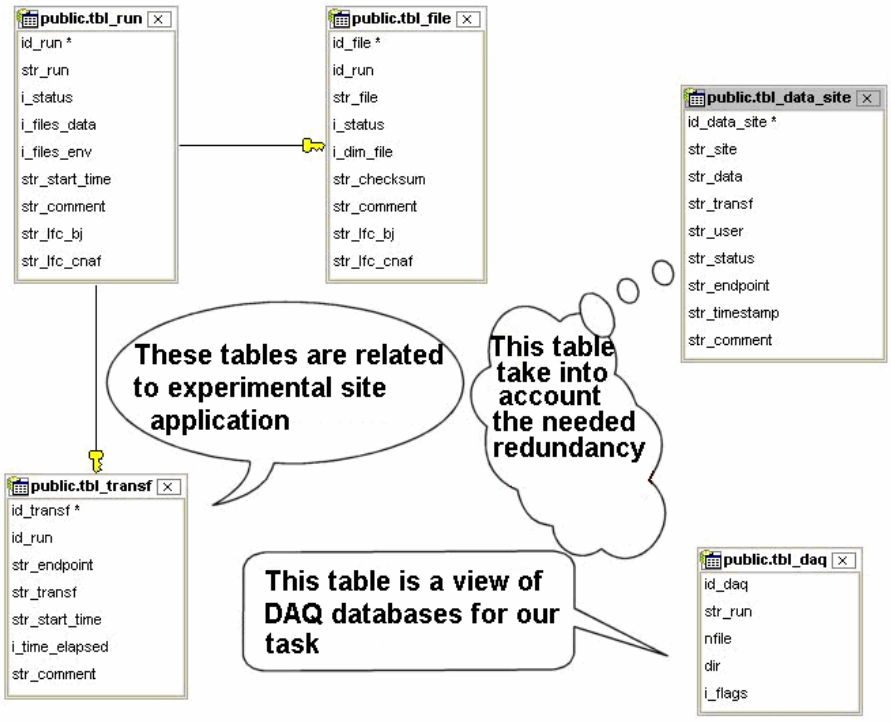


Fig. 3. DB structure

All these applications are supported by a textual monitoring that writes on a logfile (with the possibility to enable different levels of detail) and performs also status checks of the main services to produce warnings and alarms.

To complete the work we also developed a Java Graphical User Interface (GUI) application that enables a very useful and easy check on all the working modules.

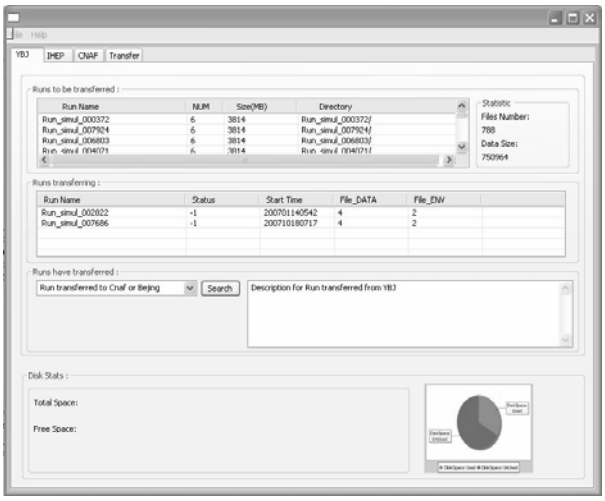


Fig. 4. Graphical User Interface

Porting of data processing activities to GRID

Why to use GRID

The GRID technology gives us the possibility to use more distributed computing resources to take advantage of High Throughput Computing and thus reduce the total wall-clock processing time. This approach is particularly suited for MonteCarlo simulations, which typically consist of CPU-bound jobs for which the amount of needed input data is generally limited and the executable can possibly have very little dependencies on external software. However, the GRID approach should also be considered for massive data processing applications, especially if the CPU time needed to process a block of input data is much larger than the time needed to make those data available to the remote worker node via the network. In our specific case, the time needed to reconstruct 1 GB of raw data is of the order of several hours, to be compared to the few minutes needed to transfer 1 GB over WAN.

Definition of a common environment

A working environment, for a geographically spread community, is based on the concept of Virtual Organization (VO). The experiment's VO was created, as first step, and roles were defined: then, all collaborating sites were asked to support such VO. The mutual acceptance of the Digital Certificates issued by the Chinese and Italian Certification Authorities was also established.

An important benefit which stems naturally from the adoption of GRID technologies is the enforcement of the definition of a common environment, meaning anything from the definition of common policies, such those on how to organize files or to label software versions, to the definition of procedures for data reconstruction and the use of the same calibrations.

A script managing software installation at remote sites was prepared. Each software package is distributed as a tar archive and has an associated tag which will be published by the CE upon successful installation. The tag name has a simple structure consisting of a set of strings with obvious meaning, separated by the '-' (dash) character for easy parsing, like:

VO-<experiment>-<test|prod>-<program_name>-<program_version>-<architecture>

The installation script is executed as a GRID job by the software manager: it accepts switches to perform such operations as software installation, validation and removal. After each step a temporary tag composed by the concatenation of the software tag and the status is published by the CE: this temporary tag is used by the installation script to make sure that the operations are performed in the right order.

The experiment's software, as well as other software needed to satisfy dependencies, was installed both at sites which are collaborating with the experiment and at sites which support the Argo VO even though the collaboration is not present. All software was installed under the path pointed to by the environment variable VO_<experiment>_SW_DIR, as is usual for EGEE-like 12 sites.

As far as the file organization is concerned, a common logical naming convention was defined for all kinds of files, for raw data as well as files produced by data reconstruction or Monte Carlo simulation. The format of the logical names resembles that of a physical filename, and consists in a set of strings separated by the '/' character. The tree-like structure was chosen because:

- it can be easily mapped to a physical file-system as well as to the logical file-system provided by tape systems;
- it allows for easy navigability;
- it can be defined in such a way as to limit the maximum number of entries at each level and at the same time be “descriptive”, making some characteristics of the data files apparent at a glance, like the time the file was taken, or the release used for reconstruction.

The logical name uniquely identifies a file at any site and in any environment, be it a disk or a tape file-system, or the LFC catalogue, for example: the logical to “physical” mapping is accomplished by pre-pending the logical name with a site-specific prefix (meaning there will be one prefix for mapping to disk at each site, another one for mapping to tape, yet another one for mapping to LFC catalogue and so on).

Application porting

After the experiment's official software had been installed at a bunch of sites, we started porting the experiment's applications to GRID: we focused on the data reconstruction application and the Monte Carlo simulation.

Both applications require submitting a large number of jobs which are very similar to one another, differing only in a small number of parameters (run number, input and output filenames, energy range, calibration file,...), most of which can be computed dynamically (for example, directly by querying a database, or indirectly by combining other parameters together). The main differences between submitting one such job to the GRID or to a local farm are that in the former case:

- one needs also an accompanying Job Description Language (JDL) 13 file;

- the scripts should be a little more general, e.g.: the scripts should not rely on a specific absolute path for the application executable, but rather make use of the appropriate environment variable;
- keeping track of the jobs can be more difficult, since the Resource Broker (RB) or the Workload Management System (WLMS) 14 dynamically and independently schedule jobs to potentially any available CE which matches with the Job requirements.

For these reasons, the GRID case can be regarded as a generalization of the submission to a local farm, so we developed general procedures which could work in both environments.

Such procedures rely on a few Perl scripts, some configuration files and a number of template files. Configuration files contain lines with the format:

VARIABLE := value

where "value" can be an arbitrary string: a small Perl library contains all routines needed to use variables in the definition of other variables, or within the scripts, or for making substitutions in the template files.

Such a scheme is extremely powerful, as many important changes can be performed without the need for editing the Perl code, but just simply editing the template files. The porting of the applications to GRID required us to extend the procedures to handle the generation of the JDL file, and to modify the template files.

To allow for simple tracking of the jobs, the experiment's production database was extended in such a way as to store the jobID returned at submission time: the jobID is stored as a string, whose value clearly distinguishes between local and GRID jobs. For the initial simplified version, we decided to have just one master production database, which is mirrored by remote sites. The production database can be accessed from remote nodes so that each one can update the status flag relevant to its own running job. Should the worker node be unable to contact the database for any reason, a recovery procedure was prepared: this procedure runs on any User Interface and periodically checks and updates the status flag for all jobs running since too long time.

Sites structure

In this specific case, the experiment's two main computing centres are peers, so we defined a symmetrical architecture with two GRID production sites for raw data reconstruction, both storing a copy of the raw and reconstructed data, continuously aligned.

The GRID architecture we are going to use is shown in Figure 4. Each production site will keep a copy of the raw data and of the reconstructed data, a LFC catalogue, a BDII information systems and the UIs.

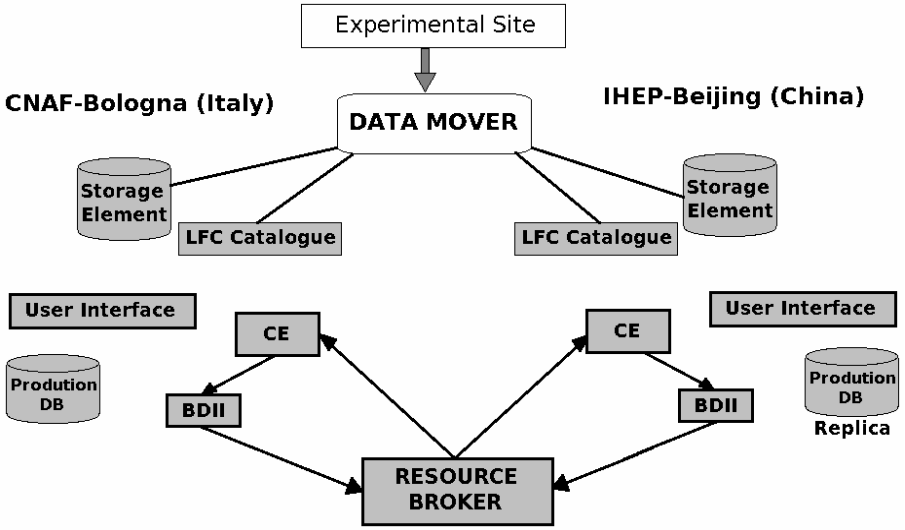


Fig. 5. Grid architecture

As already mentioned, there will be only one active production database that will be inquired to discover the raw data files to be processed: the request for computing resources will be forwarded to a Resource Broker (RB). The jobs will be submitted according to the availability of computing resources and automatically the data files will be read in input and written in output on the local SE. The computing center at CNAF is using a SE of the type SRM/CASTOR 15, while at IHEP the SE is of the type SRM/dCACHE 16.

Once the job is finished the database will be updated and the reconstructed events files will be copied to the other computing centre's SE through the same procedure for LFC catalogues alignment described earlier for the Data Mover. For safety reasons the production database will be mirrored, thus also allowing users at different sites to enquire and select data for physics analysis independently of the status of the network links.

Current status

The procedures for Data Mover have been tested on a test-bed including the sites of INFN Roma Tre, CNAF and IHEP. Some minor configuration problems for FTS servers were detected and solved. Some performance problems still

persist during the file transfers and should be solved by the network managers through better routing procedure.
The job submission was tested using different user interfaces and computing resources both in Italy and in China.

Conclusions

A general Grid approach to the data transfer and processing has been presented, which can be applied to many cases of scientific and non-scientific data which need to be analysed in a geographically distributed environment.

The automatic procedure to transfer the experimental data using the GRID middleware tools allows both a good control and monitoring of the operations and the fast availability of the data to the processing system, through the use of LFC catalogues.

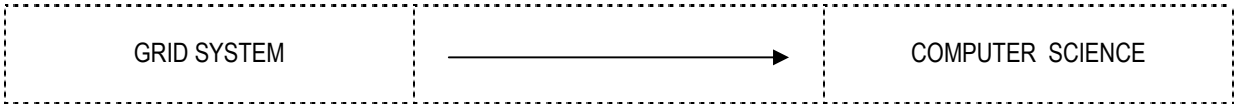
Acknowledgements

The authors are grateful to the ARGO-YBJ experiment for the supportive collaboration in the work.
The presented activity has been performed in the framework of EUChinaGRID, an European co-funded Project.

References and Glossary

1. EUChinaGRID Project: http://www.euchinagrid.eu
2. SpecINT2000: http://www.spec.org
3. gLite: http://glite.web.cern.ch/glite/
4. Storage Element (SE): https://twiki.cern.ch/twiki/bin/view/LCG/DpmGeneralDescription
5. File Transfer Service (FTS): http://www.gridpp.ac.uk/wiki/GLite_File_Transfer_Service, http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS/
6. LCG File Catalogne (LFC): http://www.gridpp.ac.uk/wiki/LCG_File_Catalog, https://twiki.cern.ch/twiki/bin/view/LCG/LfcGeneralDescription
7. User Interface (UI): http://glite.web.cern.ch/glite/documentation/R3.0/default.asp, http://www.gridpp.ac.uk/deployment/users/ui.html
8. Storage Resource Management RM: http://sdm.lbl.gov/srm-wg/
9. crontab: http://www.opengroup.org/onlinepubs/009695399/utilities/crontab.html
10. Perl: http://www.perl.com/
11. Java Language: http://java.sun.com/
12. EGEE: http://public.eu-egee.org/
13. Job Description Language (JDL): http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0102-02-Document.pdf, http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0142-0 2.pdf
14. Resource Broker & WLMS: https://edms.cern.ch/document/572489/, https://edms.cern.ch/document/674643/
15. The CASTOR Project: http://castor.web.cern.ch/castor/
16. dCache: http://www.dcache.org/

Glossaries of Grid terms:
http://www.gridpp.ac.uk/gas/
http://egee-jra2.web.cern.ch/EGEE-JRA2/Glossary/Glossary.html
http://grid-it.cnaf.infn.it/fileadmin/users/dictionary/dictionary.html



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 39-43

HIGH THROUGHPUT PROTEIN STRUCTURE PREDICTION IN A GRID ENVIRONMENT

GIOVANNI MINERVINI[*], GIUSEPPE LA ROCCA[**], PIER LUIGI LUISI[*], FABIO POLTICELLI[*x]

[*]*Department of Biology, University Roma Tre, 00146 Rome, Italy;* [x]polticel@uniroma3.it
[**]*INFN Sezione di Catania, 95123 Catania, Italy*

Running title: PROTEIN STRUCTURE PREDICTION IN GRID

Abstract : The number of known natural protein sequences, though quite large, is infinitely small as compared to the number of proteins theoretically possible with the twenty natural amino acids. Thus, there exists a huge number of protein sequences which have never been observed in nature, the so called "never born proteins". The study of the structural and functional properties of "never born proteins" represents a way to improve our knowledge on the fundamental properties that make existing protein sequences so unique. Furthermore it is of great interest to understand if the extant proteins are only the result of contingency or else the result of a selection process based on the peculiar physico-chemical properties of their protein sequence. Protein structure prediction tools combined with the use of large computing resources allow to tackle this problem. In fact, the study of never born proteins requires the generation of a large library of protein sequences not present in nature and the prediction of their three-dimensional structure. This is not trivial when facing 10^5-10^7 protein sequences. Indeed, on a single CPU it would require years to predict the structure of such a large library of protein sequences. On the other hand, this is an embarassingly parallel problem in which the same computation (*i.e*. the prediction of the three-dimensional structure of a protein sequence) must be repeated several times (*i.e*. on a large number of protein sequences). The use of grid infrastructures makes feasible to approach this problem in an acceptable time frame. In this paper we describe the set up of a simulation environment within the EUChinaGRID infrastructure that allows user friendly exploitation of grid resources for large-scale protein structure prediction.

Introduction

Simple calculations show that the number of known natural proteins is just a tiny fraction of all the theoretically possible sequences. As an example, the latest release of UniProtKB/Swiss-Prot (release 52.5, 15 May 2007) contains 267354 sequence entries [1], many of which are evolutionary related. On the other hand, considering random polypeptides of just 100 amino acids in length (the average length of natural proteins being 367 amino acids [1]), with the 20 natural amino acids co-monomers it is possible to obtain 100^{20} chemically different proteins. This is an astronomically large number which leads to the consideration that there is a huge number of protein sequences which have never been exploited by nature. In other words a huge number of "never born proteins" (NBP) [2]. This arises the fundamental question if the set of known natural proteins have particular features which make them eligible for selection, in terms, for example, of particular thermodynamic, kinetic or functional properties. One of the key features of natural protein sequences is their ability to fold and form a stable and well defined three-dimensional structure which in turn dictates their specific biological function [3]. From this viewpoint, the study of the structural features of NBP can help to answer the question if the natural protein sequences were selected during molecular evolution because they have unique properties and which are such properties (for instance a peculiar amino acid composition, hydrophobic/hydrophilic amino acids ratio, etc.). Such a problem cannot be easily tackled with an experimental approach which would require the production and structural characterization of a large number of random polypeptides. Attempts have been made in this direction [2], however we chose to tackle the problem using a computational approach to generate a large number of random proteins sequences with no significant homology with natural proteins (see accompanying paper by Evangelista et al.,) and to study their structural properties by means of the well known *ab initio* protein structure prediction software Rosetta *abinitio* [4]. However, to obtain statistically significant results the size of the sequence data base to be analysed must be sufficiently large (at least 10^5 to 10^7 sequences). This is a highly demanding problem from a computational viewpoint. In fact on a single CPU it would require years of computing time to predict the structure of such a large number of protein sequences. On the other hand, from a computational viewpoint this is an embarassingly parallel problem in that the same computation (*i.e*. the prediction of the three-dimensional structure of a protein sequence) must be repeated several times (*i.e*. on a large number of protein sequences). Grid infrastructures are highly suitable tools to approach this kind of problems in that a large number of grid computing elements can be used to execute relatively simple calculations. In this paper we describe the deployment of the

Rosetta *abinitio* software on the GILDA testbed (see below), as a first step towards porting of the software in the EUChinaGRID grid infrastructure. The development of a user friendly working environment within the GENIUS portal is also described, which allows the submission of a large number of protein structure prediction simulations with the final aim of structurally characterizing a large database of NBP sequences.

Methodological issues

The Rosetta software

Rosetta *abinitio* is an *ab initio* protein structure prediction software which is based on the assumption that in a polypeptide chain local interactions bias the conformation of sequence fragments, while global interactions determine the three-dimensional structure with minimal energy which is also compatible with the local biases [4]. To derive the local sequence-structure relationships for a given amino acid sequence (the query sequence) Rosetta *abinitio* uses the Protein Data Bank [5] to extract the distribution of conformations adopted by short segments in known structures. The latter is taken as an approximation of the distribution adopted by the query sequence segments during the folding process [4].

In detail, Rosetta workflow can be divided into two modules:

Module I - Input generation - The query sequence is divided in fragments of 3 and 9 amino acids. The software extracts from the data base of protein structures the distribution of three-dimensional structures adopted by these fragments based on their specific sequence. For each query sequence a fragments data base is derived which contains all the possible local structures adopted by each fragment of the entire sequence. The procedure for input generation is rather complex due to the many dependencies of module I. In fact, to be executed the first Rosetta *abinitio* module needs the output generated by the programs Blast [6] and PSIPRED [7] in addition to the non redundant NCBI protein sequence database [8]. On the other hand this procedure is computationally inexpensive (10 min of CPU time on a Pentium IV 3,2 GHz). Thus it has been chosen to generate the fragments database locally with a perl script that automatizes the procedure for a large dataset of query sequences. The script retrieves query sequences from a random sequence database in FASTA format (see accompanying paper by Evangelista et al.) and executes Rosetta *abinitio* module I creating an input folder with all the files needed for the execution of Rosetta *abinitio* module II. Approximately 500 input datasets are currently being generated weekly with this procedure.

Module II - Ab initio protein structure prediction – Using the derived fragments database and the PSIPRED secondary structure prediction generated by module I for each query sequence, the sets of fragments are assembled in a high number of different combinations by a Monte Carlo procedure by Rosetta *abinitio* module II. The resulting structures are then subjected to an energy minimization procedure using a semi-empirical force field [4]. The principal non-local interactions considered by the software are hydrophobic interactions, electrostatic interactions, main chain hydrogen bonds and excluded volume. The structures compatible with both local biases and non-local interactions are ranked according to their total energy resulting from the minimization procedure. A single run with just the lowest energy structure as output takes approx. 10-40 min of CPU time, for a 70 amino acids long NBP and depending on the degree of refinement of the structure. Rosetta *abinitio* Module II has thus been deployed on the GILDA testbed through the use of the GENIUS interface (see below) with the option of parametric jobs submission to run a large number of jobs, as required for the study of the large library of NBP generated.

The GILDA testbed

GILDA (which stands for Grid Infn Laboratory for Dissemination Activities) is a virtual laboratory of the Italian National Institute of Nuclear Physics (INFN) to demonstrate and disseminate the capabilities of grid computing [9]. Within the GILDA virtual laboratory, the GILDA testbed is a series of sites and services (Resource Broker, Information Index, Data Managers, Monitoring tool, Computing Elements, and Storage Elements) on which the latest version of the INFN Grid middle-ware, compatible with gLite, is installed.

Results

Integration of Rosetta Module II on the GILDA grid infrastructure

Single job execution on GILDA - A single run of Rosetta *abinitio* Module II consists of two different phases. In the first phase an initial model of the protein structure is generated using the fragment libraries and the PSIPRED secondary structure prediction. The initial model is then used as input for the second phase in which it will be idealised. A shell script has been prepared which registers the program executable (pFold.lnx) and the required input files (fragment libraries and secondary structure prediction file) on the LFC catalog, calls the Rosetta *abinitio* Module II executable and proceeds with workflow execution. A JDL file was created to run the application on the GILDA working nodes which use the gLite middleware [10].

Integration on the GENIUS web portal - A key issue to attract the biology community towards the exploiting of the Grid paradigm is to overcome the difficulties connected with the use of the grid middleware by users without a strong background in informatics. This is the main goal that has to be achieved in order to disseminate the use of grid services by biology applications. To achieve this goal and allow a wide biologists community to run the software using a user friendly interface, Rosetta *abinitio* application has been integrated on the GENIUS (Grid Enabled web eNvironment for site Independent User job Submission) Grid Portal [11], a portal developed by a collaboration between the italian INFN Grid Project [12] and the italian web technology company Nice [13]. Thanks to this Grid portal, non-expert users can access a grid infrastructure, execute and monitor Rosetta abinitio application only using a conventional web browser. All the complexity of the underlying grid infrastructure is in fact completely hidden by GENIUS to the end user. In our context, given the huge number of NBP sequences to be simulated, an automatic procedure for the generation of parametric JDL files has been set up on the GENIUS Grid Portal. With this procedure, exploiting the features introduced by the last release of the

gLite middleware, users can create and submit parametric jobs to the grid. Each submitted job independently performs a prediction of the protein structure.

Hereafter is described in detail the workflow adopted to run Rosetta *abinitio* application on GENIUS. After the user has correctly initialized his personal credentials on a MyProxy Server, he can connect to the GENIUS portal and start to set up the attributes of the parametric JDL that will be created "on the fly" and then submitted to the grid. First the user specifies the number of runs, equivalent to the number of amino acid sequences to be simulated (Figure 1). Then, the user specifies the working directory, the name of the shell script (Rosetta *abinitio* executable), to be executed on a grid resource, loads a .tar.gz input file for each query sequence (containing the fragment libraries and the PSIPRED output file) and specifies the output files (initial and refined model coordinates) in parametric form (Figure 1). The parametric JDL file is then automatically generated and visualised in order to be inspected by the user and submitted (Figure 1). The status of the parametric job as well as the status of individual runs of the same job can be also checked from within the GENIUS portal.

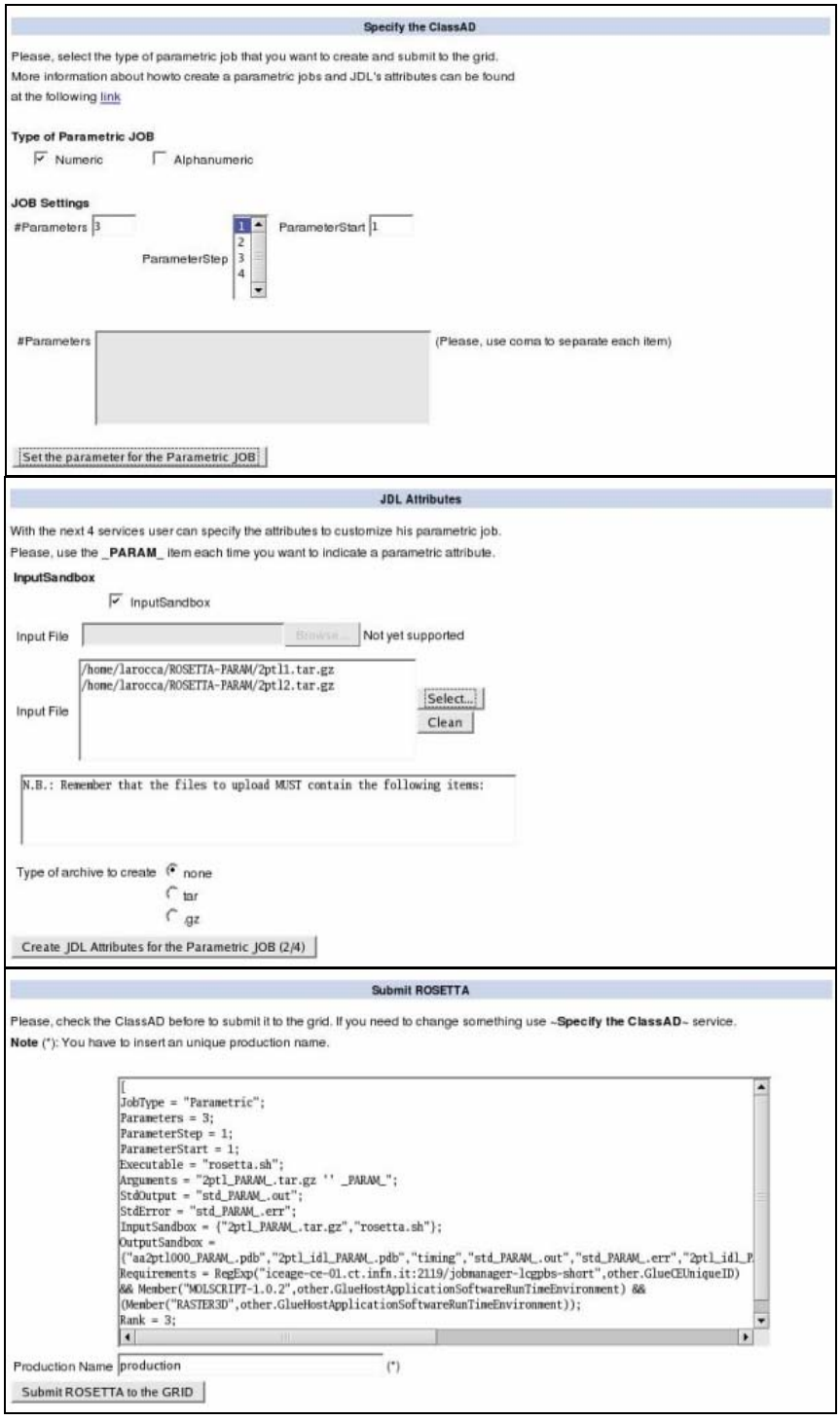


Fig. 1. Screenshots of the GENIUS grid portal showing services for the specification of the number of structure predictions to run (top panel), of the input and output files (middle panel) and for the inspection of the parametric JDL file (bottom panel).

When the prediction is done it is also possible, using the portal, to inspect the output produced in graphics form. Figure 2 shows the graphical output of the predicted structure in "spacefill" representation generated by Raster3D [14] in .png format. In addition, in order to allow the user to analyse the predicted NBP structural model, the JMOL Java applet [15] has been embedded into the GENIUS portal. A JMOL representation of a predicted NBP structure is also shown in Figure 2.

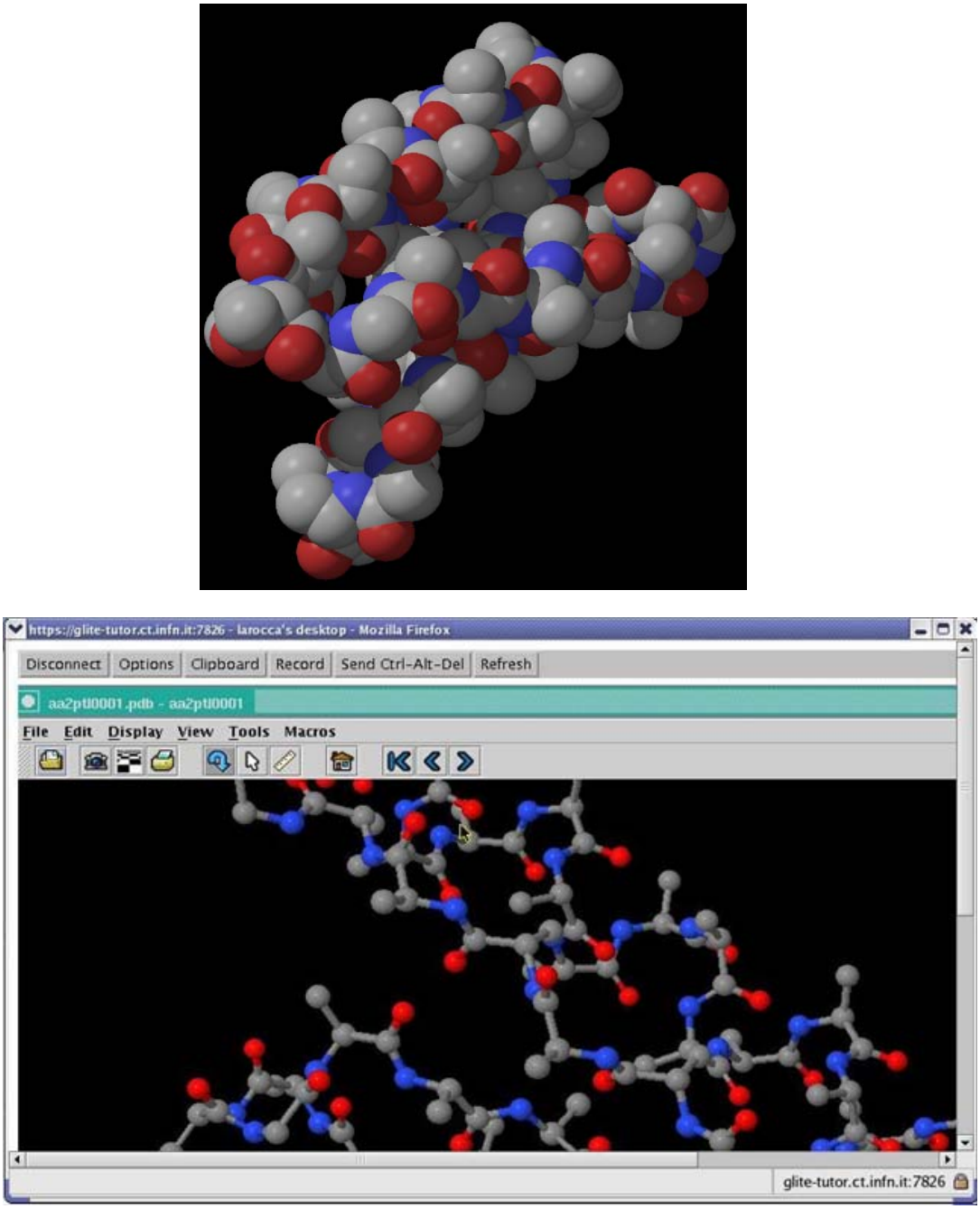


Fig. 2. Graphical output of a protein structure prediction generated from within the GENIUS grid portal using Raster3D (left) and JMOL (right).

Conclusions

Grid technologies are attracting increasing interest in the biology community due to the possibility to approach computational biology problems highly demanding in terms of both computing and data storage resources. Protein structure prediction is one of the major challenges in computational biology in that a huge amount of data are available for protein sequences while this is not the case for the corresponding three-dimensional structures. On the other hand, knowledge of the three-dimensional structure of a protein opens up the way for the comprehension of its function and molecular mechanism, a critical step in key areas of biomedical research. From this viewpoint, the importance of the deployment of Rosetta software in grid goes beyond the study of NBPs. In fact, the same tool can be used to tackle equally

complex and demanding biological problems, such as for instance the prediction of the structure and function of the entire set of proteins of a bacterial pathogen or a viruses, allowing the selection and study of suitable targets for drug design.

Acknowledgements

This work has been supported by a European Commission grant to the project "EUChinaGRID: Interconnection and Interoperability of grids between Europe and China" (contract number: 026634).

References

1. Bairoch A., Boeckmann B., Ferro S., Gasteiger E.: Swiss-Prot: Juggling between evolution and stability, Brief. Bioinform. 5, 39-55, 2004
2. Chiarabelli C., Vrijbloed J.W., De Lucrezia D., Thomas R.M., Stano P., Polticelli F., Ottone T., Papa E., Luisi P.L.: Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display, Chem. Biodivers., 3, 840-859, 2006
3. Branden C., Tooze J.: Introduction to protein structure, Garland Publishing, New York, 1999
4. Rohl C.A., Strauss C.E., Misura K.M., Baker D.: Protein structure prediction using Rosetta, Methods Enzymol., 383, 66-93, 2004
5. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., The Protein Data Bank, Nucleic Acids Res., 28, 235-242, 2000
6. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J. Mol. Biol., 215, 403-410, 1990
7. McGuffin L.J., Bryson K., Jones D.T., The PSIPRED protein structure prediction server, Bioinformatics, 16, 404-405, 2000
8. Wheeler D.L., Barrett T., Benson D.A., Bryant, S.H., Canese K., Church D.M., et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 33, D39-D45, 2005
9. GILDA – https://gilda.ct.infn.it/
10. gLite middleware - http://glite.web.cern.ch/glite/
11. GENIUS Portal – https://genius.ct.infn.it/
12. INFN Grid Project – http://www.infn.it/
13. Nice – http://www.nice-italy.com/
14. Merritt E.A., Bacon D.J., Raster3D Photorealistic Molecular Graphics, Methods in Enzymol., 277, 505-524, 1997
15. Jmol: An open-source Java viewer for three-dimensional molecular structures. http://www.jmol.org/

EnginFrame Framework – http://www.enginframe.com/
RASTER-3D - http://skuld.bmsc.washington.edu/raster3d/

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 45-49

AN APPROACH TO PROTEIN FOLDING ON THE GRID – EUCHINAGRID EXPERIENCE

M. MALAWSKI[*], T. SZEPIENIEC[**], M. KOCHANCZYK[***], M. PIWOWAR[***], I. ROTERMAN[***].

[*]*Institute of Computer Science, AGH, Al. Mickiewicza 30, 30-059 Krakow, Poland*
[**]*Academic Computer Center CYFRONET, ul. Nawojki 11, 30-950 Krakow, Poland*
[***]*Department of Bioinformatics and Telemedicine, Jagiellonian University, Collegium Medicum, ul. Sw. Anny 12, 31-008 Krakow, Poland*

Abstract: Contemporary pharmacology in its quest for more relevant and effective drugs needs to examine large range of biological structures to identify biological active compounds. We consider large grid environment the only platform to face such a computational challenge.
In our project, the search is focused on peptide-like molecules containing about 70 amino acids in a single polypeptide chain. The limited number of proteins existing in the nature will be extended to those, which have not been recognized in any organisms ("never born proteins"). The assumption is that those which do not exist in the nature may also render biological activity, which directed on pharmacological use may correct some pathological phenomena.
As the function results from the structure, two approaches are applied to predict cartesian coordinates of proteins' atoms: sophisticated Monte Carlo structure creation, elimination and refinement using the Rosetta program and our own program for simulation of the protein folding process.
As a computing platform we use the EuChinaGRID project resources, which are currently a part of EGEE infrastructure and are expanding to include Chinese resources as well. We describe the approach for porting the application to the grid and the prototype portal developed for simulation management and results analysis.

Key words: protein folding simulation, grid system, pharmacology, drug design

Introduction

Fundamental research for individualized therapy

Contemporary pharmacology, which is expected to be ready to design the therapy in the individual manner for each patient, is facing a large challenge. The fast pace of drug design, which is assumed to satisfy all specific expectations related to particular disease and to particular patient, is the critical issue. The know-how in chemical disciplines seems to be developed on the satisfactory level. Computer equipment and software are also ready to be applied. The only missing link in simulation of biological processes is theoretical and then numerical ability to predict three-dimensional structure of protein. These are the molecules responsible for most of the processes in each living organism. It was evidenced that the function of protein molecule is determined entirely by its structure. This is why the search for reliable numerical models allowing correct structure prediction on the basis of known amino acids sequence is necessary to make the progress in fast new drug design aimed to correct disfunctional proteins. All steps of so called "central dogma of molecular biology": from nucleic acids to biological function, seem to be recognized to the extent of understanding the mechanism of disfunction called disease. Instead of "drug design" understood as correction of proteins' activity, the "therapy design" extends all over the steps of biological dogma and is placed in the focus of modern pharmacology. The processes of larger and larger systems need to be simulated in silico. The experiments (forbidden in vivo) are possible in silico and seem to be unlimited. The solution of this problem in its complete form seems to attainable in the relatively close future.

Simulation of the protein folding process

The only important step: accurate automated prediction of the three-dimensional structure is still unavailable in the in silico form. This is why many efforts are undertaken to solve this problem. This is also why the very large computer resources of grid-like size are exploited for protein structure prediction programs. Despite of 30-years long history, the correct model able to recreate the path according to which the unique spatial structure of polypeptide chain is formed, is still missing. The problem seems to be a hot one for life sciences nowadays. There are some tools like Rosetta which applied to particular amino acid sequence are able to suggest (only suggest) the native conformation of protein. The outcome of this method (treated so far as the best one) is the structure of limited confidence. The question: How do the proteins fold ? remains still unanswered. The probability based models are not able to give the reliable answer to this question. As long as the mechanism of folding is unrecognized, the models for its modification are also unavailable.

The "never born proteins" project

The task of massive protein structure prediction for 70 amino acids long polypeptides (10^7 of them) has been undertaken by the international team within IST EUChinaGrid Project [1]. The team joins experts from two disciplines: biochemistry specialists in protein structure prediction and computer science specialist in grid system. O predict the structure, the Rosetta method is applied and also a technique elaborated recently in JUCM [2], which is an attempt to simulate the protein folding process rather than protein structure prediction, is harnessed. Within the project the application was prepared to run on European and Chinese grid-resources.

In this paper we describe the protein folding application, focusing on the step which we needed for porting it to the grid. We also give the overview of the additional tools based on a portal, which were developed to simplify the process of management of running such a large-scale application on the grid and to aid biochemists interested in result analysis.

Related work

Distributed processing infrastructures such as grids or peer-to-peer systems has been used for protein folding since a relatively long time. The examples include early experiments using CHARMM software and the grid infrastructure [3]. There are also widely recognized projects that exploit the power of thousands of PC machines voluntarily offered by the participants via the BOINC platform clients [4]: pioneering Predict@home [5], Folding@home [6], that performs distributed molecular dynamics simulations, and Rosetta@home, which is powered by Rosetta software [7]. Predictor@home was recently taken offline, Folding@home concentrates on simulations of the folding pathways of single proteins etiologically related to specific diseases, and Rosetta@home is devoted to bringing to perform their jigsaw-puzzle-game-alike method. The initiative of Human Proteome Folding Project [8], running on the infrastructure of Grid.org and World Community Grid and also using Rosetta software, produced a database of predicted structures of all human protein domains that were not yet resolved experimentally.

Running the application on the grid

The initial application comprised software developed by the JUCM team were prepared to run on a single machine or local cluster. When porting it to the grid such as EGEE, where the basic processing unit is a batch job, it is necessary to analyze application workflow in order to identify basic tasks and their data dependencies. The task should be possibly coarse-grained, since the overhead of job submission and batch system execution is considerable.

Stages of simulation

The JUMC protein folding application consists of three main stages shown in Fig.1.: early and late stage folding followed by the active site recognition. Given a sequence, creation of the early stage is entirely polypeptide backbone dependent [9] and requires a large contingency table with precomputed locations of tetrapeptides i a limited conformational subspace [10]. Additionally at the step, eventual steric clashes between distant amino acids are detected and resolved. The late stage works with such an intermediate structure by the introduction of side chain interactions that are extended by an external force field expressing hydrophobic character of some amino acids [11]. Its impact on the structure is evaluated as the discrepancy between actual and expected hydrophobicity, dependent on the distance from the centroid according to the gaussian distribution and assigned due to the own normalized scale distribution ("fuzzy-oil-drop"). Alternating with internal energy minimizations in the ECEPP/3 force field prevents atoms from overlapping. Distance relations between rigid elements of such small peptides hinder the molecule to cover itself thoroughly with hydrophylic residues providing hint for the location of the active site [12].

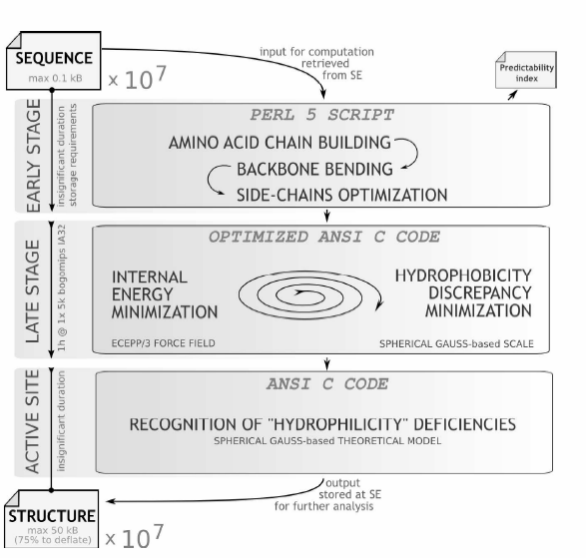


Fig.1. Stages of the protein folding process simulation

Steps for porting to the grid

After identification of logical stages of the application, it is necessary to consider also the technical side of the software, such as executables, library dependencies, and input/output files. The following steps were needed to grid-enable the folding application.

1. For all programs used in workflow all the required packages, which were not available on the grid worker nodes, were collected. For example, the library of sequences required for early-stage and code dependencies for late-stage were included.
2. The main script was created for running the application. It is responsible for proceeding with workflow execution checking if results of each stage are available. The parameters of this script are a sequence string and an identifier of the sequence.
3. It was decided to register results of computation for single sequence in a separate file on the grid storage, namely in LFC catalogue. The file name includes the sequence identifier and the resulting protein is stored in PDB format.
4. A self-containing bundle of programs and libraries needed for executing the application was created. This bundle was also registered in LFC catalogue.
5. A script was created to prepare installation of application on site each time when job is started and to spawn the main scripts with appropriate parameters.

6. Finally, the JDL (Job Description Language) file for gLite middleware was created. Performing these steps resulted in a self-contained application, which can be executed on the grid infrastructure without any pre-installation required. This is especially convenient for running it on many virtual organizations within EGEE (euch as Euchina and VOCE VOs) and also on Chinese grid infrastructure.

UI/Portal

As the number of simulation tasks and produced structures is huge and the way they are processed and finally interpreted is homogenous, we have developed a portal for job submission and monitoring and for data analysis, that appreciably simplifies the interaction of the average user with the complex infrastructure of the grid. Portal was developed in GridSphere Portal [13] using GridwiseTech LCG-API package [14] to cooperate with grid infrastructure.

Job submission and monitoring

In Fig. 2 the most important features of the portal and data-flow in the application were presented. Using the application portal job submission is performed (step 1). Typically this is done by uploading a file, in which up to several thousand sequences are listed with their's identifiers. The portal creates a separate grid job for each sequence and it adds them to the submission queue. Jobs from the queue are submitted to grid using LCG-API Job Monitor. This is done according to specified policies that could prevent from flooding VOs with to many grid jobs. A single job running on grid computing element download an application package to the working node (step 2), compute the results and save them to the grid storage system (step 3). Results of the jobs are validated by the portal with a post-processing analysis routines. In case of positive validation the results are registered in the results database, otherwise decision what to do next are left for the portal operator. Portal services analyze also the results of jobs failures and decide whether to resubmit a job or rather to ask the operator what to do next.

At runtime the operator can monitor computations using the application portal. Monitoring in the portal was designed to face the large amount of jobs running at the same time. The portal implements features like grouping, browsing by various criteria, viewing statistic and listing of current problems-to-solve.

Finally, basing on database, the results of computations can be browsed and accessed (step 4) for analysis by a set of tools described in next paragraphs.

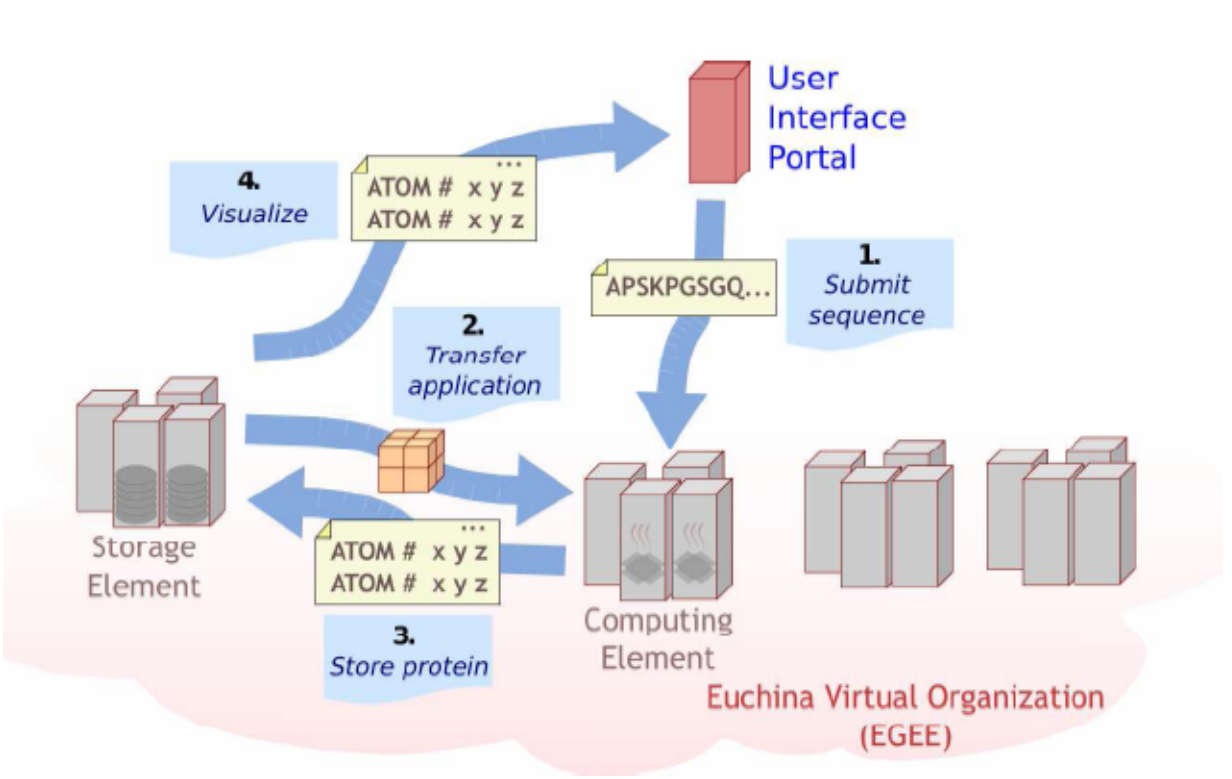


Fig. 2. User Interface portal in the context of the grid infrastructure

Result analysis

In the part of the portal devoted to the result analysis we provide conventional tools that are familiar to biochemists and biophysicists dealing with proteins. After choosing the id of a resulted structure, remote secondary structure assignment is performed remotely in DSSP [15] and presented graphically.

Using JUMC Structural Bioinfo Toolkit, that operates on the server-side and generates images to the virtual framebuffer using Java2D, the Á/A map with preferred areas and contact map with different distance cut-offs are displayed.

We also reingeineered the MBT Protein Workshop [16] in order to enable immediate visualization in the classic cartoon-like representation. On the basis of our Toolkit we developed a specialized molecular viewer that is able to point the location

of the probable active site using a color scale. Molecular surface is computed remotely in MSMS [17] and retrieved via Java RMI. If a protein was synthetized in a wet biology laboratory and has undergone a 2D electrophoresis, within the portal it is also possible to get an estimate location of the molecule in the gel (portlet with a Curl wrapper to the ExPASy.org service).

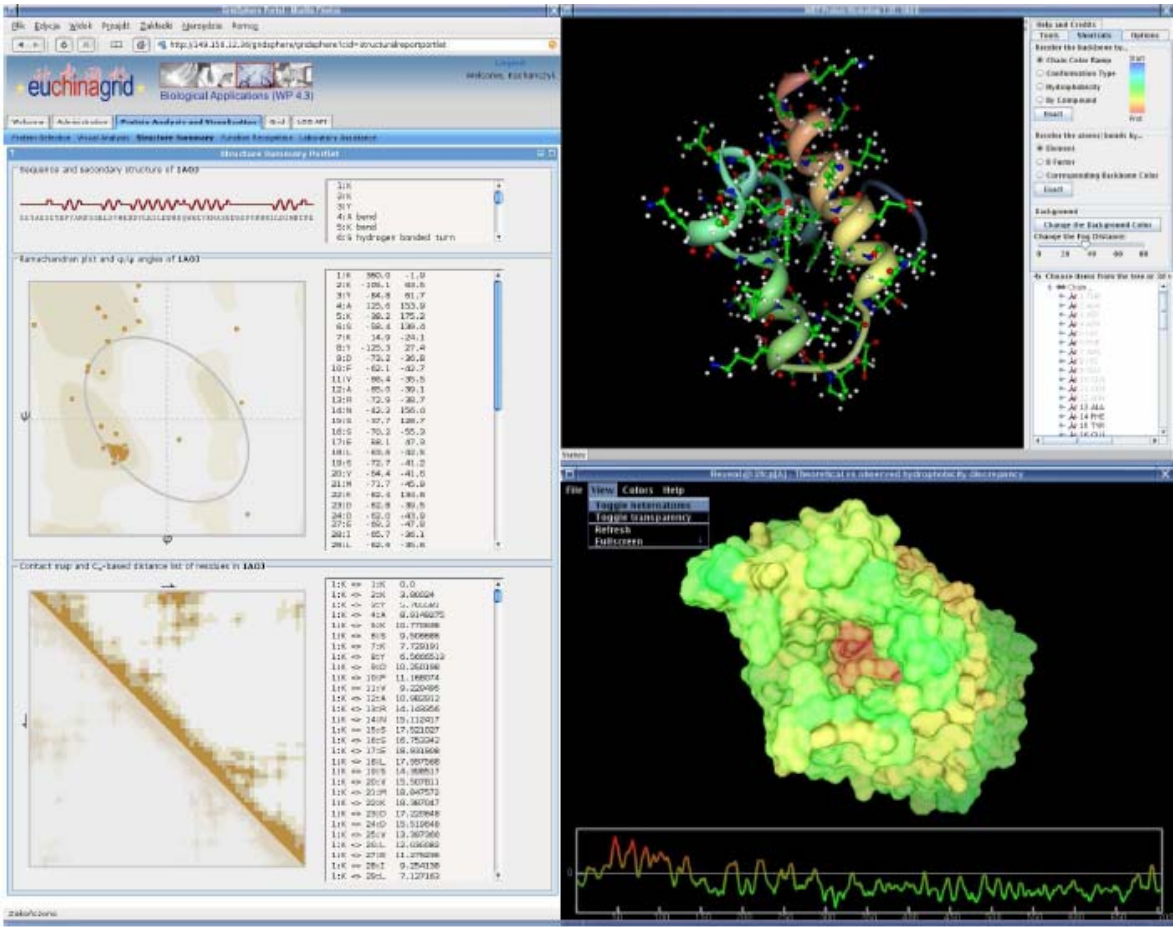


Fig. 3. Part of the portal for result analysis. Molecular viewers can be launched via JavaWebStart.

Summary and future work

The possibility to fold the proteins on such a scale applying two different methods is the great opportunity to test both of them. The mutual comparison of obtained results (according to Rosetta in cooperation with University Roma Tre [18] and basing on our mechanistic approach) is assumed to help to understand the nature of proteins in respect to their behavior in natural environment.

Moreover, the possible synthesis of the protein of assumed as pharmacologically active allows (recognized on the basis of predicted structure) immediate verification of obtained computational results (experimental partners in the Beijing University) and laboratory tests of harnessing them as potential new drugs.

The approach to running the application on the grid was tested on a sample batch comprising 10000 sequences and the prototype portal was used for demonstration purposes. Current work focuses on the development of a database for management of simulations and improving the usability of portal. Performing more tests will allow to verify both the simulation model and our portal toolkit.

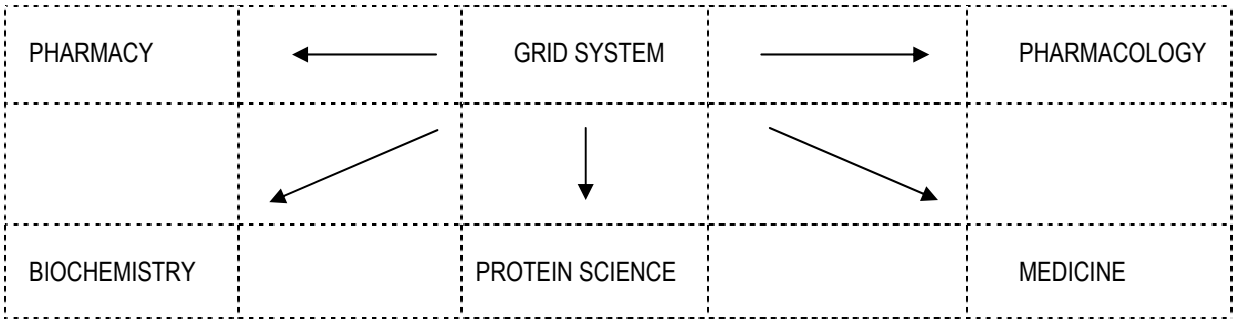
Acknowledgements

This work was partly funded by the European Commission, Project EUChinaGRID and by the related Polish SPUB-M Grant. Maciej Malawski kindly acknowledges the support from the Foundation for Polish Science.

References

1. IST EUChinaGRID Project. Project website: http://www.euchinagrid.org.
2. Brylinski M., Konieczny L., Czerwonko P., Jurkowski W., Roterman I.: Earlystage folding in proteins (in silico) – sequence to structure relation, J Biomed Biotechnol, 2 (2005) 65-79.
3. Natrajan A., Crowley M., Wilkins-Diehr N., Humphrey M., Fox A., Grimshaw A., and Brooks III C.: Studying Protein Folding on the Grid: Experiences using CHARMM on NPACI Resources under Legion. Proceedings of the HPDC Conference (2001), San Francisco, USA.
4. Berkeley Open Infrastructure for Network Computing. Project website: http://boinc.berkeley.edu.
5. Taufer M., An C., and Kerstens A., Brooks III Ch. L.: Predictor@Home: A 'Protein Structure Prediction Supercomputer' Based on Global Computing, IEEE Transactions on Parallel and Distributed Systems, 17, 8 (2006) 786-796.

6. Shirts M.R., Pande V.S.: Screen Savers of theWorld, Unite! Science, 290 (2000) 1903-1904.
7. Rohl C.A., Strauss C.E., Misura K.M., Baker D.: Protein structure prediction using Rosetta, Methods Enzymol, 383 (2004) 66-93.
8. Human Proteome Folding Project. Project website: http://www.grid.org/projects/hpf.
9. Jurkowski W., Brylinski M., Wisniowski Z., Roterman I.: The conformational subspace in simulation of early-stage protein folding, Proteins, 55 (2004) 115-127.
10. Brylinski M., Jurkowski W., Konieczny L., Roterman I.: Limited conformational space for early-stage protein folding simulation, Bioinformatics, 20 (2004) 199-205.
11. Brylinski M., Konieczny L., Roterman I.: Fuzzy oil-drop hydrophobic force field – a model to represent late stage folding (in silico) of lysozyme, J Biomol Struct Dyn, 23 (2006) 519-528.
12. Brylinski M., Kochanczyk M., Konieczny L., Roterman I.: Sequence-structure-function relation characterized in silico, In Silico Biol, 6 (2006) 0052.
13. GridSphere Portal. Product website: http://www.gridsphere.org.
14. GridwiseTech LCG/EGEE API – GridSphere Integration Kit. Product website: http://www.gridwisetech.com/content/view/91/96/lang,en/.
15. Kabsch W., Sander Ch.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22, 12 (1983) 2577-2637.
16. Moreland J.L., Gramada A., Buzko O.V., Zhang Q., Bourne P.E.: The Molecular Biology Toolkit (MBT): A Modular Platform for Developing Molecular Visualization Applications, BMC Bioinformatics, 6 (2005) 21.
17. Sanner M.F., Olson A.J., Spehner J.-C.: Reduced Surface: An Efficient Way to Compute Molecular Surfaces. Biopolymers, 38 (1996) 305-320.
18. Chiarabelli C., Vrijbloed J.W., De Lucrezia D., Thomas R.M., Stano P., Polticelli F., Ottone T., Papa E., Luisi P.L.: Investigation of de novo Totally Random Biosequences, Part II, Chemistry and Biodiversity, 3, 8 (2006) 840-859.



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 51-52

MASSIVE IDENTIFICATION OF SIMILARITIES IN DNA MATERIALS ORGANIZED IN GRID ENVIRONMENT

MONIKA PIWOWAR[*], TOMASZ SZEPIENIEC[*,**], IRENA ROTERMAN[*]

[*]*Department of Bioinformatics & Telemedicine Collegium Medicum UJ, Kraków, Poland*
[**]*ACK CYFRONET AGH, Kraków, Poland*

Introduction

The EUChinaGrid project is focused on the structure prediction of proteins of potential pharmacological application. Sequences of 70 amino acids long polipeptides (10^7 of them that are classified as 'never born protein') are used to generate their three dimensional structures [9, 10].

The aim of genomic part of the project is to search all accessible genetic information (listed Materials and Methods) completely sequenced as well as in progress to identify stretches of genomic sequence about biological function potential that were not identified to be exist in nature (proteins "never born" in evolution). Innovation of genomic part of the project rely on finding information about proteins in genomic regions where it theoretically should not be (in case of humane genome it is big amount genetic materials (about 97%) that does not encode any known proteins). Our attention is focused especially on regions including protein-coding gene fragments but also regions including other functional elements such as part of RNA genes and regulatory regions. The main task is defining the localization of similar nucleotide sequences (genome, chromosome, locus, structure of genetic sequence e.g. part of gene or known repetitive sequences) and statistical characterizing by cluster analysis and comparison of amino acids combination (kind of amino acid, type of physical and chemical properties).

Materials and methods

The complete DNA sequence is analyzed in respect to the presence of similar sequences to selected ones ("never born proteins") in noncoding region. The sequences of "never born proteins" for 10^7 polypeptides are generated randomly. The sequences of high significant sequence similarity to real proteins are excluded.

Entire genetic information is taken from National Center of Biotechnology Information (ftp.ncbi.nih.gov) to find similarities of sequences. It was searched not only humane genome but also other Eukaryotes genomes e.g. genomes of animals, plants, fungi and Protists genomes and organelles (listed in Table 1).

Table 1.

Mammals - Homo sapiens (human) - Mus musculus (mouse) - Rattus norvegicus (rat) - Bos taurus (cow) - Canis familiaris (dog) - Sus scrofa (pig) Other Vertebrates - Danio rerio (zebrafish)	Invertebrates - Anopheles gambiae (mosquito) - Caenorhabditis elegans (nematode) - Drosophila melanogaster (fruit fly) Invertebrates - Arabidopsis thaliana (thale cress) - Avena sativa (oat) - Glycine max (soybean) - Hordeum vulgare (barley) - Lycopersicon esculentum (tomato) - Oryza sativa (rice) - Triticum aestivum (wheat) - Zea mays (corn)	Fungi - Saccharomyces cerevisiae (baker's yeast) - Schizosaccharomyces pombe (fission yeast) - Magnaporthe grisea (rice blast fungus) - Neurospora crassa (orange bread mold) Protozoa - Plasmodium falciparum (malaria) Organelle genomes - Mmitochondrial genomes (757 in Metazoa) - Plasid genomes (57 in Eukaryota)

Genome sequences are translated to amino acids sequences in three reading frames. Because of diversity in type of translating genetic material among genomes the different genetic codes to proper genomes are applied (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c)

Database storing contigs that have information about nonbored protein is created to assist in fast finding information e.g. genomes of different organisms, size of contigs, accession number and create sub-bases that can be useful in further analysis. It is important also to perform statistical analysis describing genetic materials which is analyzed e.g. type of genomes, genetic codes, numbers and kind of repetitive elements, number of different gene structure, composition of sequences.

The sequences of not existing proteins that are found in genetic material are separated for further description and characterization. One of the aim of analysis under considera-

tion is to describe region in genomes with align sequences and type of structure that create. Important was description whether that sequences are part functional elements e.g. genes or pseudo genes, promoters, start codons, splite sites, introns, eksons, stop codons, polyadenylation sites that indicate the presence of a gene nearby. It is done by using in slico technique like evidence-base approach [8]. To obtain assumed results is used selected gene finding softwares (http://www.nslij-genetics. org/gene/programs.html/)

Another aim is description whether sequences are part of repetitive elements (mikrosatelites, minisatelites) and to which population of repetitive elements belong to [1,2,3]. Repetitive sequences are taken from Giri Institute server (http://www.girinst.org/repbase/update/index.html).

Cluster analysis methods for grouping homologous into sequences families are planned. When all searching results will be collected similar sequences will be used to construct hierarchical tree [7, 6, 5]. It is expected revealing groups of sequences which are related and give information about amount sequences in particular cluster, in the same way like in analysis that were done during genome-wide expression patterns analysis [4].

Scripts for translating nucleotide sequences to amino acids sequences are created in C programming language. Program for searching protein database is BLAST (http://www.ncbi.nlm.nih.gov/BLAST/).

Organizing Computations in Grid Infrastructure

All the DNA materials prepared for described above experiments are 15GB is size. All those data should be analyzed against large amount of sequences. Processing all the computation on average single-CPU machine would require about 200 days. Therefore, we organized computations in parallel mode on resources provided by EUChina Virtual Organization (VO). Resources in size of about 300 CPU enabled us to complete all computations in two days. Additionally, the results of our work was a framework, that would be used easily for any further computations in which BLAST package and selection of gather DNA material are used. Below we described the steps which enabled these computations on gLite-based grid.

1. We use LFC catalog available for VO, that we were using, to store all the materials that was prepared for the experiment. This enabled access to them from all machines taking part in computations.
2. Sequences that were subject of our experiment, were group in batches of about 100000 and stored on LFC catalog.
3. For the results of computations, a special space on storage was created.
4. The main script that automatize all necessary work (getting application copy, data transformation to common format, running blast, copying the output files to destination space etc.) was prepared. The script was parameterized by location of DNA materials, location of sequences batch. The names of output files stored in repository were created as combination of parameters.
5. A template grid job was constructed to process one batch and one material file
6. In grid portal, based on GridSphere framework (www.gridsphere.org) and LCGAPI services, we developed user friendly application submission. In the process of submission, the user is asked to specify the selection of application-related files, next, portal prepares and submit jobs accordingly.
7. User can observe the process of computations as well as analyze results of computation both in the portal and on the local machine after downloading them previously from grid storage.

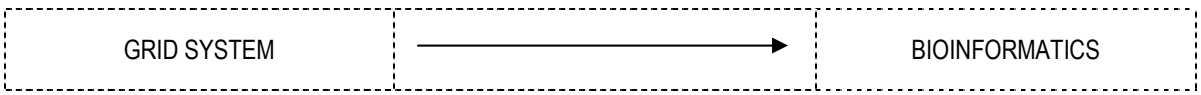
Having the framework prepared, we were able to carry out the whole experiment in 38 hours. Average usage of resources was 126 CPUs.

Summary

In this paper we presented, how computations related to Genomics, could be organize on gLite-grid. The grid environment gives opportunity to radically speed-up computational part in research on this field. The authors believe that similar methodology would be adapted to other applications as well.

References:

1. Jurka J., Kapitonov V.V., Smit A.F.A.: Repetitive elements: detection. Nature Encyclopedia of the Human Genome (Cooper, N.D., ed.), vol. 5, 9-14, Nature Publishing Group, London, New York and Tokyo 2003.
2. Jurka J.: Repetitive DNA: detection, annotation, and analysis. In Introduction to Bioinformatics: A Theoretical and Practical Approach. (Krawetz, S.A. and Womble D.D., eds), chapter 8, 151-167, Humana Press, Totowa NJ 2003.
3. Jurka J.: Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 9:418-420, 2000.
4. Eisen M.B., Spellman P.T., Brown P.O., Botstein D.: Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci U S A. 8;95(25):14863-14868, 1998.
5. Heyer L.J., Kruglyak S., Yooseph S.: Exploring Expression Data: Identification and Analysis of Coexpressed Genes, Genome Research 9:1106-1115, 1999.
6. Huang Z.: Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, 2, 283-304, 1998.
7. Jardine N., Sibson, R.: The construction of hierarchic and non-hierarchic classifications. The Computer Journal 11:177, 1968.
8. Saeys Y., Rouzé P., Van de Peer Y.: In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. Bioinformatics 23 (4):414-420. DOI:10.1093/bioinformatics/btl639, 2007.
9. Brylinski M., Jurkowski W., Konieczny L., Roterman I.: Limited conformational space for early-stage protein folding simulation. Bioinformatics 20, 199-205, 2004.
10. Brylinski M., Konieczny L., Roterman I.: Fuzzy-oil-drop hydrophobic force field - a model to represent late-stage folding (in silico) of lysozyme. J Biomol Struct Dyn. 23, 519-528, 2006.



BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 53-56

COMPUTERS IN MEDICINE

J. K. LOSTER, A. GARLICKI, M. BOCIĄGA, P. SKWARA, A. KALINOWSKA-NOWAK

Chair of Gastroenterology, Hepatology and Infectious Diseases – Collegium Medicum – Jagiellonian University – Cracow , Poland

Introduction

The computers have appeared in medicine since the 80th. The early models were difficult to be used. Nowadays, the tools produced by specialists in computer science are user friendly making the doctor work easier and much faster.

The opinion concerning computers given by medical doctor does not happen frequently. I would like to share my opinion – the opinion of the user active in practical medicine. Computers – so far – are mostly used as typing machine. Unfortunately, this exploitation is far from the ideal one omitting the powerful tools which are already available. In this paper I would like to present the areas in which – according to my opinion – the computers may be very helpful. The traditional hand-written documentation has been collected. This way not eliminates loosing of some documents, multiple rewriting of some information etc. The critical for electronic documentation is the education. The appropriate tools are ready and wait for adaptation. The education seems to be necessary in form of permanent education. The organization of education is difficult due to large dispersion of staff members and high differentiation of their duties. It seems that the education shall be taken into consideration during the preparation of work sheets. It makes possible the education for the complete staff of the hospital. The collaboration between many specializations is possible to be shown in such case.

Specificity of the work in hospital

The work in hospital is recognized as quite strange and complicated. This opinion is based on the computer programs, which appear as completely not applicable to the doctors expectations. The proposal is to invite the computer scientists to hospital to accompany the doctors and to recognize his work from his point of view. The aim of this technique is to recognize the specificity of the doctors' work.

The elimination of time consuming hand writing seems to be difficult because of the first step syndrome. The start to be fluent in any new discipline seems to the always the high barrier, the passing trough is the challenge for beginners. The training makes the master is the appropriate way to adopt the computers in medical practice.

Medical documentation

Computers are quite closely related to medical documentation. The extent to which his relation shall be depends on the hospital administration which is responsible for the work organization in hospital. The electronic medical documentation requires even country wide law regulation to ensure the validation of electronic documents by insurance companies, health ministry and rental system. The wide collaboration of high terdyinscyplinary character is expected.

Ambulatory

The place of first contact with new patent is the ambulatory. The revision and analysis of documentation making possible the first aid decision and additional diagnostic procedures is of high importance. This place requires special computer tools.

Registration

The registration procedure is highly time consuming for patent. The visit in the physician's office is unpredictable in respect of waiting time. The 30-40 minutes lasting visit requires few hours (or even the whole day) of waiting in the queue before doctors interview. This disorder is usually caused by the emergency patients, who are not able to wait. The easy way to patients registration is in private cabinets, where no emergency patients are accepted.

The time consuming operation is the documentation of personal data, which can be introduced in computer system before the visit starts. The commonly available computer system eliminates also the multiple introducing of common data (personal data) in different places in health care system. The access to the National Health System resources makes the administrative work much easier. Introduction of internet-mediated registration system could also help to solve the critical problem in medical care system organization.

The nation wide health care system enables also patients registration in the hospital localized in a certain distance versus the first aid doctor place. The free access too the doctor makes possible search for the less occupied medical care centers lowering the waiting time which is the wasting time.

National system of med care

The nation wide electronic system is working in scandinavian countries. Unfortunately the system in Poland is based on paper documentation. The only electronic related are sometimes the scanned documents (PDF format), which being uneditable are close for any text changes. The forms of multiple choice in electronic version could be of high acceptance and expectation by medical doctors.

Additional examination

The medical examination ordered by particular specialist must be very often extender also by the additional examination of highly specific character. For example the radiological examination is very often necessary to make the proper decision in many disciplines. It is very convenient to registry particular examination visible (available) for the other department, which is ready to reserve the time and place for the examination of patient, who did not appeared yet in the radiology department.

E-prescription

The prescription documentation is quite stressful for doctor, who are obliged to fulfill once again the next form writing the name address and other identification data of patent currently examined by the doctor. Electronic prescription could print all known data concerning the patients making the doctor work easier. The library of dosing system related to particular drug (as well as information about drugs interaction) available during the prescription preparation could make the doctor work easier and faster. The significant advantage is the serial prescription tool, which is able to print the prescription without the additional visit to the doctor. Analysis of the patients record during the therapy makes possible the correction introduced selectively by the doctor according to specially defined attributes for each user of the system.

Hospital

Appearance of patent in hospital is accompanied by the creation of patients' record. It should cover all possible diagnostic and therapeutic procedures. It ensures the availability of any procedure during the hospital therapy. The easy system of registration of procedures is the main advantage for medical doctor. The system activates also the administration record, the cost record. The patient appears immediately on the hospital drug store record, where all consumed drugs and injections are registered. The system summarizes the disappearance of all drugs per day eliminating possible lack of particular drug.

The controlled accessibility of "players" (doctors of different specializations, nurses, pharmacists and many others) must be controlled by system eliminating the intervention of non-profesionalists into the system.

Patients registration in hospital

The contact with the external world (anything else than hospital) making possible the cost calculation and staff engagement. The additional equipment (for handicapted patients requiring the every day help) may be traced by the insurance companies and particularly the National health System. The contact with the institution responsible for pension founds makes the information transmission of permanent form.

Medical interview

Medical interview is the step which also can be performer in electronic system particularly in the initial step of interview, when standard questions are asked. The availability of the complete record of the patient making available information describing the medical events, which happened years ago is the time saving procedure. The ideal solution in this case could be the movable computer, which can work independently on the local conditions.

Manual examinations

The physical examination produces the result expressing the presence of pathological symptoms. From the medical point of view important is the registration of pathology as well as their absence. The automatism of this procedure is expected to make simple the acceptance of non existing pathological events.

Reports

The therapeutic procedure starts when particular physical or pharmacological treatment is ordered for patent. The therapeutical treatment represents the dynamic event accepting corrections and modification depending on patients' reaction to the treatment. the procedures are usually conducted by other specialists including nurses. The electronic registration of procedural steps seems to be very convenient. On the other hand the new problem appears which is the responsibility. The electronic signature is expected to the integral part of the electronic procedure. The introduction of electronic signature is bottle neck of electronic system of registration. Some medical systems expect the doubled documentation: paper based and electronic.

The electronic registration of drug distribution equipped additionally by the analytical tool oriented on drug interaction seems to be of very helpful nowadays, when many newly introduced drugs can cause some troubles in this field.

The programs automatically printing the labels which pasted to the probes allow avoid the mismatch in clinical materials.

The substitution of temperature monitoring table (traditionally fixed on the patients bed) seems to be very useful especially when the possibility to save also other physiological parameters is available.

Additional duty of hospital doctor is the registration of hospital infections detected in particular hospital. The electronic transmission of information about such event seems to be also a useful solution.

Clinical analysis

Patients record of information

The tool of particular importance is the electronic record of information describing the patients history in hospital. The installation of electronic system for patients registration made the very time consuming procedure significantly shorter. The possible errors in data transmission are eliminated what allows avoiding of many misunderstandings in doctor work.

Time is important for therapeutic procedures

The traditional system of information distribution and contacts between particular departments is time consuming. The electronic communication between different departments becomes significantly faster. Some clinical analyses need traditional documentation what can be done independently according to traditional system. The accessibility to the information in electronic way is able to speed up the therapeutic procedures significantly.

Instant Access to the patients record

The common organization of patients record seems to be important. The accessibility to the data from different localizations in hospital (emergency room, diagnostic department, therapeutic department etc) is very useful for the doctor, whose work requires high mobility. The accessibility to the tools like Blackbery, Palm or PDA working in the system push-email seems to be the good solution. Such equipment is expensive however few computers of this kind per clinic could solve the problem.

Medical data analysis

The computers are very useful for medical data collection. The medical record of the patent with complicated disease after few weeks of hospitalization (diagnostic and therapeutic procedures) can be quite large. The storage of these data is of significant importance. The fast analysis of these data seems to be even more important. The tool allowing different analyses (not particularly including the statistical methods, just reports of different form) is very helpful and makes the work much easier. The tables giving the summary of symptoms with selected information important for particular patient shall be compatible with administrative documentation to avoid the redundancy in documentation.

Artificial intelligence

This discipline seems to be the most interesting one for computer programmers. The expert systems may be very useful for non professionalists. According to medical doctors opinion, it is not the most important part of programs applicable in medicine. The expert systems work well on the self-diagnosis field for example in alergology. The identification of some allergens may be achieved without the immediate contact with the doctors, although at the end of such procedure the advice to see the doctor shall be present.

Problems

Medical specializations and their influence on the programs

The project of computer applications in hospital shown above is not complete. The specializations of particular hospitals require also the special solutions for computer systems working in hospital. The misunderstanding between the author of the program and expectations of medical doctor makes the collaboration with computer program more difficult and even impossible. The both collaborators (medical doctor and computer science specialists) must exchange their opinions to make the programs tolls compatible to the specificity of medical discipline. The difficulties in mutual comprehension shall be patiently solved on the basis of discussion and fast as possible removing of not appropriate program procedures.

Education

The introduction of new system is possible only on the condition of permanent education of hospital staff. The form of workshops seems to be the best educational form. The education shall be organized on the basis of already available data bases to demonstrate the program in its working form. It seems that the long educational system (few months) with the exam at the end of the course is the best form to ensure the proper program applicability. The e-learning technique seems to be the best form of permanent education of the members of hospital staff.

Advantages of electronic systems in practical medicine

The computer systems make the work of medical doctors optimal. The optimization of practical treatment seems to be in close relation to optimal costs ensuring good organization. The elimination of redundancy of some administrative documentation makes possible focusing on the medical problems. The issue of highest difficulty is the complexity of the system. It seems that the all-or-none procedure (the complete system instead of step-wise introduction of some parts of system) is the correct solution. It is of particular importance for hospital infections, which may be present in different hospitals and transported by particular patients. The worst solution is to implement the program without the consultation with medical doctors. Such solution may cause more disadvantages than advantages.

New discoveries available through the Internet

The access to the internet resources speeds up the application of new diagnostic and therapeutic procedures independently on the localization even in the globe scale. The availability of medical scientific libraries makes the innovative procedures popular. The possibility to share the experiences and opinion is of high importance in practical medicine. The study of professional publications becomes the standard in every day life of hospitals.

Permanent monitoring of epidemiological data

The large scale medical data collection in unified system applied for many countries seems to be critical for instant recognition of the events of epidemiological character. The data bases of appropriate size may help to give prognoses and allow the tools for preventive medicine. This kind of system covering all states in United States of America has been working since few years with the excellent feedback on the field of prevention.

Summary

The presence of computers with the specific medical tools implemented is increasing permanently. The specificity of medical doctors work shall be taken into account by specialists in programming thus the close collaboration is necessary to make the programs compatible to doctors expectations. The most complicated is the lack of systems communication. The network system satisfying doctors

expectations seems to make the collaboration with programmers much easier particularly in the issue of confidence. The discussion and mutual exchange of opinions seems to be critical for interdisciplinary collaboration in this field.

Large scale computing is of special interest for infectious diseases. The access to the net system and to data bases in the scale of the whole globe is of particular importance for AIDS outbreak. The variability of this virus traced in the planet scale may make possible the prediction of the prospective mutants enabling preparation of therapeutic (pharmacological) treatments in advance. This is why the access to the large scale net system based on grid system seems to be promising in the strategy of the AIDS therapeutic system.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 57

GRIDS AND THEIR ROLE IN SUPPORTING WORLDWIDE DEVELOPMENT

FEDERICA TANLONGO

GARR, Roma, Italy

Grids are a set of services over the Internet, allowing geographically dispersed users to share computer power, data storage capacity and remote instrumentation. Although Grids are still in a prototype phase, experts believe that they will have a dramatic impact, comparable to WWW, in the next few years; but while nowadays everybody knows the www and every day millions of people use it to share information all over the Internet, only a few of them are aware of the potential of Grid technology.

The basic concept is in the very word "grid", usually meaning the electric distribution system in English-language countries: electric power is indeed distributed to final users who are not aware how and where it was produced, nor they need to use it. With grid computing, it is just the same for remote resources.

In the next future, the global network of computers will become a whole, wide, computational resource that anyone may access on demand: users will exploit the same computing power of an enormous supercomputer just connecting from their PC.

Grid computing is in fact a particular example of distribute computing based on the idea to share resources on a global scale. Of course, in order to work properly, Grids need :

- An Authentication and Authorization system, providing secure access to resources, to guarantee data privacy and integrity (a critical factor in several application fields such as biomedicine);
- A mechanism (the so-called middleware) able to manage and allocate resources in an optimal way to all users and applications who need them, just like the Operative System does with programs running on your PC;
- A reliable, high-performance network connection amongst resources, ensuring that the time taken for data transfer is negligible in comparison with the benefit of quicker processing obtained thanks to distributed computing.

First Grids were developed in the framework of the so-called eScience, an innovative approach to research, thanks to the use of advanced technologies of communication and regardless to geographical location of instruments, resources and last but not least, brains.

A number of scientific applications characterized by very demanding requirements in terms of data processing, can benefit from this technology, which enables different computing centres, wherever located, to collaborate at the same computation with ideally (almost) the same effectiveness that they would reach if all their CPUs were in the same room.

Currently, Grid paradigm is being adopted in several application fields, such as astrophysics, theoretical chemistry, biomedicine, high-energy-physics, Climate, Earth Science, Archaeology, natural disaster prevention and so forth.

As you may imagine, the potential of such technology is enormous and affect not only a few scientist, but, in principle, each person or organization using or not computing and storage devices. Indeed, in the last few years, not only research institutions, but several private companies and major software houses as well as governments are approaching grids and investing on them.

Granting access to large resources with comparatively small investments in infrastructure, Grids may effectively contribute to bridge digital divide and fight the drain brain in developing countries, allowing researchers to participate in large collaborations by providing their intellectual contribution only.

OECD recognised the Grid approach's potential OECD in 2005, recommending "*The creation of new mechanisms (or the strenghtening of existing ones) to facilitate access to Grids for researchers and research organizations in developing countries, plus other appropriate measures to broaden international participation in grid projects*" [from 2005 OECD Global Science Forum].

The EC Research Infrastructures Programme supports this recommendation in the framework of the FP6 for Research and Scientific Development, trough funding a number of Grid infrastructure and applications projects aiming at promoting cooperation between EC and worldwide emerging countries.

Such projects set out to integrate the European Grid Infrastructure with other regions', in order to create one wide resource for scientists working on existing or future collaboration and involve scientific partners from all around the world.

EUChinaGRID (www.euchinagrid.eu), EU-IndiaGrid (www.euindiagrid.eu), EUMEDGRID (www.eumedgrid.eu) and EELA (www.eu-eela.eu) target respectively the collaboration with China, India, the Southern Mediterranean and Latin America.

In line with the support of the international extension of the so-called European Research Area (ERA), such projects aim at integrating the European Grid infrastructure with the ones available in other world regions, often using different middleware, in order to converge towards a whole and providing wide resources for the benefits of researchers working in international collaborations.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 59

GRIDS AT 4300 METERS OVER THE SEA LEVEL: ARGO ON EUCHINAGRID

C. STANESCU[*], F. RUGGIERI[*], Y.Q. GUO[**], L. WANG[**], X.M. ZHANG[**]

[*]*INFN Roma Tre, Roma, Italy.*
[**]*IHEP, Beijing, China.*

ARGO-YBJ is a Cosmic Ray Experiment in Yangbajing – Tibet (P.R. China), which has reached the full configuration by the end of year 2006 and is taking remarkable amounts of data. The YangBaJing Laboratory is pretty unique due to the altitude position of around 4300 m over the sea level which makes it the ideal laboratory for the study of Extensive Air Showers (EAS) in the region from 500 MeV (millions of electron Volts, measuring the energy) to few TeV (millions of MeV). On the other hand the position at such a highness makes difficult to maintain a stable crew of researchers to monitor and control the experiment. The usage of remote controls via network and remote access to processes and data is then mandatory to achieve an acceptable duty cycle in the long term. The foreseen data that will be accumulated by the end of 2007 is of the order of 100 TBytes (1012 Bytes) and larger amounts of data are expected every forthcoming year.

The needs of data exchange with the laboratory in Tibet and of strong collaboration with Chinese institutes for the analysis of these data is well in line with the typical applications of a world-wide GRID. Today several hundreds of Gigabytes of data are exchanged via Network while the usage of the magnetic tape cartridges is being limited to the final archive. Thanks a combination of Grid technology and high-bandwidth network connectivity, EUChinaGRID greatly improved the efficiency of the data transfer and dramatically helps the coordinated analysis of those large quantities of data between Europe and China.

Data and analysis results of the ARGO-YBJ experiment are of widen importance for a large worldwide community related to the investigation of the Gamma Ray Bursts (GRB). Gamma-ray bursts are short-lived bursts of gamma-ray photons, the most energetic form of light. At least some of them are associated with a special type of supernovae, the explosions marking the deaths of especially massive stars.

The Yang-Ba-Jing laboratory in Tibet is a part of the ARGO experiment

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 61

EUCHINAGRID: A HIGH-TECH BRIDGE ACROSS EUROPE AND CHINA

FEDERICA TANLONGO

GARR, Roma, Italy

Co-Funded by the European Commission in the framework of FP6, EUChinaGRID (www.euchinagrid.eu) aims at integrating major Grid infrastructures in Europe (EGEE) and China (CNGrid) for the benefits of eScience applications, thus facilitating existing and future collaboration between Europe and China. EUChinaGRID is also promoting the exchange of expertise with Chinese counterparts towards the deployment of new advanced services and applications in Grids, in line with the support of the intercontinental extension of the European Research Area (ERA).

With a total budget of 1.636.000,00 €, the project is coordinated by the Italian INFN (National Institute of Nuclear Physics) and involves several high-profile partners in Europe (CERN, Department of Biology, University of Roma Tre, GARR, GRNET, Jagiellonian University medical College) and China (Beihang University, CNIC - Chinese Academy of Sciences, IHEP, Peking University).

The project is supporting the implementation of an intercontinental pilot infrastructure, using in a first place EGEE-supported applications in order to validate the infrastructure, then facilitating the migration of new ones on the European and Chinese infrastructures. EUChinaGRID's first results were therefore to facilitate scientific data transfer and processing: pilot Physics (LHC), Astrophysics (ARGO), and Biology (Early/Late Stage, Rosetta), applications are already exploiting the new infrastructure, while helping in validating it.

EUChinaGRID Project officially started on 1st January 2006 and, during the first year of works already achieved several goals.

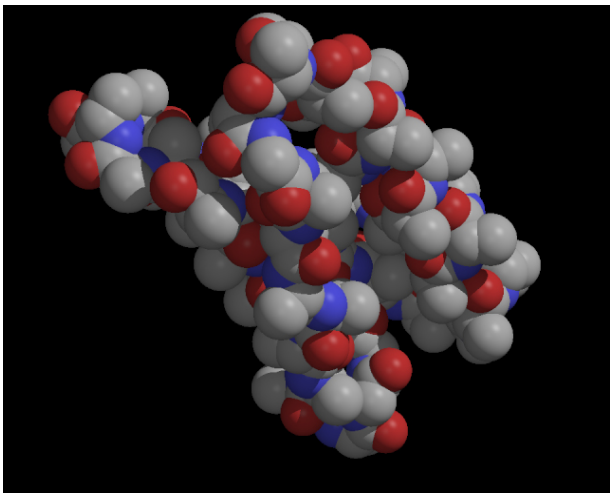
Started well ahead the foreseen plans, a first pilot infrastructure is up and running with 9 sites, 3 of which in China. All relevant Grid services were started and are maintained to facilitate the access of users and Virtual Organizations (VO) through the web portal.

A major challenge for the project is interoperability between European and Chinese middleware: reaching it, EUChinaGRID will provide the International scientific community a transparent access to a set of resources much greater than separately available in each environment. After a year of activity, a first version of the gateway between EGEE and CNGrid is already available.

A pioneering work has being carried out as well in order to achieve "vertical" interoperability, i.e. between Grid middleware and the different versions of the IP protocol, thus enabling the deployment of grid nodes in an IPv6 environment. First results of IPv6-Grid Middleware compliance test were published and widely disseminated to middleware developers; this activity brought to the deployment of a version of GOS, the middleware used by CNGrid, which is fully compliant with IPv6.

During the last quarter of 2006 EUChinaGRID project also supported with significant computing resources the WISDOM Data Challenge. This virtual screening challenge of the international WISDOM (World-wide In Silico Docking On Malaria) initiative started on 1st October and targeted compounds of interest for drug discovery against neglected diseases.

EUChinaGRID had an intense dissemination activity with about 500 students, users, developers and system administrators in Europe and China were involved in international conference, workshops and targeted training events.



Graphical output of a protein structure predicted with ROSETTA

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 63

RADIOLOGY ON GRID

ANDRZEJ URBANIK

Chair of Radiology – Collegium Medicum – Jagiellonian University – Cracow , Poland
Chair of Radiology – University Hospital , Krakow, Kopernika 21, Poland

The radiology is the discipline of medicine which deals with large databases of medical information. These are mostly of graphic form. The computer resources to conduct the standard medical analyses like RTG, USG, mammography, CT and MRI are appropriate to these analysis of this type of data.

There is only one medical diagnostic measurement which requires large scale computer resources. This measurement, very important for neurological diagnostics, is the visualization of brain functionality. This technique is based on the magnetic resonance phenomena (fMRI). The analysis of the source data (which is of large size) requires the massive calculation to be performed in a relatively short time scale to make possible the diagnosis complete.

According to our experiences, such measurements are performed rather unfrequently, what makes the availability of large scale computational resources and access to grid net of important meaning for practical medicine particularly in the radiology specialization.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 3, No. 5, 2007, pp. 65

GRID MONITORING IN EUCHINAGRID INFRASTRUCTURE

LANXIN MA

European Organization for Nuclear Research, Geneva - Switzerland
Institute of high energy physics, Beijing-China
Lanxin.Ma@cern.ch

EUChinaGrid is a Project which is founded by EU in 6th framework. Until now, EUChina grid infrastructure contains 10 sites which cover 4 counties in China and in Europe. There ara more than 1000 CPUs, 10 SEs, 10 CEs etc. in EUChinaGrid infrastructure to support EUChina issues. The project has many applications, such as high energy physics application, CMS , ATLAS, astroparticle physics applications (ARGO-YBJ / GRB) , biology application etc. So, it is important to provide reliable grid services, improve the reliability of the grid infrastructure and provide stakeholders with views of the infrastructure allowing to understand the current and historical status of the service. For this purpose, gridIce and SAM are used to monitoring EUChina grid infrastructure. In this paper, we present tools which are able to to check if a given grid service works as expected for a given user or set of users on the different resources available on a grid.

Selvita is a product and solution provider for the Life Sciences Industry. We employ a world-class team of dedicated medicine, chemistry, pharma, molecular biology, biotechnology and information technology professionals and enjoy a very good cooperation with leading Polish, European and U.S. universities and research institutes. We deliver comprehensive solutions to the customers from Life Sciences industry targeted at lowering the cost of introducing innovative therapeutical compounds to the market. We support our customers in the following ways:

- ✓ Implementation of innovative and cost effective information technology platforms accelerating research process and decreasing the risk of its failure
- ✓ Enabling access to databases with ready, preprocessed knowledge for research organizations which allows them to concentrate on their own creative process, rather than on mechanical knowledge acquisition
- ✓ Outsourcing of qualified R&D staff, specialized in subsequent phases of introduction of innovative pharmaceuticals to the market

We also develop our own innovative biologically active structures, which are the effect of research at the Polish universities, as candidates for further commercialization by our customers. The core strengths of the company are:

- ✓ molecule discovery & development
- ✓ bioinformatics
- ✓ contract research

Selvita implements advanced data processing solutions for the customers from biotechnology and pharmaceutical industry, including bioinformatics applications (proprietary products, products from our partners and customized solutions) in the area of genomics, proteomics, pharmacogenomics, PK/PD modeling and cheminformatics.

The functionality of our solutions comprises all activities of data processing: molecular modeling, storage, retrieval, data warehousing, data mining, integration and cleaning as well as applications for simulation, sequence analysis, clusterization, phylogenetic prediction, parallel processing, agent technologies, grid processing and data visualization.

Thanks to our solutions it is possible to design new compounds in silico based on quantitative structure-activity relations, computer analysis of receptor interactions, conformation analysis, calculation of physicochemical parameters of organic compounds. We also support construction of pharmacodynamics and pharmacokinetics models of biologically active compounds.

We also provide outsourced drug design and screening services based on customers' specifications with the utilization of our own IT infrastructure and software as well as high quality laboratory infrastructure. We also offer integration and access services to biological and other data over the Internet in compliance with international standards and compatible with publicly available databases.

Selvita also offers integration services in the area of IT infrastructure cooperating with commercial research equipment and bioreactor facilities. We also conduct our own research projects for the pharmaceutical customers in collaboration with leading universities and institutes. We are also interested in research cooperation on commercial and publicly funded (e.g. European Union) projects.

Please visit our website at www.selvita.com