GeneGrid: Grid Service Based Virtual Bioinformatics Laboratory

P.V. Jithesh,
Noel Kelly, Sachin Wasnik,
Paul Donachy, Terence
Harmer, Ron Perrott

*Belfast e-Science Centre,
Queen's University of
Belfast
{ p.jithesh, n.kelly, s.wasnik
p.donachy, t.harmer,
r.perrott}@qub.ac.uk*

Mark McCurley, Michael
Townsley, Jim Johnston

*Fusion Antibodies Ltd,
Belfast*

*{mark.mccurley,
michael.townsley,
jim.johnston}
@fusionantibodies.com*

Shane McKee

*Amtec Medical Ltd,
Belfast*

*shanemckee
@doctors.org.uk*

Abstract

GeneGrid is a collaborative industrial R&D project initiated by the Belfast e-Science Centre, under the UK e-Science Programme, with commercial partners involved in the research and development of antibodies and drugs. GeneGrid provides a platform for scientists, especially biologists, to access their collective skills, experiences and results in a secure, reliable and scalable manner through the creation of a 'Virtual Bioinformatics Laboratory'. It enables the seamless integration of a myriad of heterogeneous applications and datasets that span multiple administrative domains and locations across the globe, and present these to the scientist through a simple user friendly interface. This paper presents how the grid services of GeneGrid are involved in the integration of bioinformatics applications as well as in the creation and execution of in silico experiments. A real use case scenario is also presented, involving the identification of novel members belonging to a protein family, for demonstrating the capabilities of GeneGrid. Experiences from the adoption of standards such as OGSA and the integration of third party programs, are also presented.

1. Introduction

Genome sequencing and post-genomic technologies such as microarrays, are creating an explosion in the number of biological datasets to be managed, integrated and analysed, pushing bioinformatics to the forefront of disciplines that need huge computing power and highly collaborative environments. The emergence of grid computing technologies has opened up an unprecedented opportunity for biologists to integrate data from multiple sources, in spatially distant locations, which can be seamlessly analysed leading to a greater chance of knowledge discovery.

GeneGrid is a UK e-Science industrial project with the involvement of companies interested in antibody and drug development. The aim of GeneGrid is to provide a platform for scientists to access their collective skills, experiences and results in a secure, reliable and scalable manner through the creation of a 'Virtual Bioinformatics Laboratory' [1]. GeneGrid accomplishes the seamless integration of a myriad of heterogeneous resources that span multiple administrative domains and locations and provides the scientist an integrated environment for the streamlined access of a number of bioinformatics and other accessory programs through a simple interface. It allows biologists to create, execute and manage workflows that represent bioinformatics experiments. Such workflows automate and hence accelerate the experiments, preventing errors that usually creep in because of manual interventions.

This paper presents the architecture of GeneGrid and its implementation based on the existing international standards. Experiences from developing Open Grid Services Architecture (OGSA) [2] based grid services using Globus Toolkit for the integration of bioinformatics applications and databases as well as the use of GeneGrid in the creation and execution of *in silico* experiments are discussed.

2. GeneGrid Architecture

GeneGrid consists of a number of cooperating Grid services developed based on the OGSA and using Globus Toolkit ver 3 (GT3). GeneGrid services may be categorised logically into different components, namely Workflow Management, Resource Monitoring & Service Discovery, Data Management, Application Management and the Portal, which are discussed below.

2.1. Application Integration

Access to the bioinformatics applications available on various resources is provided by the GeneGrid Application Manager (GAM) [3, 4]. GAM achieves this integration through two types of OGSA-based grid services: GeneGrid Application Manager Service Factory (GAMSF) and the GeneGrid Application Manager Service (GAMS).

GAMSF is a persistent service, which extends the standard interfaces or Port Types, like GridServiceFactory of the Open Grid Services Infrastructure (OGSI) [5] to integrate one or more bioinformatics applications to the grid, and exposes them to the rest of the GeneGrid. The primary function of GAMSF is to create transient instances of itself called GeneGrid Application Manager Services (GAMS) which facilitate clients to interface with the applications.

Any client wishing to execute a supported application will first connect to the GAMSF and create an instance - the GAMS. This newly created GAMS then exposes to the client the operations which allow the client to execute the supported application as an extension to the operations provided by the OGSA Grid Service interface. Each GAMS is created by a client with the intention of executing a given application, and after completion of this task the GAMS is destroyed. Currently GeneGrid integrates a number of bioinformatics applications including BLAST [6], TMHMM [7], SignalP [8], ClustalW [9] and HMMER [10]. In addition, GAM also integrates a number of custom programs developed to link the tasks in a workflow. Figure 1(a) gives an overview of the components that provide the GAM functionality.

2.2. Database Management

The GeneGrid Data Manager (GDM) is responsible for the integration and access of a number of disparate and heterogeneous biological datasets, as well as for providing a data warehousing facility within GeneGrid for experiment data such as results [11]. The data integrated by the GDM falls into two categories. 1). Biological data consisting of datasets available in the public domain, e.g. Swissprot [12], EMBL [13] etc. and proprietary biological data private to the companies. 2). GeneGrid data consisting of data either required by, or created by GeneGrid, such as workflow definitions or results information.

GDM has used OGSA-DAI (http://www.ogsadai.org) as the basis of its framework, enhancing and adapting it as required, such as for providing access to flat file databases. GDM consists of two types of services, replicating those found in OGSA-DAI. The GeneGrid Data Manager Service Factory (GDMSF) is a persistent service configured to support a single data set. The main role of the GDMSF is to create, upon request by a client, transient GeneGrid Data Manager Services (GDMS) which facilitate interaction between a client and the data set (Figure 1b).

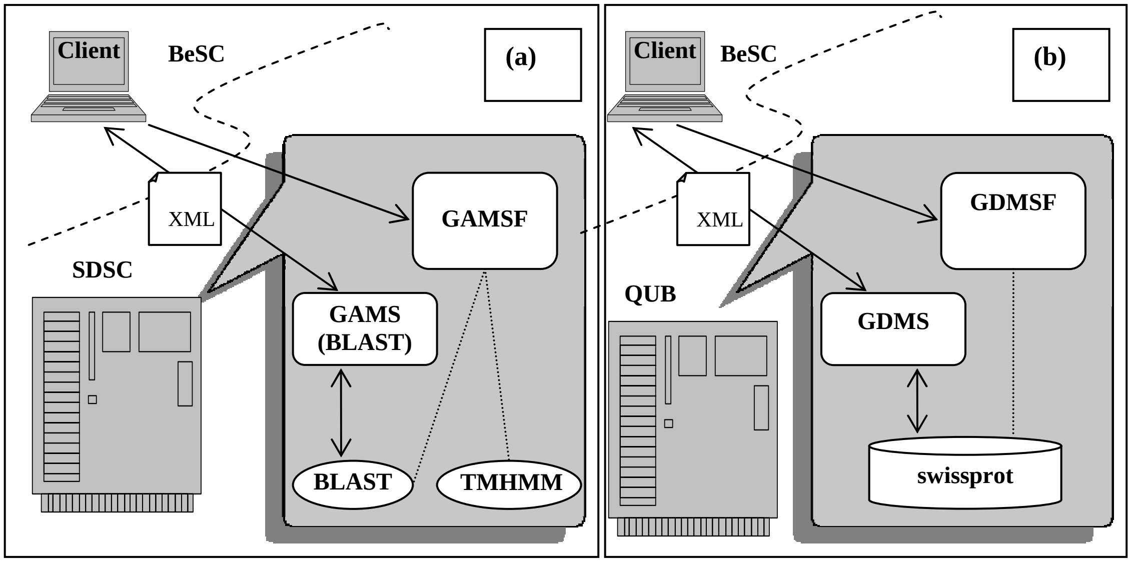


Figure 1. A client accessing (a) an application, e.g; BLAST on another resource through GAMS and (b) a database e.g: Swissprot through GDMS

2.3. Workflow Management

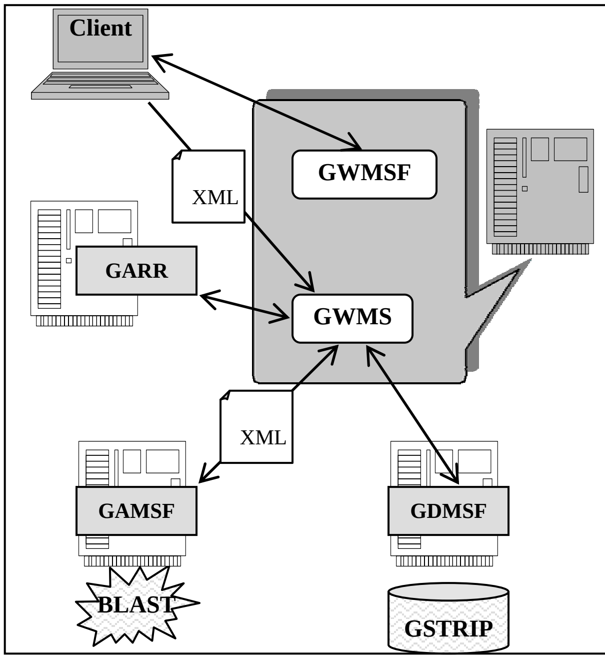
GeneGrid Workflow Manager (GWM) is the component of the system responsible for the processing of all submitted experiments, or workflows, within GeneGrid (Figure 2). As in the case of GAM, there are two types of services in the GWM. The first, the GeneGrid Workflow Manager Service Factory (GWMSF) is a persistent OGSA-based grid service. The main role of the GWMSF is to create GeneGrid Workflow Manager Services (GWMS), which will process and execute a submitted workflow across the resources available. Each GWMS is a transient grid service which is active for the lifetime of the workflow it is created to manage. The main roles of this service are to select the appropriate resources on which to run elements of the workflow, as well as to update the GeneGrid Status Tracking and Result & Input Parameters (GSTRIP) Database with all status changes. GWMS gets information on resources, databases, GDM services and GAM services through the GeneGrid Application & Resources Registry (GARR).

2.4. Resource Monitoring & Service Discovery

GARR is the central service in GeneGrid that mediates service discovery by publishing information about various services available in GeneGrid. A lightweight adaptor present on all the resources called GeneGrid Node Monitor (GNM) updates the GARR with the status of the resources, such as load average and available memory. In addition GNMs may also be configured to advertise details of the services deployed on the resources, such as service name, type, location and the database or application they integrate.

2.5. Portal

The GeneGrid Portal provides a secure central access point for all users to GeneGrid and is based upon the GridSphere product [14]. It also serves to conceal the complexity of interacting with many different Grid resource types and applications from the end users' perspective, providing a user friendly interface similar to those which our user community is already familiar with. This results in a drastically reduced learning curve for the scientists in order to exploit grid technology.



Submission and execution of a workflow containing a BLAST task. The client will first connect to the GWMSF to create a GWMS instance, before forwarding the workflow XML to the newly created GWMS. The GWMS identifies the BLAST task within the workflow, and queries the GARR for the location of all suitable GAMSF. The GWMS will then submit the task XML to the most appropriate GAM for execution. The GWMS will also find the location of the GDMSF serving the GSTRIP database from the GARR in order to submit updates as to the status of this BLAST task.

Figure 2. Workflow management by the GeneGrid Workflow Manager (GWM)

3. GeneGrid Component Integration

GeneGrid Environment (GE) is the collective name for the core distributed elements of the GeneGrid project, which allow the creation, processing and tracking of workflows. Contained within the GE is at least one GeneGrid Portal, at least one deployment of both the GARR and the GWMSF, an implementation of each of the GeneGrid Workflow Definition Database (GWDD) and the GSTRIP database, as well as at least one GDMSF configured to each of these databases. All instances of any factory services mentioned above may also be considered elements of the GE. By allowing users to access a GE, we create a Virtual Organisation (VO), and hence each GE may be considered as a single installation of GeneGrid.

Bioinformatics applications and datasets are exposed to the GeneGrid Environment by GAMSF and GDMSF respectively. These GAM and GDM services make up the GeneGrid Shared Resources. Each GAMSF and GDMSF advertises its existence and capabilities to a GE via GNM on their hosting nodes registering with the GARR. It is possible for GNM to register with many GARR services across multiple GE allowing the resources to be shared between multiple organisations. Therefore, organisations have complete control over what resources, if any, they wish to share with other GeneGrid organisations, forming dynamic virtual organisations.

4. GeneGrid Operation

Standard GeneGrid operation is workflow driven with scientists interacting with the system via the GeneGrid Portal to both generate and track workflows.

4.1 Workflow Creation

Having created GDMS for accessing both the GSTRIP and the GWDD, and having uploaded the Master Workflow Definition Document, the GeneGrid Portal is ready to create and submit new workflows. Users interact with the Portal to select required tasks, and fill out web based forms as presented by the Portal in order to generate a new Workflow XML document. The Portal will also upload information to the GSTRIP database as it is provided by the user. Once the user is happy with the workflow they have created, they may submit it for processing. The Portal will then connect to the GARR to retrieve the location of the GWMSF, and request it to create a GWMS instance. This newly created service is then sent the workflow XML for processing.

4.2 Workflow Execution

Having received the Workflow XML, the GWMS will proceed to break the workflow into its constituent tasks. The GWMS will connect to the GARR to find the locations of suitable GAM or GDM services capable of executing any tasks ready for execution. Tasks which rely upon the results of others are placed in a queue until such time as the information upon which they are dependant becomes available. The GWMS will also update the record for each task, and hence the workflow, within the GSTRIP database as appropriate.

4.3 Usability

GeneGrid development has seen a strong emphasis placed upon making the use of Grid technology as easy as possible for biological scientists. Currently, to perform an involved bioinformatics experiment with publicly available web sites would be a long tedious task with the user being asked to take a very “hands-on” approach at all times. This hands-on approach can be both time consuming and error prone. GeneGrid has automated this process considerably. Users may create a complex experiment from a single standard interface. As the linking of tasks together is automated, this considerably reduces the time required by the scientist to sit at a terminal! This point is emphasised further when we consider the recycling of experiment workflows by the end user to run the same experiment repeatedly with different input data each time. The absence of the need for manual intervention also reduces considerably the amount of errors that may “creep” into the process. Such automation has required the development of a number of custom “linker” applications for transforming the results of one application or database operation into something which can be understood by another. However, the GeneGrid Portal is designed to intuitively include such linker tasks into the workflow, allowing the scientists to concentrate on the applications and database with which they are familiar.

5. Use Cases

Scientists from the partner companies have tested the GeneGrid prototype by way of executing workflows which are biologically relevant. Such use cases have proved invaluable in providing feedback to the developers leading to bug fixes and further improvements. The use cases also set new requirements which led to the evolution of GeneGrid to a robust and versatile system. One such use case for the identification of novel protein family members is described briefly here.

Siglecs are a family of cell surface proteins belonging to the large Immunoglobulin superfamily, with a number of characteristic features shared among the members. They are involved in cell–cell interactions and signalling functions in the haemopoietic, immune and nervous systems [15]. The effort in this case study is to find new members of this promising family among the genomes using GeneGrid.

5.1 Use Case Workflow

In order to run such an experiment with existing technologies and resources would be a tedious time consuming task with quite a high risk of error. The scientist would take a known siglec sequence as the input for BLAST to find alignments, and when the results were available, extract all the accession numbers of interest to obtain the protein sequences. Each accession number would then have to be queried on line against the SwissProt database, with the resulting sequence being passed by the scientist into TMHMM. At this point, the experiment has forked from one experiment to multiple parallel experiments which must be tracked, and depending upon the BLAST configuration in the first step, the scientist could be tracking the progress of sequences numbering into the hundreds! The resulting TMHMM files must then be examined by the researcher, and if the file is of interest, the sequence used as input must be forwarded to SignalP, and again, the resulting files checked for success.

Using GeneGrid the scientist may create the experiment described above as a single workflow by supplying all the required input parameters. GeneGrid will automatically execute each task within the workflow once all the required data for that task becomes available. GeneGrid simplifies the experiment further by automatically including linker tasks on demand which are used to pre-process the results of one stage so that they are compatible with another e.g. BLAST results are processed to find all accession numbers before querying the SwissProt database.

This automation cuts the time required by the scientist to set up and track experiments considerably with GeneGrid managing the tracking of all experiment threads, and also eliminates the errors which creep in via manual intervention by the researcher. Finally, through the GeneGrid Portal, the scientist may track the progress of the experiment, examining the input and output files used at each point in the experiment.

Execution of the above workflow resulted in six uncharacterized and potentially new siglecs, which are currently being characterized using further procedures. Execution of the workflow, which would have taken about a day with conventional methods involving manual access to applications, took about 20 minutes in GeneGrid. This acceleration is largely due to the automation and parallelization of task execution, as well as the optimal use of available resources.

6. Discussion and conclusion

The use of grid technology and standards which are in the infancy has made GeneGrid development a challenging one. Furthermore, bioinformatics programs and databases usually have different proprietary formats and generally do not follow any standards. This has made the integration of multiple programs and data sources in GeneGrid quite difficult. A simple workflow which a biologist wishes to execute may not often suggest the underlying complexity in joining the tasks in the workflow.

The use of products from other projects also presented obstacles as those products are either in developmental stages or do not address GeneGrid’s requirements in its entirety. Sometimes such problems were alleviated by the extension of the third party products with custom solutions. For example, as GeneGrid required access to a number of flat file databases, OGSA-DAI product was extended to provide the required functionality.

The ability of GeneGrid to overcome these issues and provide automation of workflows shows distinctive advantages over conventional methods, a few of which are listed below.

- GeneGrid’s secure single access point provides users with a means to easily access many diverse applications and datasets without the need to visit many web sites.
- Automatic monitoring and selection of resources removes a major burden from the user, while ensuring an efficient allocation of resources.

- Considerably less time spent by the user creating and managing workflows.
- Errors which may creep in as a result of manual intervention are avoided.
- The ease of use of the GeneGrid front end means that scientists may exploit the promising potential of Grid technology while being insulated them from the inherent complexity of new underlying technology.

Thus, the development of a functional prototype of GeneGrid and its use in the problem of identifying new siglecs have clearly illustrated the viability of utilising grid services for integrating heterogeneous Bioinformatics programs with diverse requirements on different resources while following a workflow based approach.

7. References

[1] P. Donachy, T.J. Harmer, R.H. Perrott *et al*, “Grid Based Virtual Bioinformatics Laboratory”, *Proceedings of the UK e-Science All Hands Meeting (2003),* 111-116

[2] I. Foster, C. Kesselman, *et al.*, “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration”, *Open Grid Service Infrastructure WG, Global Grid Forum (2002)*

[3] P.V. Jithesh, N. Kelly, D.R. Simpson, *et al* “Bioinformatics Application Integration and Management in GeneGrid: Experiments and Experiences”, *Proceedings of UK e-Science All Hands Meeting (2004),* 563-570

[4] P.V. Jithesh, N. Kelly, Paul Donachy *et al* “GeneGrid: Grid Based Solution for Bioinformatics Application Integration and Experiment Execution”, *IEEE Symposium on Computer Based Medical Systems, Dublin (2005).*

[5] S. Tuecke, K. Czajkowski, I. Foster *et al.,* Open Grid Services Infrastructure (OGSI) Version 1.0. *Global Grid Forum Draft Recommendation, (6/27/2003).*

[6] S.F. Altschul, *et al*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.,* vol. 25, pp. 3389-3402, Sep 1. 1997.

[7] A. Crogh *et al,* “Predicting transmembrane topology," *J.Mol.Biol.,* vol. 305, pp. 567-580, Jan. 2001.

[8] J.D. Bendtsen, H. Nielsen, G. von Heijne and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *J.Mol.Biol.,* vol. 340, pp. 783-795, Jul 16. 2004.

[9] J.D. Thompson, D.G. Higgins and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.,* vol. 22, pp. 4673-4680, (1994).

[10] S.R. Eddy, “Profile hidden Markov Models,” *Bioinformatics,* 14, 755-763 (1998)

[11] N. Kelly, P.V. Jithesh, D.R. Simpson *et al*, “Bioinformatics Data and the Grid: The GeneGrid Data Manager”, *Proceedings of UK e-Science All Hands Meeting (2004),* 571-578

[12] R. Apweiler, *et al*, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.,* 32, D115-9, 2004.

[13] C. Kanz, P. Aldebert, N. Althorpe *et al*, "The EMBL Nucleotide Sequence Database," *Nucleic Acids Res.,* vol. 33 Database Issue, pp. D29-33, Jan 1. 2005.

[14] J. Novotny, M. Russell, O. Wehrens, “GridSphere: An Advanced Portal Framework”, *Proceedings of EuroMicro Conference (2004), 412-419*

[15] P.R. Crocker Siglecs: sialic acid binding immunoglobulin-like lectins in cell-cell interactions and signaling. *Curr. Opin. Struct. Biol.*, 12, 609-615 (2002).