GeneGrid: Grid Based Solution for Bioinformatics Application Integration and Experiment Execution

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

Belfast e-Science Centre, Queen's University of Belfast

{ p.jithesh, n.kelly, 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.

1. Introduction

The improvement in genome sequencing and post-genomic technologies such as microarrays has led to the generation of vast amount of biological data. The requirements of storage and the analysis of such large volumes of data have pushed 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, viz., Fusion Antibodies Ltd. and Amtec Medical Ltd., interested in antibody and drug development. The aim 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. GeneGrid is built upon the state-of-the-art technology in distributed computing, namely grid computing, which is concerned with the coordinated resource sharing and problem solving in dynamic multi-institutional virtual organisations [2]. GeneGrid 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 how the grid services components 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, involving the identification of novel members of a protein family, for demonstrating the capabilities of GeneGrid, is also presented. Identification and characterisation of novel proteins encoded by the genomes may either help in isolating the role of these proteins in disease conditions or provide potential targets for the development of new drugs. However, the computational procedures required for such discovery are quite elaborate and complex involving a number of bioinformatics programs as well as computing and data resources. The ability of GeneGrid to circumvent these problems was tested using the case of a protein family called Sialic acid-binding Immunoglobulin-like lectins (Siglecs).

2. GeneGrid Architecture

GeneGrid consists of a number of cooperating Grid services developed based on the Open Grid Services Architecture (OGSA) [3]. Grid services provide a standardised way of programmatic access to resources, data and applications using XML based messages. 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

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

The OGSA-compliant 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. A single GAMSF on a resource can generally interface all the applications available on that resource. The primary function of GAMSF is to create instances of itself called GeneGrid Application Manager Services (GAMS) which facilitate clients to interface with the applications.

GAMS is transient in nature unlike the parent GAMSF. Any client wishing to execute a supported application will first connect to the GAMSF and create an instance - the GAMS. Upon creation, the GAMS inherits configuration from the GAMSF which inform it as to how to access and utilise the required application. 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]. 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 data private to the companies. 2). GeneGrid data consisting of data either required by, or created by GeneGrid, such as results information or workflow definitions.

GDM has adapted OGSA-DAI(http://www.ogsadai.org) as the basis of its framework, enhancing and adapting it as required. GDM consists of two types of services, replicating those found in OGSA-DAI. The GeneGrid Data Manager Service Factory (GDMSF) is a persistent OGSA-compliant 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).

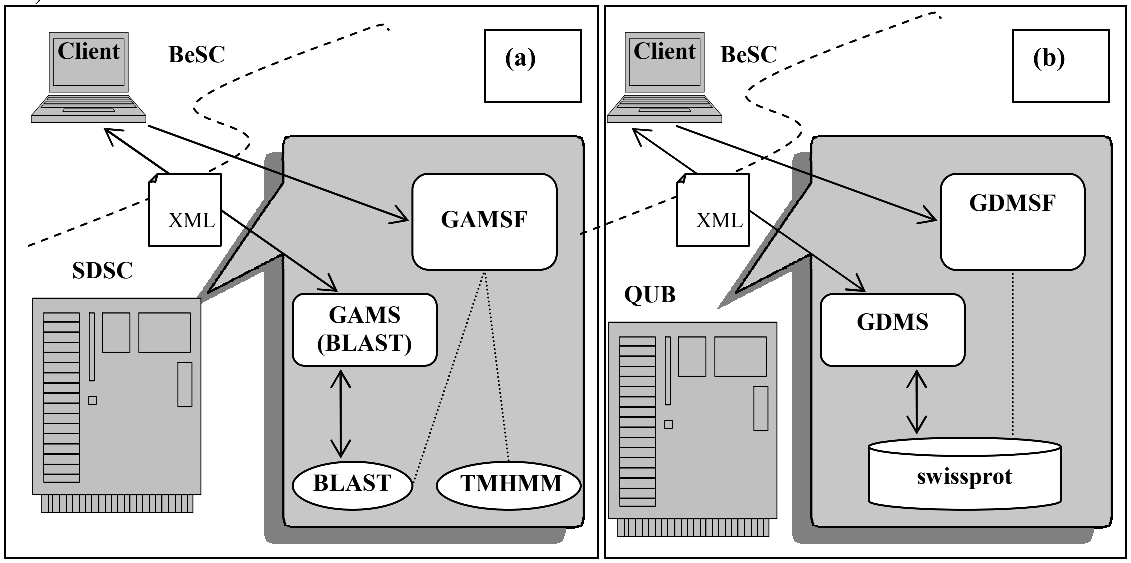


Figure 1. A client accessing (a) an application, e.g; BLAST on another resource through GAMS and (b) a database e.g: Swissprot through GDMS

2.3. Workflow Management & Service Discovery

The 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-compliant 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).

2.3. 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.

2.4. 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.

3. GeneGrid operation and a use case: identification of novel protein family members

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.

3.1. Design of workflow

Initially a workflow for the experiment was designed by the biologists and bioinformatician capturing the requirement for identifying new siglecs, which is discussed below. The experiment starts with protein sequences of known members of the family and finding homologues in various databases using BLAST [6]. The next phase eliminates those sequences which did not contain characteristic transmembrane and signal regions using TMHMM [7] and SignalP [8] respectively.

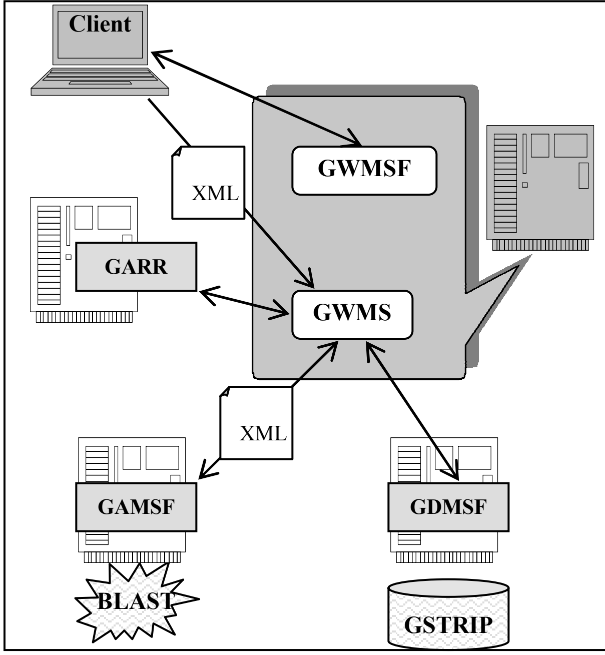
3.2. Creation of workflow

User interaction with GeneGrid is facilitated by the portal, which also masks the underlying complexity of the grid from the user. Each application within the GeneGrid system is described in detail in an XML file called the Master Workflow Definition (MWD) present in a central repository, the GeneGrid Workflow Definition Database (GWDD). This MWD defines all relevant details for each application including details such as input and output and various other parameters. When a user creates a workflow that consists of a series of tasks, the portal creates a web form containing the correct input fields for each task based on the definition for that application within the MWD. Once the user has supplied all the appropriate input parameters for all the tasks within the workflow, a single XML file is created that describes the scientist's full experiment.

3.3. Execution of workflow

This step of the procedure is entirely accomplished by the GeneGrid system, without any user intervention. A workflow XML file created in the previous step is then processed by the GeneGrid Workflow Manager (GWM) service, which in turn identifies and extracts the tasks required to complete the requested workflow.

After identifying the tasks, the GWM service contacts the GARR to get the information on the services that provided the required application programs and databases. The tasks are then passed on to the appropriate service on the specific resources for execution in parallel or sequentially based on the interdependency of tasks. Once the GAM service received a task in the form of an XML file from the GWM, it creates the actual command to be executed on the resource using the parameters from the task XML and a local XML configuration definition specific to the resource. GAM accesses the input sequence from the GSTRIP database and once the task was executed, this database was updated with the output files.



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)

As is the case with many workflows, the above workflow also needed some additional programs to link the output of one task with the input of the next task. For example, after executing the BLAST task it was required to extract the accession numbers of all the hits to use in the next step to retrieve the FASTA format sequences from a database. Such linking was accomplished by developing a suite of linker grid services exposing routines implemented in Perl and BioPerl (http://bioperl.org). GeneGrid automatically tries to add such linking tasks into the workflow based on certain rules and conditions.

3.4. Examining the results

Once the user initiates a workflow, it is possible to get the status at any time. GWM frequently updates the status of each task in a workflow which is stored in the GSTRIP database. A facility on the portal also allows the user to view or download the intermediate input parameters/files and output files as they become available.

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.

4. Discussion and conclusion

Using GeneGrid for *in silico* experiments provides distinctive advantages over the conventional methods as described below.

GeneGrid integrates numerous bioinformatics programs and databases available on different resources across various sites allowing the scientists to easily access the diverse applications and data sources without bothering to visit many web servers. This reduces the overall time for execution of the experiment. As the system takes care of monitoring and selection of appropriate resource for the tasks requested, it relieves the user of such selections and more importantly, utilises the resources in an efficient way. This not only reduces the overall time of the experiment, but the individual tasks are also sometimes accelerated. Errors which may creep in as a result of manual intervention are avoided by automation.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.

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.

5. 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, S Tuecke, “The Anatomy of the Grid: Enabling Scalable Virtual Organisations”, *International J,* Supercomputer Applications (2003), 15(3)

[3] 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)*

[4] 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

[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).