

# Bioinformatics Application Integration in GeneGrid

**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  
Townasley, Jim Johnston

Fusion Antibodies Ltd,  
Belfast

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

Shane McKee

Amtec Medical Ltd,  
Belfast

shanemckee  
@doctors.org.uk

## Abstract

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 the improvements and modifications made to the GeneGrid Application Manager (GAM) since its last release. GAM is the Globus Toolkit 3 based grid service responsible for the integration of Bioinformatics applications and other accessory programs present on heterogeneous resources, within the GeneGrid environment. A major thrust was given to make its functionality as extensible as possible by making it highly generic. This has helped in the easy and seamless integration of new applications that are heterogeneous in their requirements and outputs, making it possible to perform a number of real biological workflows.

## 1. Introduction

GeneGrid is a UK e-Science industrial project with the involvement of companies, viz., Fusion Antibodies Ltd. and Amtec Medical Ltd., interested in the development of antibodies and drugs. Its 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 allows biologists to create, execute and manage workflows that represent bioinformatics experiments.

This paper presents the improvements and modifications made to the GeneGrid Application Manager (GAM) since its last release [2]. GAM is the Globus Toolkit 3 based grid service responsible for the integration of Bioinformatics applications and other accessory programs present on heterogeneous resources, within the GeneGrid environment. A major thrust was given to make its functionality as extensible as possible by making it highly generic. This has

helped in the easy and seamless integration of new applications that are heterogeneous in their requirements and outputs, making it possible to execute a number of real biological workflows.

## 2. GeneGrid Architecture

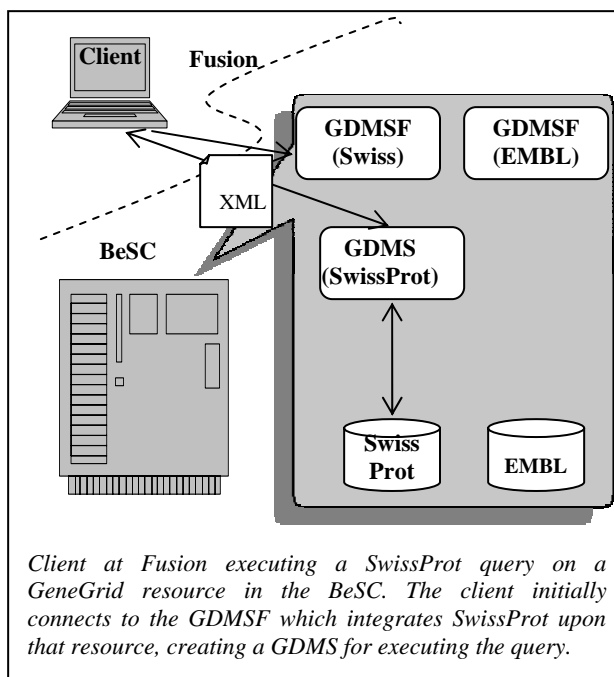
GeneGrid consists of a number of cooperating Grid services developed based on the Open Grid Services Architecture (OGSA) [3]. GeneGrid services may be categorised logically into different components, namely Workflow Management, Resource Monitoring & Service Discovery, Data Management, Application Management and the Portal. This paper focuses on the Application Management component, which will be described in detail in Section 3. All the other components are discussed briefly below.

### 2.1. 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 [4]. The data integrated by the GDM falls into two categories. 1). Biological data consisting of datasets available in the public domain, e.g. Swissprot [5], EMBL [6] 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.

OGSA-DAI (<http://www.ogsadai.org>) was used as the basis of GDM, 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 1).



**Figure 1. A client accessing a database e.g: Swissprot through GDMS**

### 2.2. Workflow Management

The GeneGrid Workflow Manager (GWM) is responsible for the processing of all submitted workflows within GeneGrid (Figure 3). 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 the central access point for all users to GeneGrid and is based upon the GridSphere product [7]. 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.

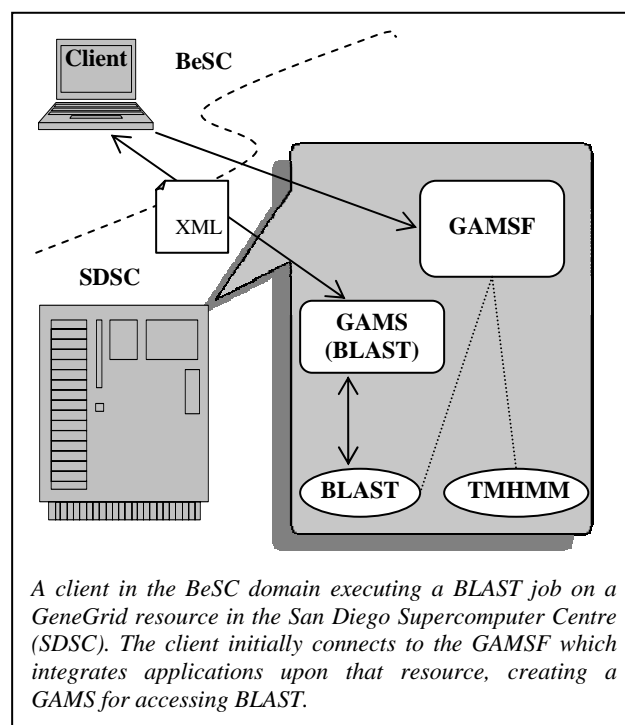
## 3. Application Integration and Management

### 3.1 Services

GeneGrid Application Manager (GAM) provides programmatic access to the bioinformatics and other accessory applications available on various resources [2]. GAM achieves this integration through two types of grid services: GeneGrid Application Manager Service Factory (GAMSF) and the GeneGrid Application Manager Service (GAMS).

The GAMSF is a persistent service, which extends the standard portTypes, like the GridServiceFactory of the Open Grid Services Infrastructure (OGSI) [8] 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.



**Figure 2. A client accessing an application e.g:BLAST through GAMS**

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. Hence GAMS differs from its parent GAMSF in being transient in nature (Figure 2).

As mentioned in the previous section, GARR is the central service in GeneGrid responsible for the resource and service publishing. The Node Monitors present on any resource with GAM also sends the GSH of the GAMSF to the GARR. In addition, it also transmits the information about the applications integrated by the GAM on the particular resource. This helps in the discovery of appropriate GAMSF by a client wishing to execute a specific application.

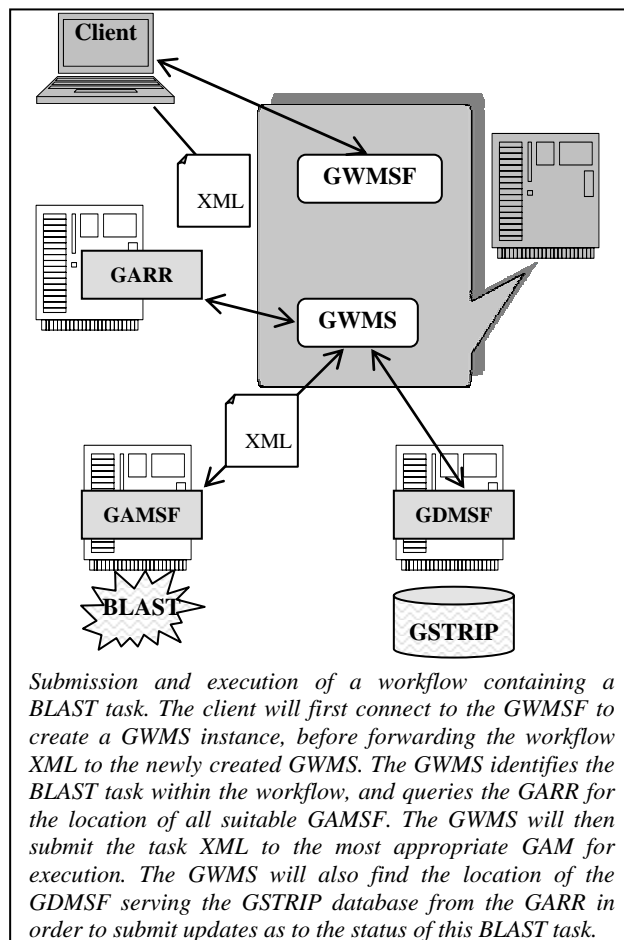
### 3.2 Interaction with other GeneGrid Components

The GeneGrid Portal displays the details of applications available to the user by reading a Master Workflow Definition file. Information about each application, such as the input and output requirement as well as other parameters are present in this XML file. Once the user creates an experiment through the portal by entering the values of the parameters, the workflow is submitted to a GeneGrid Workflow Manager service (GWMS). The GWMS is responsible for identifying the tasks involved in the workflow and finding the appropriate resources for the task execution. Hence the GWMS usually acts as a client to the GAMSF (Figure 3).

After acquiring the GSH of the appropriate GAMSF from GARR, the GWMS directly contacts and requests an instance of the GAMSF. The factory creates a GAMS instance and returns the handle to the invoking service. The GWMS then contacts the GAMS instance with a properly formatted task XML file that has all the parameter values entered by the user for the specific task. GAMS retrieves any input file required to execute the task from the GSTRIP database and stores locally. Based on the task XML and a configuration XML present on the resource, which contains the resource specific details of the application, GAMS creates the command for execution of the application.

The command is then executed on the resource and if successful, GAMS reads the

output files and updates the GSTRIP database. In addition, GAMS also updates some metadata about the files which provide provenance information. As mentioned earlier the interaction with databases are mediated by the GeneGrid Data Manager services.



*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 3. Workflow Management in GeneGrid

The GAMS also notifies the successful or unsuccessful completion of tasks to the GWMS instance which requested the task execution. If the task execution was successful, a pointer to the output data stored in GSTRIP is also forwarded to GWMS, which can then be displayed in through the Portal.

### 3.3. Applications and Accessory Programs

GeneGrid integrates a number of popular bioinformatics applications commonly used by

our user community. The applications were selected based on the requirements from the scientists, provided through use cases. Currently GeneGrid integrates the following Bioinformatics applications:

- Basic Local Alignment Search Tool (BLAST) [9]
- TMHMM - Transmembrane prediction program [10]
- SignalP – Signal Prediction program [11]
- ClustalW – Multiple sequence alignment program [12]
- HMMER – Profile HMM creation and database search tools [13]
- Many of the programs from the EMBOSS suite (<http://www.emboss.org>)

Integration of new applications to GeneGrid is an easy and straightforward procedure and does not require any modifications to the codes. GAM takes care of the heterogeneous nature of the applications.

In addition to the Bioinformatics applications, GAM also integrates a number of accessory programs which help in the linking of tasks in a workflow. As Bioinformatics applications are highly heterogeneous in their input and parameter requirements, it was necessary to develop such programs. Such utility programs were developed particularly following the use cases provided by the scientists.

#### 4. Use Case: Automated Antigenic Region Detection

One of the routine Bioinformatics analyses carried out by scientists from the partner company is finding the regions with antigenic properties in proteins starting from the gene sequences. The procedure involves a number of Bioinformatics programs which are to be accessed from different sources such as publicly available web servers and those present on resources internally. Manual execution of such experiments are tedious, time-consuming and

error prone, especially considering the volume of analyses need to be carried out in a day.

GeneGrid has provided a solution to automate the experiment through reusable workflows and substantially reduced the time for execution by distributing the tasks across optimal resources.

The workflow for automated antigenic region detection involves a number of steps (Figure 4):

1. Gene sequences are translated into the corresponding protein sequences
2. Various characteristics of the protein such as presence of transmembrane regions, signal regions etc are examined.
3. Regions of the proteins are extracted with the desired characteristics and those regions which do not contribute to antigenicity are removed.
4. The sequence fragments searched against a database such as Swissprot to eliminate the chance of having highly homologous sequences
5. The unique fragments selected are further analysed to see whether they can have primers for polymerase chain reaction.

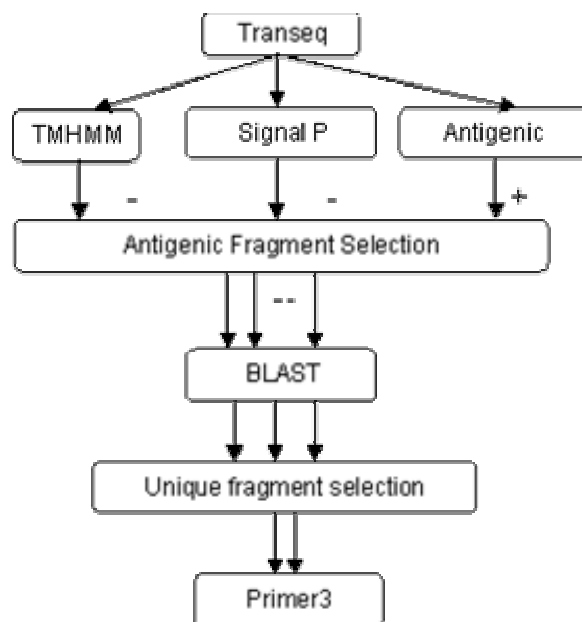


Figure 4. Outline of the use case. See text for details.

Manual execution of the experiment took about 30-60 minutes per gene for analysis. In contrast, the automated antigenic region detection using GeneGrid took about 90 minutes for the analysis of 100 genes. This acceleration is largely due to the automation and parallelization of task execution, as well as the optimal use of available resources.

## 5. 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 GeneGrid Application Manager, which integrates heterogeneous Bioinformatics applications present on disparate resources, has been modified since its last release to make it more flexible and robust. This has helped the integration of new applications and utility programs easy. More use cases were tested successfully with promising performance figures.

## 6. 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] 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.
- [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] 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
- [5] R. Apweiler, *et al*, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.*, 32, D115-9, 2004.
- [6] 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.
- [7] J. Novotny, M. Russell, O. Wehrens, "GridSphere: An Advanced Portal Framework", *Proceedings of EuroMicro Conference (2004)*, 412-419
- [8] S. Tuecke, K. Czajkowski, I. Foster *et al.*, Open Grid Services Infrastructure (OGSI) Version 1.0. *Global Grid Forum Draft Recommendation*, (6/27/2003).
- [9] 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.
- [10] Crogh *et al*, "Predicting transmembrane topology," *J.Mol.Biol.*, vol. 305, pp. 567-580, Jan. 2001.
- [11] 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.
- [12] 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).
- [13] S.R. Eddy, "Profile hidden Markov Models," *Bioinformatics*, 14, 755-763 (1998)