

# The GeneGrid Portal: A User Interface for a Virtual Bioinformatics Laboratory

Noel Kelly,  
P.V. Jithesh, Sachin Wasnik, Roy  
McLaughlin, Fabiola Fragoso, Paul  
Donachy, Terence J. Harmer, Ron H.  
Perrott

*Queen's University of Belfast*

*{n.kelly,p.jithesh,s.wasnik,m1590102,  
p.donachy,t.harmer,r.perrott}@qub.a  
c.uk, fabiolafragoso@gmail.com*

Mark McCurley,  
Michael Townsley,  
Jim Johnston

*Fusion Antibodies Ltd*

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

Shane McKee

*Amtec Medical Ltd,*

*shanemckee@doctors.org.uk*

## Abstract

*GeneGrid is a collaborative industrial grid computing R&D project initiated by the Belfast e-Science Centre (BeSC) under the UK e-Science programme and supported by the Department of Trade & Industry. The project aims to provide a platform for scientists to access their collective resources, skills, experiences and results in a secure, reliable and scalable manner by creating a "Virtual Bioinformatics Laboratory".*

*GeneGrid provides seamless integration of a myriad of heterogeneous applications and datasets that span multiple administrative domains and locations across the globe, and presents these to scientist through a simple user friendly interface.*

*Despite bringing substantial benefits to our commercial partners, the learning curve required to prepare biologists to use Grid Technology presents a major deterrent. A major challenge during GeneGrid development is to expose the full computational and collaborative potential of the Grid, while keeping our users' need for knowledge of the underlying technology to a minimum.*

*GeneGrid meets this challenge through the GeneGrid Portal – the user interface for the project. In this paper we present an overview of GeneGrid, concentrating on the role of the GeneGrid Portal. We also discuss functionality provided by the GeneGrid Portal, along with a snapshot of the future plans for the project.*

## 1. Introduction

The growth of bioinformatics has stemmed from the avalanche of data generated in biological fields thanks to the development and sophistication of technologies that were unimaginable just a few years ago. The major breakthrough in this field came from the genome sequencing projects such as the human genome project, as well as equivalent projects for other animal, plant and bacterial genomes (see <http://www.genomesonline.org>). Post-genomic technologies such as microarrays, which help in the genome wide analysis of the expression of thousands of genes in a single assay, are creating an explosion in the number of datasets to be managed and integrated. To make sense of all this data, the bioinformatics community has come up with a variety of software systems to mine and analyse it. However, the variety and complexity of data formats and the heterogeneity of the requirements and outputs of individual programs have complicated their use in bioinformatics experiments. A typical experiment may

require following a specific sequence of steps to integrate data from a variety of public and private resources. Moreover, the actual time taken for such experiments often far surpasses the individual component task execution times since such procedures require an expert Bioinformatician to access the resources and data at remote locations, or via web servers requiring much manual intervention. Often the calculations are so computationally intensive that they require high performance computing to accomplish the task in reasonable time limits, leaving the biologists with moderate compute facilities at a loss.

GeneGrid [3] is a novel and pragmatic solution to address the above problems. It accomplishes the seamless integration of a myriad of heterogeneous resources that span multiple administrative domains and locations, providing the scientist with an integrated environment for collaborative discovery and development via a user friendly interface. GeneGrid is built upon state-of-the-art technology in distributed computing, namely grid computing, which coordinates resource sharing and

problem solving in dynamic multi-institutional virtual organisations [1].

An industrial UK e-Science project, GeneGrid involves companies interested in antibody and drug development, and aims 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’ [3].

Being a workflow driven project, biologists using GeneGrid may create, execute and manage bioinformatics experiments which are represented in GeneGrid as XML workflow documents. These workflows automate, and hence accelerate, experiments, while providing relief from errors that usually creep in due to manual interventions.

## 2. GeneGrid Component Architecture

GeneGrid consists of five cooperating components which independently address a subset of the main requirements of the project, and by cooperating provide scientists with an integrated environment for the streamlined access of a number of bioinformatics applications & databases through a simple interface. These five components, namely the GeneGrid Data Manager, the GeneGrid Application Manager, the GeneGrid Workflow Manager, the GeneGrid Service Discovery Component and the GeneGrid Portal will be discussed individually in more detail below.

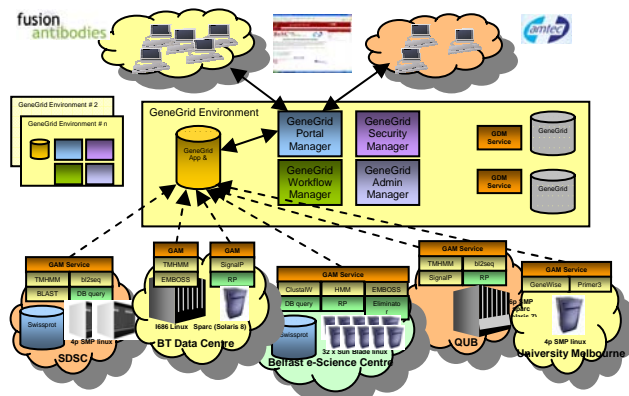


Figure 1. GeneGrid Architecture Overview

### 2.1 GeneGrid Application Manager (GAM)

The GAM is tasked primarily with the access and integration of bioinformatics applications such as BLAST [7] or ClustalW [8]

The GAM consists of two types of distributed OGSA compliant grid services [2] - GeneGrid Application Manager Service Factories (GAMSF) and the instances which they create, called GeneGrid Application Manager Services (GAMS).

Each Factory is a persistent OGSA compliant Grid Service [2] which implements the standard OGSA

Factory interface in order to integrate one or more bioinformatics applications to the Grid, and expose them to the rest of the GeneGrid architecture. Each application integrated into GeneGrid must have at least one Factory configured to support it, with the standard practice being that one Factory is available on any node offering bioinformatics applications to GeneGrid. Its primary function is to create instances which facilitate clients to interface with the applications.

Factory instances, or GAMS, like the factory which creates them, are OGSA compliant service, although unlike their parent, they are transient in nature. Any client wishing to execute a supported application will first create an instance which inherits configuration from its parent factory, informing it as to how to access and utilise the required application. This newly created GAMS then expose the available operations to the client which allow the client to execute the supported application as an extension to the operations provided by the OGSA Grid Service interface.

Within GeneGrid, in order to execute the application, the client will submit an XML document to the GAMS document containing a set of execution parameters. This document is then processed by the GAMS for the required application, and a set of resulting output files are generated. These files are then uploaded by the GAMS for persistent storage.

While the GAM currently supports a finite set of bioinformatics applications, the solution is versatile and scalable, ensuring further applications may be supported with relative ease.

### 2.2 GeneGrid Data Manager (GDM)

The GDM is a stand alone component within GeneGrid which consists of a number of distributed services based upon the OGSA-DAI product. It is responsible for integrating data from the public biological domain such as EMBL or SwissProt, as well as both GeneGrid specific data and data proprietary to the commercial partners.

The GDM contains two types of services, the GeneGrid Data Manager Service Factory (GDMSF) and the GeneGrid Data Manager Service (GDMS). These services mirror the Grid Data Service Factory and the Grid Data Service respectively described at [www.ogsadai.org.uk](http://www.ogsadai.org.uk).

### 2.3 GeneGrid Workflow Manager (GWM)

GeneGrid is a workflow driven project, with all experiments created by the scientists described in an XML workflow document. This workflow contains a set of bioinformatics jobs, or tasks, as well as all required execution parameters as selected by the scientist. The GeneGrid Workflow Manager – the central component of the GeneGrid architecture – is responsible for the

execution of this workflow, and contains 2 types of service.

The GeneGrid Workflow Manager Service Factory (GWMSF) is a persistent OGSA compliant Grid Service which implements the standard OGSA Factory interface. Its primary function is to create instances of itself called GeneGrid Workflow Manager Services (GWMS) which facilitate GeneGrid workflow execution.

The created instances are transient OGSA compliant Grid services which implement and extend the standard OGSA Grid Service interface to facilitate the management of workflows within GeneGrid, with each instance created to manage a single workflow.

The GWMS receives a workflow from the client in the form of an XML document. This XML document is then partially parsed by the GWMS to identify the individual tasks present in the workflow, and to identify any inter-dependence between these tasks. The GWMS is then responsible for the selection of the appropriate resource on which to execute the individual tasks. Dependencies between tasks are resolved by the GWMS as appropriate.

## 2.4 GeneGrid Service Discovery Component

The GeneGrid Service Discovery Component contains a single service, as well as a collection of lightweight applications used to monitor resource information.

The GeneGrid Application & Resource Registry (GARR) is a persistent OGSA compliant Grid Service which facilitates service discovery within the project. All computational nodes available to GeneGrid run a lightweight application which collects resource information relating to the GeneGrid node on which it is deployed. These applications, called GeneGrid Node Monitors (GNM), transmit the information they have collected after a configurable time period to the GARR for persistent storage.

Each Monitor may also be configured to advertise details on the Grid Services deployed on the node on which they execute, forwarding information such as the service name, type and location, as well as any databases or applications they may facilitate access to. The GARR will then store this information within a MySQL database to provide persistence of resource information.

Any client wishing to discover the location of any specific integrated application or database may query the GARR for its location. It is also capable of returning resource information associated with these node(s) hosting the requested resource.

## 2.5 GeneGrid Portal

The GeneGrid Portal, based upon the GridSphere product [6], provides a secure central access point for all users and scientists to GeneGrid. 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 are already familiar with. This results in a drastically reduced learning curve for the scientists in order to exploit grid technology.

We will discuss the GeneGrid Portal in much more detail in section 4.

## 3. GeneGrid Component Integration

Section 2 presented each GeneGrid component individually. In this section we will discuss how these components cooperate and interact to form a Grid Service Oriented Virtual Bioinformatics Laboratory.

### 3.1 GeneGrid Environment

The 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 a GeneGrid Portal, a GARR, a GWMSF, an implementation of each of the GeneGrid databases, namely the GSTRIP and the Workflow Definition, 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 GeneGrid Environment.

### 3.2 GeneGrid Shared Resources

Bioinformatics applications and datasets are exposed to the GeneGrid Environment by the GAM and GDM respectively. These GAM and GDM services make up the GeneGrid Shared Resources. The existence, capabilities and location are advertised a GeneGrid Environment via Node Monitors (see 2.4) on their hosting nodes registering with the GARR.

It is possible for GNM to register with many GARR services across multiple GeneGrid Environments 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. The GeneGrid Portal

In order to bring Grid computing to the masses, the learning curve required to utilise and run grid applications must be drastically reduced. Currently, a considerable amount of knowledge on the underlying IT technologies is needed to run the simplest of tasks. Making the Grid more usable is a challenge facing all Grid application developers and projects, and in this matter, GeneGrid is no exception.

GeneGrid users primarily come from a biological laboratory environment, and their experience of IT

technology is varying. To make GeneGrid a success, we had to make the Grid as easy to use as possible, while still exploiting the functionality fully. From reviewing our user community requirements, we quickly realised that our users currently use a lot of the applications and databases available through World Wide Web forms. Therefore, in order to keep the transition as easy as possible, we decided to adapt a similar approach for GeneGrid. The use of a Portal then became an obvious candidate due to the built in functionality such as security and user management, as well as allowing us to have a centrally hosted, and hence centrally administered, user interface.



Figure 2. GeneGrid Portal Welcome Page

As the user interface for the GeneGrid Virtual Bioinformatics Laboratory, the GeneGrid Portal provides a secure central access point for all registered users to create, manage and track experiments, while also allowing them to view generated results. Administrators of GeneGrid may also use the Portal to configure and monitor the Virtual Bioinformatics Laboratory. In all there are currently 7 custom developed portlets hosted by a GridSphere Portal [6], but with all 7 developed to be JSR 168 compliant [9], they may all be ported to any JSR 168 compliant Portal Framework.

#### 4.1 User Portlets

We currently have 4 JSR 168 compliant portlets developed for use by all GeneGrid users, allowing them to execute experiments, building them from scratch, or re-using existing templates. There's also a portlet allowing users to review all their submitted experiments, allowing them to check on the current status of the experiment as well as view both input files, and generated results from the experiment. A more detailed description of each follows.

##### 4.1.1 "New Experiment" Portlet

The "New Experiment" Portlet allows users to build new experiments from all available tasks and utilities supported by GeneGrid. It adopts a wizard based approach which guides users through the stages of setting up an experiment, prompting them on what actions to take at any given time. As they progress through the wizard stages, there are opportunities to review the work to date, along with any specific details added. The options presented to the user are driven by an XML file called the "Master Workflow Definition Document" which is read up by this portlet during initialisation. This document contains advanced information on each bioinformatics application supported by GeneGrid.

When a user is content with the experiment they have compiled, they may submit it to the GeneGrid Workflow Manager for execution with this portlet assigning a unique id to the experiment for tracking purposes. Users may also attach their own description to the experiment to make tracking easier.

An import feature of this portlet is it also allows users to save the experiments they have just created as templates for use again later with different inputs.

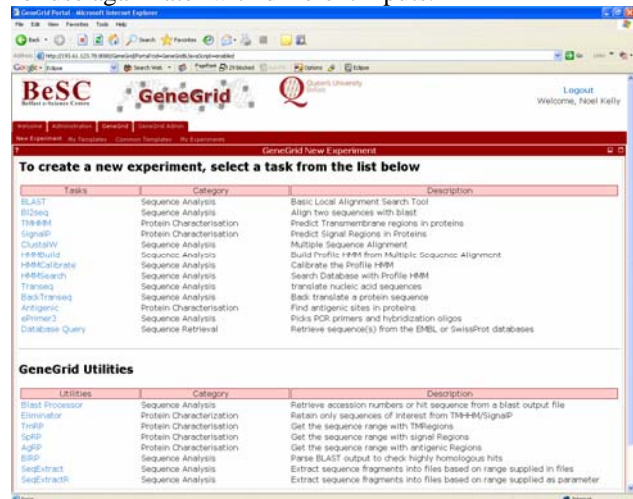


Figure 3. Initial "New Experiment" Portlet Page

##### 4.1.2 "My Templates" Portlet

In an effort to make GeneGrid as user friendly as possible, we have introduced the concept of experiment templates. This means that instead of having to rebuild an experiment over and over for different input values, users can save an experiment as a template during creation (in "New Experiment"), and run it repeatedly for different input values. Like the "New Experiments" portlet, this portlet adopts a wizard based approach which guides users through the stages of submitting the experiment. However, with much of the compilation work completed previously in creating the experiment, the user is only requested to fill in details which are still outstanding, with parameter values copied from the template.

As a security measure, when users save templates, their user name is attached. Through this portlet, the logged in user can see only the templates which they have saved, and may not see those saved by others. This is useful when you consider shared GeneGrid Environments in a commercial setting, where a user may not wish others to know anything about the work he/she is currently performing.

### 4.1.3 “Common Templates” Portlet

This portlet is identical to the “My Template” portlet described in 4.1.2, with the exception that the templates here are shared among all users of the GeneGrid Environment.

This feature alone has led to a greater uptake in the use of GeneGrid amongst our potential user community as users may work with the benefits of GeneGrid, while not having to get into the specifics of creating experiments themselves from the very beginning.

In order to keep the common templates available as concise and user friendly as possible, the ability to add common templates is restricted to GeneGrid Administrators.

### 4.1.4 “My Experiments” Portlet

In order to make the most of the GeneGrid Virtual Bioinformatics, users must be able to track the experiments they have submitted to the system for execution, and view any generated result files. This functionality is provided by the “My Experiments” portlet.

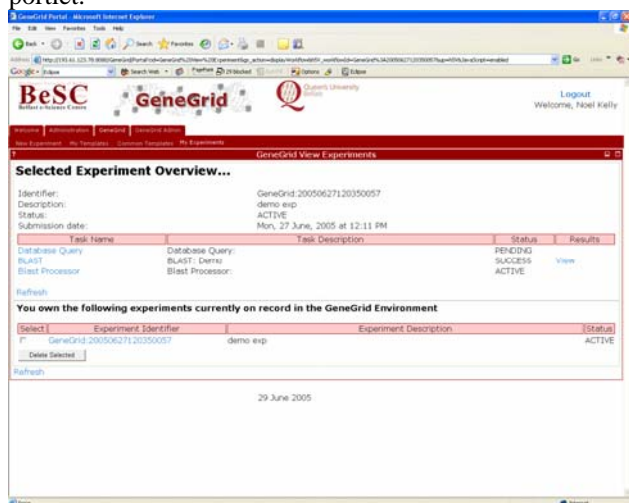


Figure 4. “My Experiments” Portlet

Upon initialisation, the portlet will retrieve a list of experiments owned by the logged in user from the GSTRIP database. The user may then browse through the list, and select to one they want to view. Further information is available from here, including an overview of the tasks in the experiment, as well as the status and error code if applicable. Users can also view individual

task overviews, which will include links to view the input and generated output if they are available.

Through this portlet, the user can examine the results generated by each step of the experiment, and also view how what inputs generated what results.

The user may also delete past experiments from this portlet.

## 4.2 Administration Portlets

As well as providing a suite of functions to all GeneGrid users, the Portal provides an excellent opportunity to allow configuration of the Virtual Bioinformatics Laboratory through a simple central secure access point. Currently, there are 3 administration portlets available.

### 4.2.1 “Resource Configuration” Portlet

With Service and Resource discovery enabled through the GARR, it is important that the GeneGrid Portal is configured with the location of the GARR service. This is possible using the “Resource Configuration” portlet, which will store the GSH of the GARR in the persistent Portal database. The Administrator can simply change this address through a form, and update it for the GARR.

This portlet will also provide additional functionality, connecting to the configured GARR and retrieving a list of all services registered. The Administrator may then add or remove services to the GARR as appropriate.

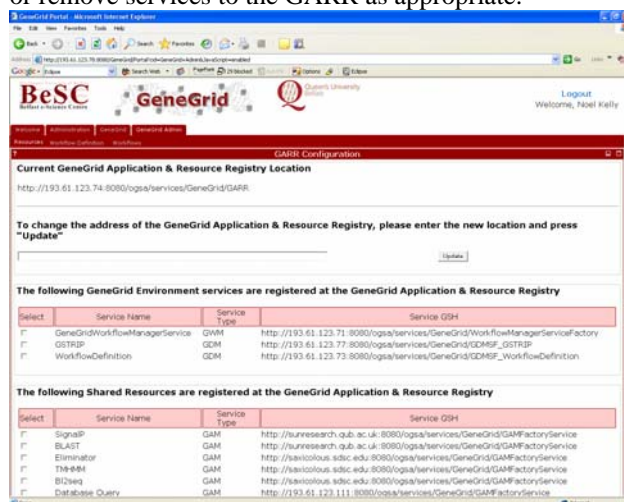


Figure 5. “Resource Configuration” Portlet

### 4.2.2 “Master Workflow Definition Configuration” Portlet

As mentioned in 4.1.1, the GeneGrid Portal reads up a document called the Master Workflow Definition Document. With changes to this file being made regularly, it is important that the Portal is able to read the latest version.

Administrators may upload the most recent Master Workflow Definition Document to the GE using this portlet. A resource name is applied to the document which

is then stored in the Workflow Definition Database, while the resource name used is saved to the Portal's persistent database.

#### 4.2.3 "Workflow" Portlet

Every time an experiment is submitted to GeneGrid for execution, a record is created in the GSTRIP database. From time to time, administrators may be required to delete records from the GSTRIP which can be quite a complex task given that each record traverses multiple tables.

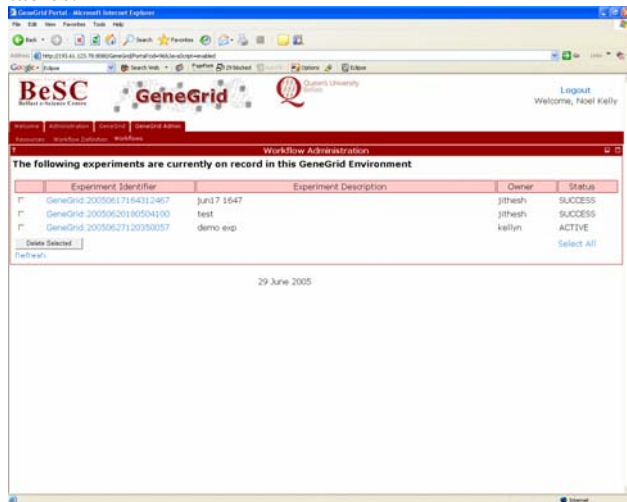


Figure 6. "Workflow" Portlet

Using the "Workflow" Portlet, the administrator can view all experiments currently recorded in the GSTRIP, and may choose to view any he / she may please. The Administrator may also use this portlet to delete any records which are no longer required.

### 5. Future Plans

GeneGrid development is still on going with improvements and additional functionality being added to all components of the project - the Portal being no exception.

With the release of version 0.5, we have users from both of our commercial partners using the Virtual Bioinformatics Laboratory on a daily basis. These users are constantly making suggestions and requesting changes which we try to incorporate.

Further to this, we have plans to integrate much more functionality into the GeneGrid Portal.

We are currently investigating using more intuitive and user-friendly graphical interfaces within our User Portlets, particularly with respect to the creation of new experiments.

There is also a considerably amount of work planned in the Administration portlets to try and make GeneGrid a more centrally configurable entity. To this end, we are looking at extending the "Master Workflow Definition

Configuration" portlet to allow administrators to view and edit the document though the portal.

We are also looking into developing a configuration option for all GeneGrid services and GeneGrid Node Monitors, allowing us to configure these entities centrally from the Portal.

Finally, the GARR (see 2.4) is based upon the Grid Manager project developed in the Belfast e-Science Centre. Grid Manager has its own versatile and intuitive user interface which we intend to migrate to the GeneGrid Portal for use within GeneGrid.

### 6. Conclusions

From talking with many potential users, the complexity and difficulty of using Grid technology is still outweighing the benefits which the Grid offers. Therefore, the challenge is there to all of us Grid Application developers to develop applications which are both easy to use, and still exploit the functionality of the Grid.

Furthermore, we in GeneGrid believe that the users' expectations and current practices should be considered when developing user interfaces for Grid Applications. Our potential users are very familiar with using web pages to exploit applications and databases, and by keeping a web based approach to our user interface, we've had an increase in the interest from our user community.

## References

- [1] I. Foster, C. Kesselman, S Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organisations", *International J, Supercomputer Applications* (2003), 15(3)
- [2] I. Foster, C. Kesselman, J. Nick, S. Tuecke, "The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration", *Open Grid Service Infrastructure WG, Global Grid Forum (June 22<sup>nd</sup>, 2002)*
- [3] 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
- [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] 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
- [6] J. Novotny, M. Russell, O. Wehrens, "GridSphere: An Advanced Portal Framework", *Proceedings of EuroMicro Conference (2004)*, 412-419
- [7] 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.
- [8] 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, Nov 11. 1994.
- [9] JSR 168 Specification
- [10] N.Kelly *et al*, "GeneGrid: A Commercial Grid Service Oriented Virtual Bioinformatics Laboratory", *Proceedings of SCC 2005*