



### Bioinformatics and e-Science

The volume and complexity of biological information in existence today necessitates computational data analysis. The emergence of high throughput technology in genomics and proteomics ensures that this complexity and volume will increase rapidly in the future. New approaches to the storage and computational analysis of biological data are required. These challenges are being tackled by the myGrid project, which will provide an e-Science laboratory for the life sciences by exploiting Grid Computing.

myGrid ([www.mygrid.org.uk](http://www.mygrid.org.uk)) is a UK e-Science pilot project funded by the Engineering and Physical Sciences Research Council. It is a collaboration between the Universities of Manchester, Newcastle, Nottingham, Sheffield and Southampton, the European Bioinformatics Institute (EBI) and IT Innovation. Industrial partners on the project include AstraZeneca Pharmaceuticals, GlaxoSmithKline, Merck, IBM and Sun Microsystems.

### myGrid middleware

The objective of myGrid is to develop an open source, service-orientated middleware framework for bioinformatics. Prototype version 1 of the myGrid middleware has been designed and implemented (Fig. 1).

The framework comprises core and support e-Science services. The core services are the myGrid Information Repository (MIR); the myGrid workflow

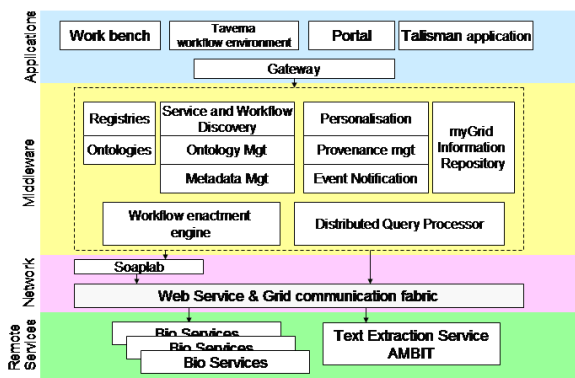


Figure 1. The myGrid stack showing the services developed by myGrid

enactment engine, which allows scientists to execute workflow processes; and the distributed query processor that integrates data stored in separate, databases.

e-Science support services aid the scientist in constructing and recording *in silico* experiments. There are services to discover bioinformatics data sources, applications and workflows, to manage provenance information and to notify scientists of changes to data or applications used previously in their experiments.

A number of bioinformatics databases and tools have been made available as web services by the myGrid project. These include the Emboss suite of bioinformatics tools, OpenBQS which provides access to the MEDLINE database, the Sequence Retrieval System at the EBI and the Gene Ontology database.

### myGrid workbench

The myGrid project has also created a demonstrator application called the myGrid workbench, which is used to show how the services in the myGrid framework can be employed in bioinformatics (Fig. 2). The workbench enables a scientist to browse the contents of the MIR, enact workflows and also

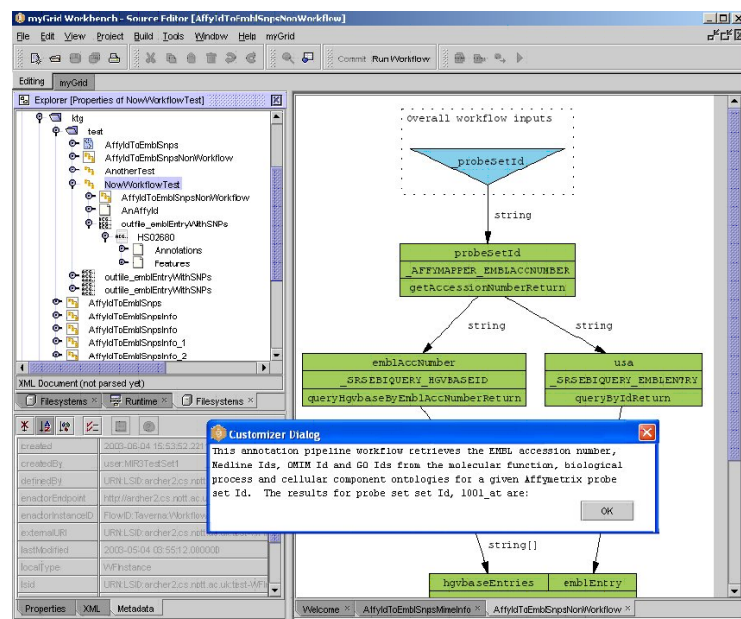


Figure 2. A screenshot of the myGrid workbench

view workflow definitions, provenance information and results generated from *in silico* experiments.

## Taverna

Workflows definitions in myGrid are scripted in the Scufi language. They can be composed using the Taverna application which enables graphical editing of the workflows. This tool can be downloaded from <http://taverna.sourceforge.net>.

## Graves' disease scenario

The technology developed by myGrid has been used to implement a scenario based on the genetic analysis of Graves' disease (GD). GD is an autoimmune disease in which an individual's lymphocytes produce auto antibodies causing cells in the thyroid gland to release higher levels of thyroid hormone.

The GD scenario implemented by myGrid has been divided into the sections shown in Figure 3.

## Annotation pipeline

A pilot microarray study identified genes differentially-expressed in GD patients. In order to understand why these genes were expressed in lymphocytes from GD patients but not in healthy individuals, biological databases such as EMBL, GO, HGBASE, OMIM, and MEDLINE need to be

queried to view information about gene structure and function, and their association with diseases. A series of annotation pipeline workflows were written in Scufi and enacted from the myGrid workflow to retrieve this information.

## Genotype Assay Design System

Single nucleotide polymorphisms (SNPs) are single base pair changes in the genome, which give rise to genetic variation amongst individuals. The differential expression of the candidate genes in GD individuals may be related to the presence of SNPs. The characteristics and frequency of the SNPs in GD patients need to be investigated using restriction fragment length polymorphism (RFLP) assays to ascertain how those SNPs are involved in GD.

Workflows were composed using myGrid services to:

1. Query databases for SNP information about differentially expressed genes
2. Design the short lengths of DNA (primers) that are used to amplify that section of the DNA sequence in the PCR experiment
3. Select a restriction enzyme that is specific to a particular SNP for the RFLP experiment

## 3D Protein Structure

Studying the structure of a protein can provide an insight to its function. myGrid services were orchestrated in workflows to:

1. Query the Protein Data Bank (PDB) database, a protein structure database, and view the structure of the protein encoded by the candidate gene if available. If so, view the protein structure to study how it relates to the function of the protein.
2. Obtain information about the protein, e.g. its function and functional domains, by querying the SWISS-PROT and InterPro databases.

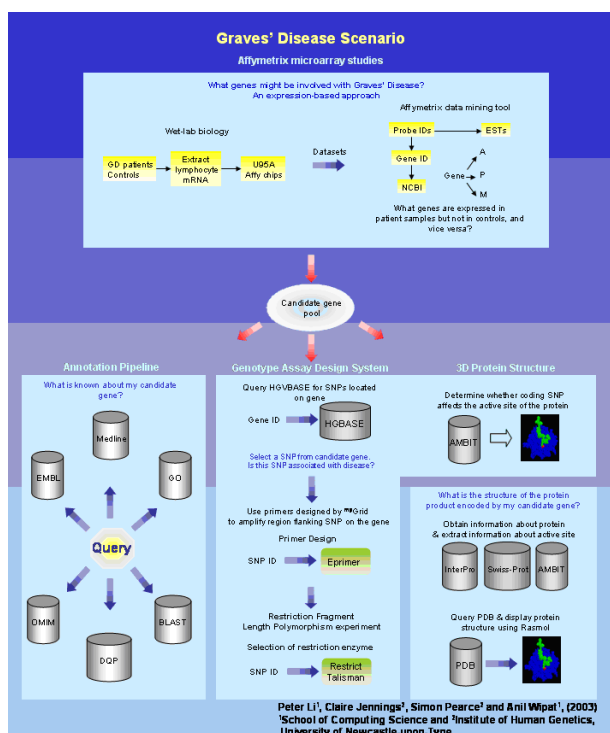


Figure 3. Bioinformatics of the Graves' disease scenario

