Performing *in silico* Experiments on the Grid: A Users Perspective

Robert Stevens[1], Kevin Glover[3], Chris Greenhalgh[3], Claire Jennings[4],
Simon Pearce[4], Peter Li[2], Melena Radenkovic[3], Anil Wipat[2].

[1]Department of Computer Science
University of Manchester
Oxford Road
Manchester
M13 9PL

[2]Claremont Tower
School of Computing Science
University of Newcastle Upon Tyne
NE1 7RU

[3]School of Computer Science and Information Technology
University of Nottingham
Jubilee Campus
Wollaton Road
Nottingham
NG8 1BB

[4]Institute of Human Genetics
University of Newcastle Upon Tyne
International Centre for life
Central Parkway
Newcastle Upon Tyne
NE1 3B2

Abstract

e-Science and the Grid are not the same; the large-scale movement of data and the exploitation of computation is not the same as the creation, performance and management of an *in silico* experiment. The notion of the marshalling of resources and creation of virtual organisations begins to bring in a flavour of science, but something more is needed over and above the classic Grid to enable e-Science. This paper looks at the requirements of e-Science from the user's perspective. The [my]Grid project aims to provide a toolkit of services that comprise the Information Grid and the applications that sit there upon. The aim is to provide a set of services that have the facilities to enable bioinformaticians (in particular) to perform *in silico* experiments using applications built upon components from a Grid enabled middleware layer. This paper introduces the [my]Grid project and explores the nature of an *in silico* experiment for the bioinformatics domain. The paper then reviews the general user requirements for an empirical e-Scientist. We then introduce a biological scenario, where bench experiments are coupled to *in silico* experiments, which we have used to drive the user requirements capture in [my]Grid. Then, the [my]Grid workbench, an application that demonstrates the functionality of [my]Grid is reviewed. Finally, we match the current status of [my]Grid to our general requirements and explore how we can use the current implementation to drive the capture of further, more detailed user requirements.

1 Introduction

e-Science is the use of electronic resources – instruments, sensors, databases, computational methods, computers - by scientists working collaboratively in large distributed project teams in order to solve scientific problems. That is, performing experiments in a computational environment. In order to do this successfully, any supporting software has to take into account the requirements of the practice of the e-Scientist. The classic Grid in itself is not e-Science – more layers are needed on top of any basic computer technology in order to support a scientist performing analyses and simulations upon a computer. In addition, not all sciences are the same. In this article, we look at the science of bioinformatics and the provision of a service based middleware toolkit call [my]Grid [8] that can be used to build Grid enabled bioinformat-

ics applications. In doing so, we will explore the nature of bioinformatics and the *in silico* experiment and along the way introduce the ^{my}Grid project, which is more fully described in a companion paper [4].

An *in silico* experiment is a procedure that uses computer-based information repositories and computational analysis to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact. ^{my}Grid [1] is an e-Science middleware project. The aim is to provide e-Science application developers a toolkit based upon a high-level middleware layer with which they can use a collection of semantically enriched services appropriate for e-Science. ^{my}Grid's aim is to support knowledge based sciences, such as biology and its sub-disciplines.

Bioinformaticians perform *in silico* experiments upon their data. In bioinformatics, our target area, *in silico* experiments are often realised as workflows that take data and pass them from resource to resource until the desired analytical goal or other outcome is achieved. A bioinformatician has a question or hypothesis and various data and analytic resources are brought together in order to explore that hypothesis; just as a bench scientist brings together chemicals, samples, glassware and machinery resources to test a hypothesis. An *in silico* experiment needs various services in order to perform the experiment. Workflows are collections of processes, each of which represents a bioinformatics resource. The output from one process can act as the input to a successor. The life-cycle of an *in silico* experiment can be seen in Figure 1. It is this life-cycle that ^{my}Grid and the applications built upon it need to support.

^{my}Grid provides a toolkit of high-level service-based middleware components to support construction, discovery, personalisation, execution and management of these kinds of *in silico* experiments in biology [9]. ^{my}Grid offers a variety of services that should enable bioinformaticians to fulfil their *in silico* goals. Figure 1 shows how the services available in the ^{my}Grid toolkit interact with the life of an experiment.

It is important to note that for ^{my}Grid, like any software, there are many kinds of user of that software. In order to understand and model the requirements of the e-Scientist, we have to make sure we know the users; this is reviewed in Section 2. Based upon the life-cycle of an experiment, we sketch out our basic requirements in Section 3. To gather requirements and evaluate our deliverables, we have been collaborating with a group of geneticists. We review the biological scenario and the associated bioinformatics in Section 4. Then, Section 4.3 looks at the ^{my}Grid demonstrator application, a workbench for *in silico* experiments, and how this fulfils our basic user requirements. The contribution of this work and future directions appears in Section 5.

---

¹http://www.mygrid.org.uk

2 ^{my}Grid Users

For a user, two views can be taken of ^{my}Grid: first, he or she can make use of an application built using the ^{my}Grid toolkit; second, there are users who build applications using the ^{my}Grid toolkit. The former are indirect users of ^{my}Grid and the latter are direct users of ^{my}Grid. All these users are important, but in this paper we concentrate upon those users that actually perform e-Science experiments.

In ^{my}Grid, we have identified a variety of users that we will need to support (Figure 2). System administrators have to install the ^{my}Grid toolkit and applications built upon ^{my}Grid. Service providers need to offer services in a manner suitable for ^{my}Grid. There are also application builders, without any bioinformatics domain knowledge, who will use the ^{my}Grid toolkit to build bioinformatics applications.

Those ^{my}Grid users with knowledge of biology fall into five categories: those rarely using bioinformatics tools; those frequently using a narrow, sophisticated set of tools; those that can perform *ad hoc* bioinformatics tasks; and those who, with additional computer science knowledge, build bioinformatics applications. ^{my}Grid's main target groups are the last two categories.

This means we have to design the ^{my}Grid tool kit and demonstrator applications so that an expert bioinformatician can perform arbitrary *in silico* experiments across a wide area of biology and record what he or she has done in a manner equivalent to that of a bench biologist.

3 Basic User Requirements

In this article we concentrate upon the requirements of the specialist bioinformatics users. For this group, the requirements are centred upon the performance of an *in silico* experiment. Taking the performance of an experiment at a high-level, we can set out some basic requirements:

- The creation or discovery of a workflow that matches the experimental goals of the scientist.

- Matching input to the workflow. In bioinformatics, input data are often collections. and the user may well need to select only a subset of those input data for input into the workflow.

- The enactment of the workflow and the monitoring of the workflow as it is enacted. One characteristic of bioinformatics experiments is the amplification of data results as the workflow is enacted. One input may give rise to many outputs. Again, the user may wish to inspect intermediate results and influence which data proceed to the next process in the workflow. All this must be recorded and traceable by users.
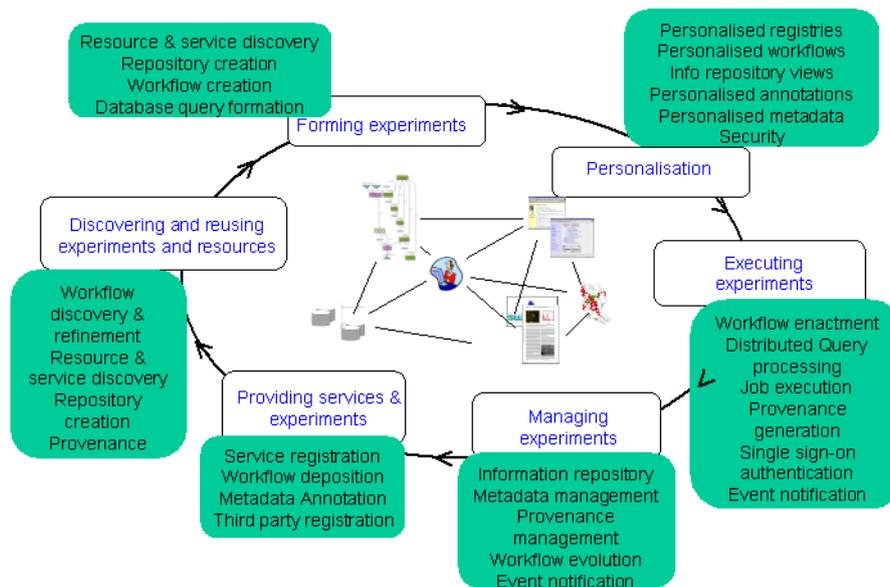
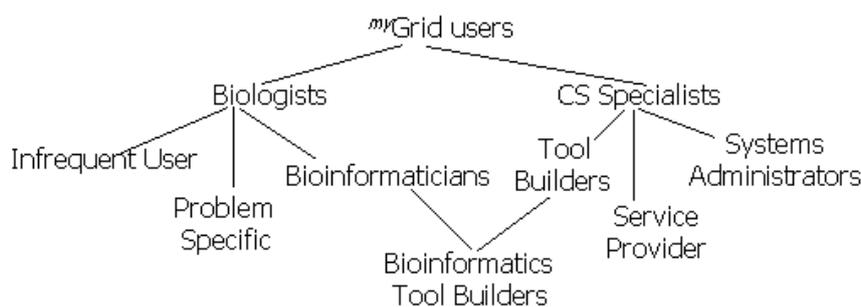Figure 1: The life-cycle of an *in silico* experiment.

Figure 2: Categories of users to be accommodated in ᵐʸGrid's design.

- Users will need to interact with third party tools, particularly the user interfaces of these tools during workflow enactment. Many bioinformatics tasks require human intervention via a user interface.

- Having completed enacting the workflow, the user will wish to view results. There is a grand management and interaction role wrapped up in this requirement: there is a plethora of information associated with any experiment; the user will need to view and manage these data. In addition, no experiment sits in isolation – all data and experiments are interlinked. Provision will need to be made to support the e-Scientist in using this information.

- As new data arrive in the user's or collaborator's file stores; analytical tools and data resources change; new workflows are created or those already used are modified; the user will wish to be notified. Having looked at the type of change, the user can make a decision about repeating, re-running or otherwise performing an experiment.

- The user, like a bench scientist, will wish to record their hypothesis; materials and methods; decisions, results and conclusions. This is provenance data about the *in silico* experiment. All the data stored by the user has provenance: the workflows; the enactments of the workflows; the input and output data; and the users themselves.

- An e-Scientist will wish to make free textual annotation/commentary of *in silico* experiments, that record his or her thoughts, over and above those mentioned above.

All the experiments' data holdings will need to be interlinked to form a Web of e-Science resources [6]. Each e-Scientist works in the context of a group; a project; an investigation; and an organisation. This requires that the Web of experimental holding can be viewed along many axes: Organisation, theme, study, project, investigation, experiment, biological concept, etc. This enables the e-Scientist to create their own personal views of their world of work. This ability, together with their own annotations and inter-linking of experiment holdings, means that the e-Scientist can be placed at the centre of the world. This is the emergent property of personalisation to which ^{my}Grid aspires.

As well as ^{my}Grid's automatic annotation and linking of holdings, the user will need to add his or her own annotations. In addition, there will be a need for third party annotation – one user may wish to add comments to another entity. this could be notes to collaborators, private comments on validity or trust or simple management information.

In summary, we have a set of basic user requirements that demonstrate the need to provide a rich level of support for a complex scientific process. In the following sections we put these basic requirements into a context by describing the biology and bioinformatics undertaken by our collaborators on the ^{my}Grid project.

4   A User Oriented Application

In order to demonstrate ^{my}Grid's prototype services as an implementation of those requirements gathered so far, as well as a tool for gathering more requirements, we have developed a collection of *in silico* experiments and a application for exhibiting those experiments. In this immediate case, the ^{my}Grid Work-Bench demonstrator application is targeted at human genetics researchers studying inherited genetic disorders. Though a demonstrator, the ^{my}Grid WorkBench is a generic application where many types of *in silico* experiments can be created and enacted. The 'workbench' metaphor provides a model by which an e-bioinformatician is offered a working area that gives him or her access to the ^{my}Grid services' functionality and application components through a GUI. We start, however, by describing the biology background of our scenario.

Single nucleotide polymorphisms ² (SNPs) are often responsible for phenotypic variations leading to disease. Classically, studies of SNPs and genetic disease involve several stages of experimental procedures, each with a computationally based component: identification of the gene/loci associated with disease, identification of the SNPs within the region, verification of the SNPs and SNP typing studies in human populations in order to find SNPs associated with the disease state.

The necessity for computational studies that are an integral part of the biological experimentation illustrates how bioinformatics is typically incorporated into the work of the molecular biologist. The ^{my}Grid team have collaborated with a group of biologists working on such a scenario to develop both the technical and user support for *in silico* experimentation as outlined above.

4.1   Graves Disease

Our demonstrator application uses ^{my}Grid technology to build in silico tools for a study that seeks to identify genes and SNPs associated with a genetic autoimmune condition known as Graves' disease. The condition is an autoimmune disease of the thyroid in which the immune system of an individual attacks the cells of the thyroid gland resulting in hyperthyroidism (thyroid overactivity). The symptoms of the disease include weight loss, increase pulse rate, trembling, heat intolerance, goitre (thyroid enlargement) and sometimes exophthalmos (protrusion of the eye-balls).

An overview of the mechanisms that are thought to lead to the diseased state are shown in Figure 3. In the normal state, the level of thyroid hormone released by the thyroid cells is controlled by a negative feedback mechanism from the pituitary gland, which releases thyroid stimulating hormone (TSH) in response to the tissue levels of thyroid hormones. TSH modulates the activity of the thyroid cells through interaction with a specific cell-surface TSH-receptor. Graves' disease is caused by the secretion of thyroid-stimulating autoantibodies by the lymphocytes of the immune system. These autoantibodies stimulate thyroid cells via the TSH-receptor and override the normal feedback mechanism, leading to hyperthyroidism [10].

4.2   The Bioinformatics Experiments for Graves' Disease

Researchers studying human genetic disease ultimately wish to establish which genes are affected in the diseased state, the changes in those genes and the underlying molecular mechanisms that lead to the autoimmune response. The hypothesis is that SNPs are instrumental in the disease mechanism and the experiments are designed to explore this hypothesis. We have established a set of ^{my}Grid *in silico* experiments in the form of workflows that will help the experimental biologist in three key areas:

i Providing systems that will aid the researcher to make hypotheses regarding the genes/loci involved in the disease;

ii Providing systems that facilitate the design of experimental tools that will be used to test these hypotheses in the laboratory;
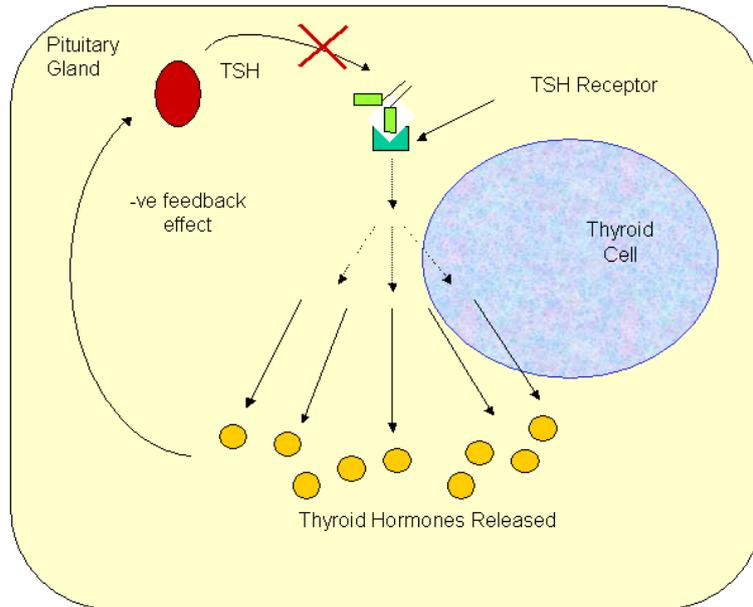

Figure 3: The normal regulation of thyroxine production and the gross mechanism seen in Graves' disease that permanently 'switches on' thyroxine production.

iii Assist the user to begin to establish the molecular basis for the disease once association has been experimentally confirmed.

Figure 4 shows the *in silico* experiments for Graves disease research that are facilitated by these workflows.

In our scenario, information about genes putatively involved in the disease is initially obtained from high-throughput expression studies using Affymetrix ^{TM}GeneChip arrays to identify genes that are differentially expressed in lymphocytes of diseased versus non-diseased subjects. The identification of differentially expressed genes is achieved through the use of standard microarray analysis technology [1]. The differentially expressed subset is uploaded into the users information repository resulting in a set of 'candidate genes termed the *candidate gene pool*.

Our first ^{my}Grid *in silico* experiment is an annotation pipeline that is designed to help the user reach a hypothesis about which genes in the candidate gene pool maybe involved in the diseased state. This workflow is used to return links to annotation data from a range of genomic databases and the literature, for each gene in the dataset. A distributed query over three Grid enabled databases is used to find ontological terms that are common to co-regulated genes. The user can assimilate the information provided and make a decision about which gene or genes that they wish to examine in more detail, and ultimately take back to laboratory studies.

Once a gene has been selected, our second *in silico* experiment allows the user to examine the genomic context of the gene and its features visually using a graphical display. Any SNPs lying in the region of the gene are displayed and SNP databases are queried to provide further information. This information is provided in order to assist the user in making a selection of the SNP that they suspect may be associated with disease and that may make a suitable candidate for typing studies in human populations. There are a number of methods by which SNPs can be tested for association with the disease state. All methods rely on determining the genotype from PCR amplified genomic regions, and one of the most common is based on Restriction Fragment Length Polymorphism (RFLP) analysis. This workflow incorporates services that will suggest suitable primer sequences for PCR amplification of genomic DNA and will find an appropriate restriction enzyme for an RFLP assay.

The third ^{my}Grid experiment is designed to help the user predict the effect of SNPs on the coding regions of a gene of interest. A protein annotation pipeline is used to derive information regarding the protein product encoded by the gene of interest, including the protein structure if it exists. The ^{my}Grid text mining service, AMBIT [3], will be used to aid the user in searching the literature for information regarding the active site of proteins and to determine whether the presence of an SNP is likely to lead to a change in protein function.
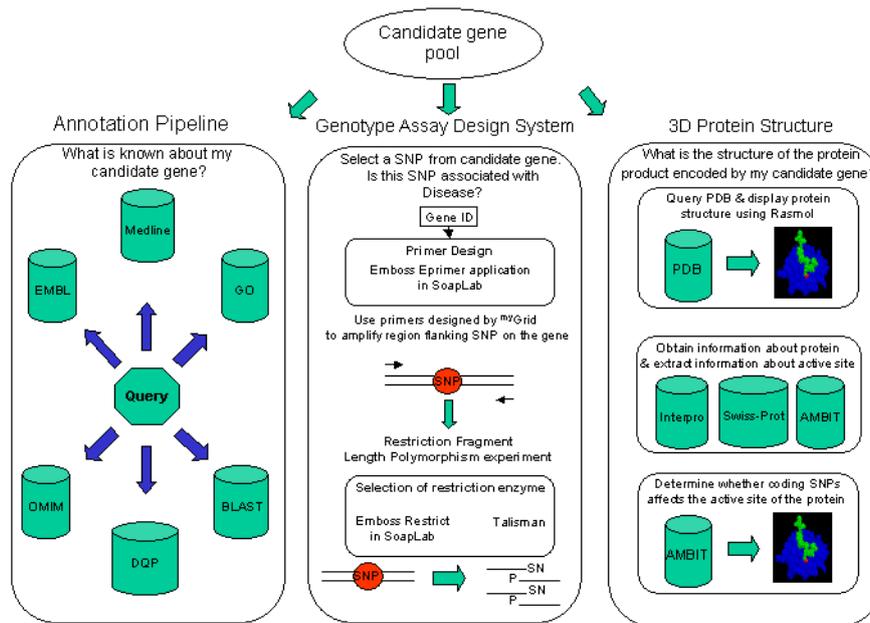
Figure 4: The bioinformatics studies used in the elucidation of the molecular mechanism of Graves' disease.

In these *in silico* experiments, we can see some classic bioinformatics. This scenario illustrates the basic needs of the bioinformatics e-Scientist. Many complex, multi-source queries have to be repeatedly applied to numerous data. Workflows representing these *in silico* experiments need to be created. Once created they need to be stored and then 'discovered' every time the user wishes to do an analysis. Each workflow is run many times, for many different data points – the results need to be recorded and organised. the scientist will need to be able to review both these results and the records of what was done and why it was done. In addition, we can see the need for the bioinformatician to use third party software to view results and potentially view intermediate results during a workflow. These workflows can deal with the results of high-throughput experiments, where many thousands of data points are analysed. Again, this means that the user needs good management of their data repository. He or she has to be able to find the results of an experiment; the data that was used as input to that experiment; to navigate between this and related experiments; and to look at the history or provenance of the data, workflows, workflow runs etc. In the next section, we will briefly examine how this has been achieved on the ^{my}Grid WorkBench.

4.3 The ^{my}GridWorkBench

The ^{my}Grid demonstration application uses this *in silico* workbench metaphor to provide bioinformaticians with a virtual WorkBench. ^{my}Grid's WorkBench is based on the NetBeans platform and provides:

1. A structured view of the organisation's information repository, which contains the data and workflows that the bioinformatician may use;

2. A notification service client, which allows the user to track changes in resources of interest;

3. custom support for key bioinformatics data types and file formats (e.g. sequences, workflows); and

4. A set of wizards that guide the user through common tasks such as workflow discovery and execution.

The basic requirements described in Section 3 have been met in the following ways:

- Workflows are created using the external Scufl Workbench which is part of Taverna. The ^{my}Grid WorkBench allows the user to upload the workflows to an MIR, and attach semantic annotation to it.

- Semantic discovery of workflows acts as the first part of the wizard used to run workflows. The wizard is started by selecting a file or set of files in an MIR and clicking the 'Run Workflow' button. The wizard uses the selected files to semantically discover any workflow that can use that type of file as an input. The workflows are presented as a list to the user to select. Highlighting any of the workflows will display a textual description of what it does, alongside a diagramatic

representation of the services used, and the flow of data between the bioinformatics services.

- Once a workflow has been selected for the chosen data, the user is then given the opportunity to fill in all the other required inputs. If multiple inputs were selected for the workflow, the ^{my}Grid WorkBench can run the workflow iteratively across all the selected inputs and coordinate the linking of all data returned. The enactor monitors the progress of the workflow enactment and provides feedback on status and which bioinformatics service is running.

- Incoming notifications are displayed in a list of notifications, with all the new ones highlighted. To subscribe for notifications, a wizard provides a list of available topics from which the user may select. The user's current subscriptions are displayed in a separate list.

- Each time a workflow is run, that enactment's results appears as a file within the mir. The user is informed by a notification that new results have returned. This file, has all the associated provenance collected by the running workflow, which can be displayed in html. Also, the provenance displays all the files used as inputs and outputs as children [5].

5 Discussion

In this article, we have described the requirements of an e-Scientist in the context of bioinformatics. An e-Scientist performs *in silico* experiments and the classic Grid, though potentially useful, is not sufficient to support this paradigm. In the ^{my}Grid project, we are attempting to provide additional layers on top of the classic Grid that will form the information Grid [8]. These layers will afford the e-Scientist the facilities that support the performance of an *in silico* experiment. The ^{my}Grid information layer services and application components meet the requirements of the e-Bioinformatician in the following ways:

1. A ^{my}Grid information repository (MIR) stores workflows, the results from workflows, input data, provenance records for each item in the MIR and the links between these data items. In addition, the MIR stores semantic annotations for each data item.

2. ^{my}Grid offers a discovery service that can work over the MIR and external repositories of Web Services and workflows. This service can use the semantic annotations in the MIR and other repositories. This discovery service is largely hidden from the user; he or she only sees a workflow wizard that offers a collection of workflow or services matched upon the data input type.

3. As well as simply re-using existing workflows, an e-Bioinformatician may wish to adapt an existing workflow or create one *de novo*. One application component of ^{my}Grid is a workflow editor, Taverna[3], in which it is intended to incorporate the discovery service to ease the creation of workflows and the binding of Web Services to workflow components.

4. Once created and supplied with data, the user will wish to run or enact the workflow. ^{my}Grid has a workflow enactment engine[2], that runs the workflow and gives feedback to inform the scientist as to progress of the enactment. The workflow language and the enactor also meet the requirement of having branch points and the ability to run parallel sub-workflows within *in silico* experiments. Finally, in the future, the user will be able to interact with a running workflow and intercede with intermediate results of a workflow to check on validity and edit, remove or otherwise alter data.

5. The world of bioinformatics is a highly volatile one. As data repositories , tools, services, workflows, *etc.* change, the user can be notified and make the appropriate decision on updating and/or re-running *in silico* experiments. Though early in its development, the ^{my}Grid notification service [7] can inform users on changes in data (within and without the MIR).

6. A major requirement, realised in ^{my}Grid, is the need to record the why, how, what and when of events during an *in silico* experiment. Provenance data are generated and linked to other experiment data holdings within the MIR, giving the e-Scientist a complete record of everything that has be done by any of the ^{my}Grid services.

It is not enough to provide one 'proscriptive' view of the complex, distributed e-Science world. Progress in science is made by gathering evidence for a point of view – our aim in ^{my}Grid is to provide the infrastructure that enables this style of *personalisation*. This will range from the ability to interlink various experiments' data holdings in order to capture a scientific point of view, down to a personalised view of which Web Services and workflows are held to be trustworthy or useful. These views need not only be held by a single user – these data and information can be organised along the axes of users, projects, investigations, organisations or any other virtual organisation enabled by Grid technology.

The ^{my}Grid project is at its halfway point at the time of writing. The ^{my}Grid WorkBench and the real-life analysis scenario have been designed to capture the basic, high-level needs of the e-bioinformatician. We aim to use this demonstration application to perform

---

³http://taverna.sourceforge.net

user studies that will drive our user requirements gathering to refine our design and enhance support for e-Scientists. In addition, we will have to extend support for application builders, system administrators and service providers. Irrespective of the e-Scientist supported, it is vital that the e-Science programme and all its projects are aware of the experimental scientists needs.

References

[1] Chipping Forecast, 2001. supplement pp 1 - 60.

[2] Matthew Addis, Mark Greenwood, Tom Oinn, Peter Li, Anil Wipat, Justin Ferris, and Darren Marvin. Experiences with workflow specification and enactment for bioinformatics. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[3] R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I. Roberts. AMBIT: Acquiring Medical and Biological Information from Text . to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[4] Carole Goble, Chris Wroe, Robert Stevens, and the ^{my}Grid consortium. The ^{my}Grid Project: Services, Architecture and Demonstrator. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[5] Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis, Darren Marvin, Luc Moreau, Tom Oinn, and Paul Watson. Provenance of e-science experiments - experience from bioinformatics. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[6] J. Hendler. Science and The Semantic Web. *Science*, page 24, Jan 2003.

[7] L. Moreau, X. Liu, S. Miles, A. Krishna, V. Tan, and R. Lawley. ^{my}Grid Notification Service. to appear in Proc UK e-Science programme All Hands Conference, 2-4 Sept 2003, Nottingham, UK, 2003.

[8] Robert D. Stevens, Alan J. Robinson, and Carole A. Goble. ^{my}Grid: personalised bioinformatics on the information grid. *Bioinformatics*, 19:i302–i304, 2003.

[9] Martin Szomszor and Luc Moreau. Recording and reasoning over data provenance in web and grid services. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE'03)*, lncs, Catania, Sicily, Italy, November 2003.

[10] B. Vaidya, H. Imrie, P. Perros, E.T. Young, W.F. Kelly, D. Carr, D.M. Large, A.D. Toft, M.I. McCarthy, P. Kendall-Taylor, and S.H.S. Pearce. The Cytotoxic T Lymphocyte Antigen-4 is a Major Graves Disease Locus. *Human Molecular Genetics 1999*, 8:1195–1199, 1999.