DDT: BIOSILICO Vol. 2, No. 4 July 2004

myGrid and the drug discovery process

Robert Stevens, Robin McEntire, Carole Goble, Mark Greenwood, Jun Zhao, Anil Wipat and Peter Li

In its early development, Grid computing has focused on providing the computational power necessary for solving computationally intensive scientific problems. However, the scientific process in the life sciences is less demanding on computational power but contains a high degree of inherent heterogeneity, and semantic and task complexity. The myGrid project has developed a Grid-enabled middleware framework to manage this complexity associated with the scientific process within the bioinformatics domain. The drug discovery process is an example of a complex scientific problem that involves managing vast amounts of information. The technology developed by the myGrid project is applicable for managing many aspects of drug discovery and development by leveraging its technology for data storage, workflow enactment, change event notification, resource discovery and provenance management.

Robert Stevens*
Carole Goble
Mark Greenwood
Jun Zhao
Department of Computer Science
University of Manchester
Oxford Road
Manchester, UK M13 9PL
*e-mail: Robert.Stevens@cs.man.ac.uk
Robin McEntire
GlaxoSmithKline
King of Prussia
PA 19406, USA
Anil Wipat
Peter Li
School of Computer Science
University of Newcastle upon Tyne
Newcastle, UK NE1 7RU

▼ The Grid is proposed as the next-generation Web, and will provide the computational power and data management infrastructure necessary to support the collaboration of people, together with data, tools and computational resources [1]. The scientific process is an example of such a collaborative process where Grid technology can facilitate virtual organizations of people, machines and instruments, data and computational resources. In this article, we introduce the myGrid project, which is developing semantically enabled Grid middleware for supporting bioinformatics applications. We use the drug discovery process (DDP) as an example of a knowledge-rich application domain that can be facilitated by technology such as myGrid. Several other Grid projects orientated towards the life sciences are underway, such as the Asia Pacific BioGrid Initiative (http://www.apbionet.org/apbiogrid), the North Carolina BioGrid (http://www.ncbiogrid.org), the Canadian BioGrid (http://www.cbr.nrc.ca), the EUROGRID project (http://www.eurogrid.org) and the Biomedical Informatics Research Network (http://www.nbirn.net). All these projects have primarily focused on the sharing of computational resources and the large-scale movement of data for simulations, remote instrumentation steerage or high-throughput sequence analysis. However, the life sciences require support for a scientific process with more modest computational needs, but with a high level of semantic complexity, of which the DDP provides many examples.

The drug discovery process

The DDP in the pharmaceutical industry is a virtual, collaborative organization and is characterized by a well-defined set of phases [2]. Key decision points are present in the DDP that enable a potential drug to progress from one step in the process to the next. Each phase of the DDP can be broken down into a distinct set of processes that define a workflow or workflows for best practice to satisfy the requirements in the given phase. The decision points, to identify targets, take targets to leads, leads to candidates and so forth, involve increasingly important decisions since each new step in the DDP means that significantly more resources from the organization will be required. As more resources are expended on a selected potential drug, there are fewer potentials that will be moved forward. It is therefore critical that the decision makers at each gate have all information necessary to make a correct judgement. In this respect, it is not surprising that the pharmaceutical industry is described as being, at its heart, a knowledge-based industry.

Over the past ten years, pharmaceutical companies have been using technologies such as HTS, combinatorial chemistry, genomics, transcriptomics, proteomics and pharmacogenomics for more efficient and effective development of new drugs. This presents a new problem because these technologies generate huge amounts of data and information that

 PII: S1741-8364(04)02412-6

must be organized, integrated, interpreted and analysed. There is also an increasing need to co-ordinate, use and re-use this distributed evidence and its associated knowledge. Although technical solutions such as data mining, data warehouses and visualization tools have been applied to this area to help with organization and analysis, the problem has not yet been fully addressed. Areas still waiting to be addressed are:

1) Integration of different types of data and information.
2) Security of data such that its ownership is kept in the hands of people responsible for its creation, while exposing the information to those in the organization who are appropriate reviewers.
3) Leveraging workflow to assist the scientist and the decision-maker in organising their work in a flexible manner, and to deliver information and knowledge to others in the organization.
4) Defining the data and information objects in a detailed and accurate manner to allow easy access to this information in a drill-down fashion.

The myGrid project is targeted at developing open source, high level, knowledge-enabled middleware to support personalized *in silico* experiments in bioinformatics on the Grid [3,4]. The issues of security and authorization that pervade distributed computing are outside the specific remit of the myGrid project and we fully expect to adopt solutions provided by other Grid projects in this role. myGrid is building middleware components to be used for running *in silico* experimentation and coordinating the resulting data and knowledge suitable for addressing points 1 and 3 above. Additional components are also being built by myGrid to support the scientific method and best practice found in the laboratory that is often neglected at the workstation in the areas of provenance management, change notification and personalization [3,4]. In the rest of this article, we describe the services offered by myGrid and how they can support the DDP.

The myGrid project

The myGrid middleware framework employs a service-orientated architecture. The various myGrid services can be used as a whole or in various combinations, depending on the needs of the application. This service model is uniform throughout the myGrid architecture, such that the networked biological resources act as services, as do all of the myGrid middleware components. These services need to be offered within a framework that accommodates their distribution and the variety of data formats within myGrid. The service model in myGrid is currently based on Web Services [5] with a later migration path to the Open Grid Services Architecture (OGSA) [6]. Thus, myGrid has many of the most important attributes of the Grid, namely creating virtual organizations of distributed people, tools, data and other resources, but is currently awaiting developments in OGSA to exploit the potential of distributed computation. The aim of myGrid is to provide the services for managing the complexities of experiments and analyses in the life sciences. Thus, myGrid does not perform the same tasks as, for instance, Globus [1] or OGSA; it adds a layer of support services above such standards to meet the needs of life scientists in creating, running and managing experiments and analyses. Many bio-services, such as computationally expensive similarity searches, are readily enhanced through parallelization and the use of Grid standards such as Globus and OGSA. myGrid expects that these and its own services will be able to use such standards as they develop.

Services for designing experiments

myGrid regards *in silico* experiments as performing queries on distributed information repositories [7] and the execution of workflows to process data by a sequential series of one or more bioinformatics analytical or database services [8]. Consequently, workflow creation, discovery and enactment form a central feature of myGrid services. Workflow or process models have a long history and widespread technology for describing and performing work. It forms a large area of work within both the Web and Grid communities, with many competing languages for the description and enactment of workflows [9].

Prototypes of bioinformatics Web Services for performing tasks in workflows available from myGrid include BLAST and the analysis tools available in the European Molecular Biology Open Software Suite (EMBOSS) [10]. These services have been created using Soaplab, a framework for deploying command-line tools as Web Services [11]. Recently, several third party services have also been made available from various organizations. Examples of these bioinformatics Web Services include XEMBL [12] hosted by the European Bioinformatics Institute, the services provided by XML Central of DDBJ [13], the KEGG suite of databases [14], a range of analysis services offered by the PathPort project [15] and those of the BioMOBY project [16]. All these services are highly heterogeneous at both the semantic and syntactic level. Outputs and inputs differ in format and composition, and data often have to be extracted from results before they are passed to the next service. This requires extra, glue-like services to be available to adapt the output of one service to the inputs of the next.

Workflows in myGrid were initially written using the Web Services Flow Language (WSFL) to define the type and order of service invocations [17]. However, WSFL and other languages such as the Business Process Execution Language

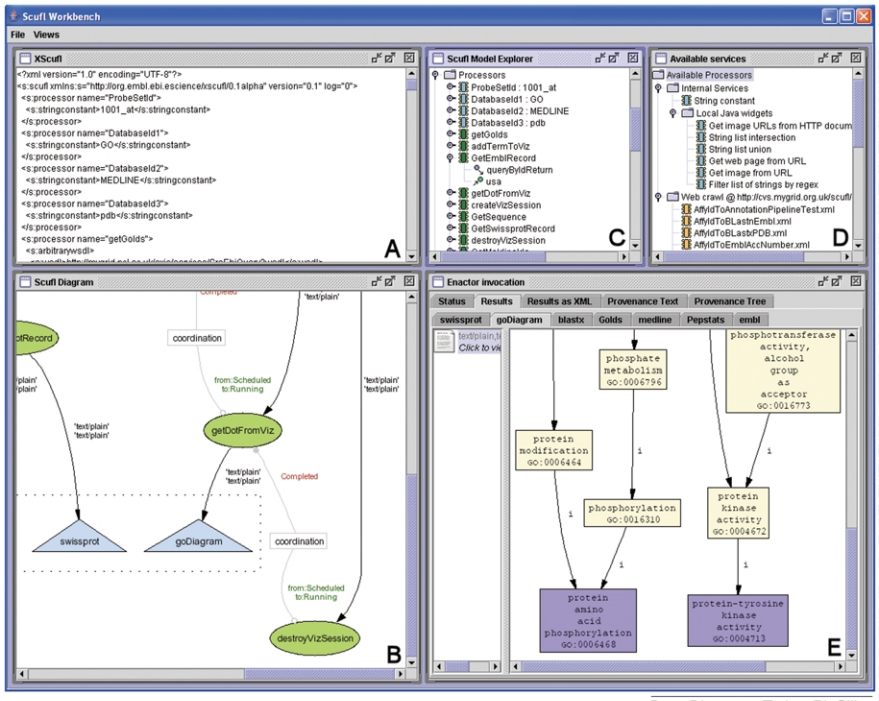

Figure 1. The Scufl workbench. A number of views are provided by this application for the composition and enactment of Scufl workflows. A workflow can be viewed **(a)** in its XML format, **(b)** in graphical format and **(c)** as a tree structure using the Scufl Model Explorer, which is also used to manipulate the workflow. Expanding a processor node reveals its inputs and outputs (c). **(d)** The workbench includes a service palette for browsing local and remote services, and workflows. **(e)** An enactor launch panel is used to display provenance information and the results generated by enacted workflows. The enactor launch panel (e) shows the result of the goDiagram workflow output (b), which is a subgraph of the Gene Ontology corresponding to terms associated with a Swiss-Prot identifier.

(BPEL) were deemed unsuitable for composing scientific workflows because they do not have the levels of user abstraction necessary for most bioinformaticians, and high-quality, free tools were not available to support standards [18]. This led to the creation of the Taverna project within myGrid to develop our own workflow language called the Simple conceptual unified flow language (Scufl) [19]. It is a high-level, XML-based, conceptual language in which each processing step of the workflow represents one atomic task. Thus it is a declarative language, where the user describes what is to be done, rather than how to do the task and, as a consequence, it is not a full programming language. A workflow enactment engine, Freefluo, has also been developed for the enactment of workflows written in either WSFL or Scufl [8]. Freefluo is able to track data provenance that is also required during the execution of *in silico* experiments [8]. Provenance is information that identifies the source and processing of data by recording the metadata and intermediate results associated with the workflow. This type of information can be a useful audit trail when investigating how results, in particular those that appear erroneous or are unexpected, have been produced by workflow processes.

Services for supporting e-Science

In the future, many hundreds of different services will be available, with massive redundancy further increasing the number. myGrid users are assisted in discovering appropriate resources during their orchestration of services in workflows. The discovery of these diverse resources is based on describing, in a formal manner that is interpretable both by humans and computer applications, the services that process data objects and the data objects themselves [20]. This enables the outputs of one service to be semantically matched against the inputs of another. Currently, most data are only described by a MIME type. Consequently, semantic matching by, for example, 'protein sequence' will facilitate the construction of workflows by guiding users to those services that match criteria selected by the user. The types of descriptions required are technical metadata about their origins and quality of service, structural descriptions of the data types or method signatures, and semantic descriptions that cover their bioinformatics or molecular biology (e.g. an enzyme or alignment algorithm) [20]. Services need to be described semantically so that a discovery service can match on inputs, outputs, task performed and resources used. In myGrid, data and services are annotated with terms based on multiple ontologies to produce semantically rich services from which workflows may be built, enacted, annotated and re-used [20]. This semantic enrichment is available during workflow construction; users do not directly write in the workflow language, but use an application called the Scufl Workbench (Figure 1), which has been developed by the Taverna project and acts as graphical editor into which the semantic discovery service has been integrated [8].

Notification

A workflow may need to be re-run when new or updated data, and analytical tools become available. myGrid has a

notification service to mediate an asynchronous interaction between services [21]. Services may register the type of notification events they produce and clients may register their interest in receiving updates. The type and granularity of notification events is defined with ontological descriptions in metadata exchanged with the notification service.

Provenance

Biologists routinely record the provenance of their bench experiments in laboratory books and this should also be true for computational experiments. To build an audit trail during the running of a workflow, myGrid services automatically record provenance information about data, services and results [22]. When a workflow is run, each service produces its own results and these are taken as inputs to another service that, again, produces more results. These fragmented results need to be coordinated such that a user can examine and validate the results of a workflow. In addition to this data derivation, there is a need to record information such as the services used in enacting a workflow; their versions, owners and locations; the workflow used, its creator, its enactor, a hypothesis, findings, etc. myGrid uses an information model throughout its components to capture and structure these provenance data and other details of the life-cycle of an experiment. In brief, the model splits into two parts: first, organizational information such as the members of the research group, data access rights, current projects and their experiments; second, information about the life cycle of a single experiment such as its design, when it has been performed, results it has produced and provenance of those results.

The information model is intended to pervade the myGrid architecture at all levels. An information repository stores information corresponding to entities in the model. Finally, we enable the user to explore the context of experimental data by providing associated metadata with each item. This metadata uses a schema derived from the information model and references other items in the repository. This stored provenance information also enables the use of notification events generated by services to determine whether a workflow needs to be re-run (e.g. if a new version of a databank used by a workflow is released).

myGrid offers middleware to bioinformatics application builders, who can choose to use any combination of myGrid services. Currently, we demonstrate myGrid's service through the Scufl workbench for the composition of workflows (Figure 1) [8]. During construction, a user can use semantic discovery and enact the workflows he or she uses. Taverna via the Freefluo enactor can also collect provenance information and coordinate collections of results. Notification is available as a Web Service and a custom client is used to subscribe to notifications.

The drug discovery process in the pharmaceutical industry

It is typical in each phase of the drug development process that a project team will be formed. The team is usually led by a champion in biological research who believes in the potential of the compound to become a drug. The project lead will need to form a team pulled from several sections of the pharmaceutical organization. These members of the team are often matrixed into the project such that each person represents a specific capability and that the group will no doubt have representation on several projects. It is essential that each participating group is organized appropriately around their task and understands how their task feeds into others. A team might be composed of:

- A biosciences sub-team to investigate the potency and selectivity of the drug compounds.
- A chemistry sub-team to analyse SARs, the tractability, and the potential for development of a particular substance.
- A pharmacokinetics group to study the metabolism of the drug compounds.
- A toxicology sub-team to investigate the toxicity of the drug compound.
- Information analysts to review the competitive landscape and to provide the team with relevant scientific information from other groups within the organization that may have a bearing on the project team's effort.
- Patent experts to judge the current landscape and leverage any patentable outcomes from the biosciences sub team.

These teams will leverage a tremendous amount of data and information, such as structural and functional information on genes, sequence searching and retrieval, toxicity, gene to function to target associations, efficacy and viability. The key problems in dealing with all these data and knowledge are:

- To provide a well-defined organizational structure, including customisable workflows that provide the team with the information relevant to each of them as soon as it becomes available. In addition, the team needs to be notified of, or otherwise find, other relevant work being done within the company and outside of the organization.
- To support decision points in the DDP which are highly dependent on scientific information. Given the number of teams working on a compound during the early phases and the amount of data and information collected, it is critical that the decision-makers have easy and sophisticated access to this information.

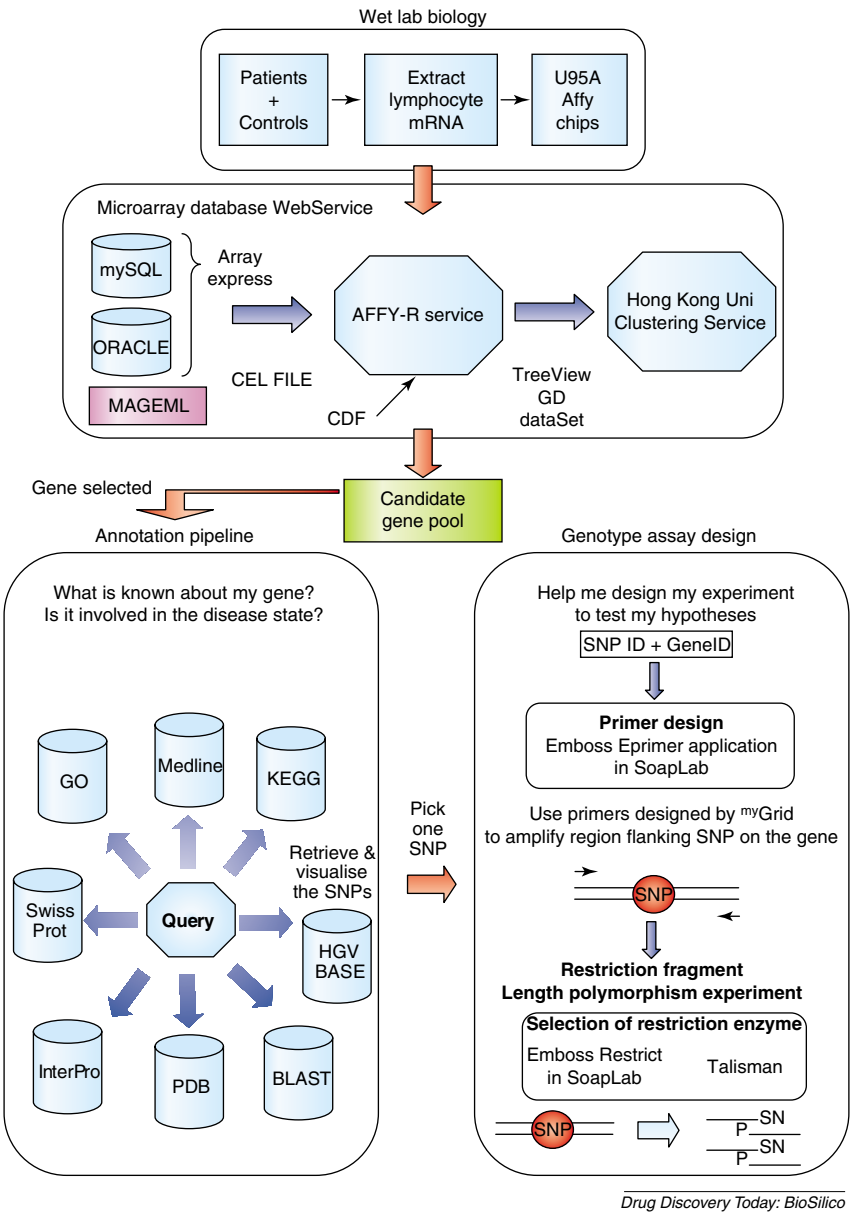

Figure 2. A high-level view of the three workflows used for the genetic analysis of Graves' disease. The first workflow involves cluster analysis of genes based on their expression levels in the microarray data set. The microarray data retrieved from the ArrayExpress database Web service is then transformed into Treeview format using the AffyR service and then processed by an algorithm by the microarray clustering service. The second workflow is a gene annotation pipeline in which a gene is selected from the output of the microarray data analysis workflow, and annotation from database services such as SwissProt, KEGG, GO and MEDLINE with analyses including BLAST are returned. In addition, information on SNPs that have been identified in the gene is retrieved from the HGVBASE database. Finally, a genotype assay design workflow takes one or more of these SNPs as input to design primer sequences and detect restriction sites for RFLP experiments using the Eprimer and Restrict programs from EMBOSS which have been deployed as Soaplab services.

There is a significant expense to the business in organizing all of this data in an orderly and user friendly manner. In addition, there is a significant level of complexity in cogently integrating the data and information because much of it has been generated and kept in multitudinous storage formats; in many places and in heterogeneous representation forms.

The DDP and myGrid services

myGrid technology has been employed in the genetic analysis of Graves' disease (GD) [23]. GD is an autoimmune disease primarily affecting the thyroid gland. In this condition, lymphocytes secrete auto-antibodies that bind to receptors on cells in the thyroid, resulting in hyperthyroidism. Symptoms of the disease include weight loss, trembling, muscle weakness, increased pulse rate, heat intolerance and expophalmos. The analysis of GD genetics involves the discovery of genes involved in the diseased state and the genotyping of single nucleotide polymorphisms (SNPs) occurring in those genes. The analysis begins with a laboratory microarray experiment where the expression levels of over 10,000 genes in lymphocytes from GD patients and healthy controls were measured. Three workflows were designed, composed and enacted as bioinformatics experiments for the analysis of the microarray data (Figure 2). Each workflow corresponds to a distinct phase in the classical *in silico* process normally carried out using several web-based resources, and has a specific function:

1. The first workflow is concerned with the analysis of the microarray data to generate a list of candidate genes which are differentially expressed in GD and in healthy individuals.

2. An annotation pipeline workflow allows the retrieval of annotated information for each gene in the list including its location in the genome, function, other similar genes, information about the gene from the scientific literature and the SNPs identified in the gene. In addition, it analyses the composition and structure of the protein encoded by the gene and also the pathways that the protein participates in.

3. A final workflow is required to help design the wet-laboratory experiments to test the hypothesis generated by the preceding workflows. This workflow aids in primer design and identifies restriction sites for use in restriction fragment length polymorphism (RFLP) experiments to genotype control and GD patient lymphocyte DNA samples for a given SNP.

When classical and myGrid approaches were compared for performing the above bioinformatics experiments in the genetic analysis of GD, both methodologies identified NF-Kappa Beta IE as a candidate gene in GD. That the same result was achieved by both methodologies shows that myGrid is capable of producing the same results as the classical *in silico* approach. The automation of the bioinformatics experiments through the use of myGrid workflow technology also enabled the analyses to be performed faster than when they are performed using the classical approach where several websites have to be visited and queries needed to be manually entered onto web pages. Since workflows can be stored as XScufl scripts, *in silico* experiments in myGrid are also repeatable with different parameters if required.

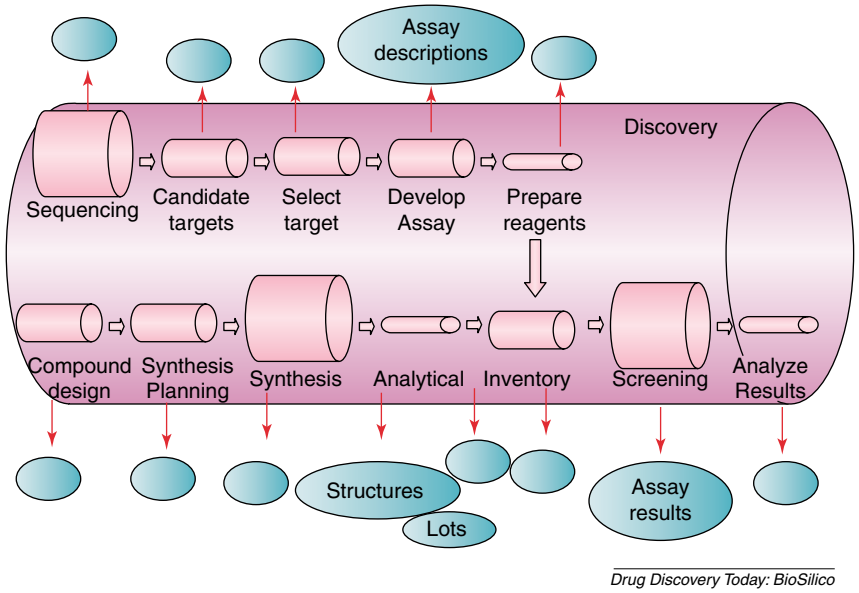

Figure 3. The drug discovery process consists of workflows involving the orchestration of people, instrumentation and computational resources. These workflows are responsible for the tasks associated with each phase of the drug discovery process such as the synthesis of compounds and assay development in the compound screening stage. Data are generated by each workflow, for example, assay results produced from toxicity screening of drug compounds.

The workflows used in the genetic analysis of GD are typical examples of bioinformatics workflows involving the querying of information repositories and the analysis of data using computational tools (Figure 2). In a similar fashion, the DDP can be considered to be one large workflow composed of sub-workflows that represent the phases of target identification, target validation, compound screening, lead optimization, pre-clinical and clinical trials (Figure 3). Each of these DDP-phase workflows is itself made up of many smaller workflows that involve laboratory, *in silico*, *in vivo* and clinical investigations performed by groups of people that are responsible for a specific task. DDP phases involve the workflow orchestration of resources that are in the form of people, analytical instruments, computers and data. Such virtual, transient organizations are the essence of the Grid and what the myGrid upper level middleware seeks to support. All these resources need to be orchestrated to achieve the goals of the DDP in an efficient and effective manner. In this respect, the DDP forms a classical target for the application of Grid computing and myGrid technology to manage the experimental process. For example, the workflows that have been used in the analysis of GD are generic so that they can be applied in the target identification phase of the DDP. The workflows will allow a disease within a pharma's therapeutic area to be substituted in place of GD to generate new potential drug targets.

Various bottlenecks in the DDP need to be addressed to accelerate and streamline the development of new drugs. Each phase of the DDP generates vast quantities of data such as provenance records, findings, conclusions and hypotheses. All this data can be stored, inter-linked and coordinated, which would also provide the access controls to ensure that different types of users, whether they are scientists or managers, with the appropriate permissions can access data belonging to other people. Every project team member will need to be informed of any significant new results when they have been generated. They also need to be notified when other teams in the organization are doing similar work, which means that the organization needs to be aware of knowledge and activities within itself. The myGrid notification service can be used so that a person with a defined role in the DDP could be notified of the availability of information required for them to achieve their task, as and when it becomes available.

It is advantageous for the working environment to be altered to suit the needs and individual preferences of users

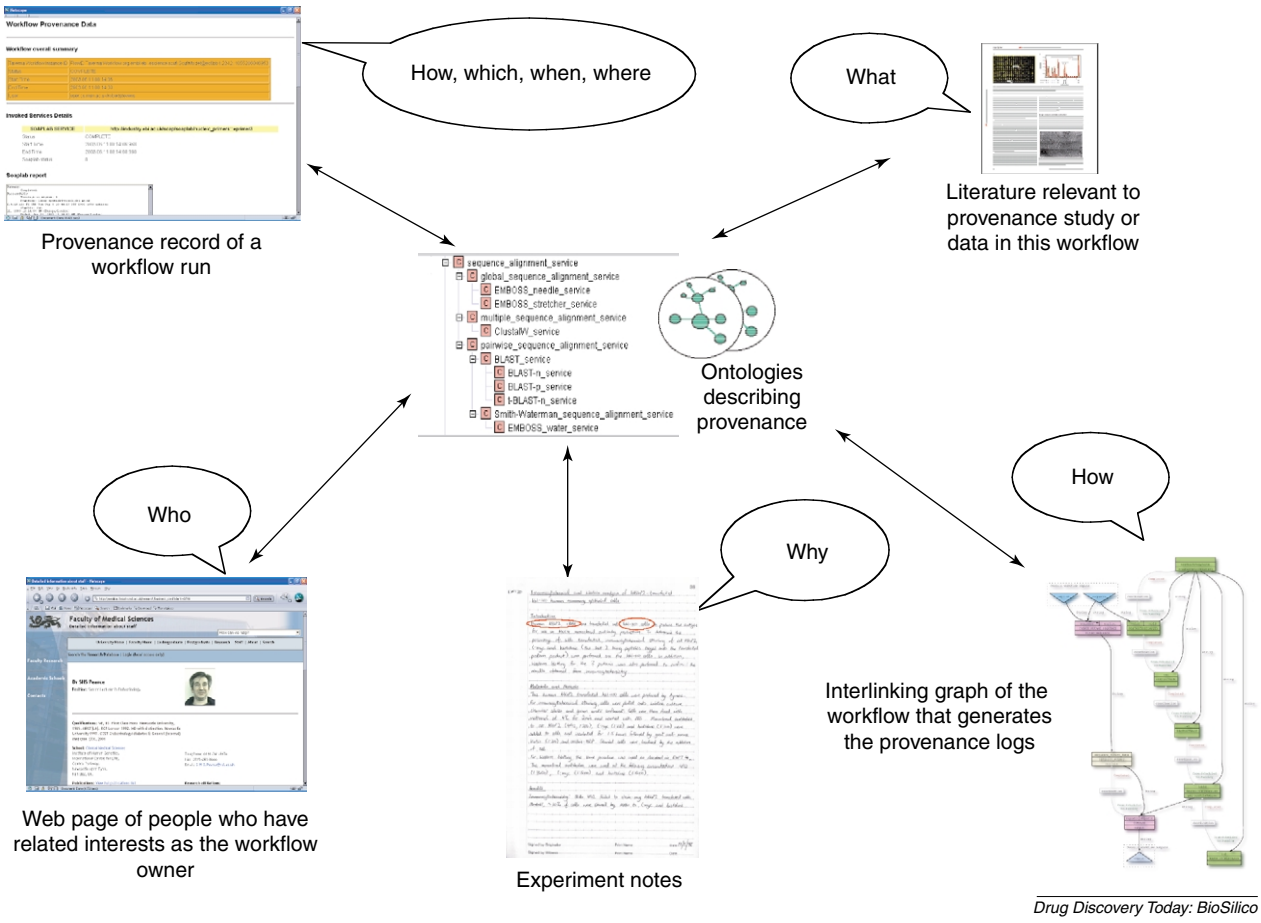

Figure 4. A web of science. A mechanism for exploring and organizing data and knowledge produced by virtual organizations.

so that they can perform their task in the DDP as efficiently as possible. The services in myGrid can be tailored according to the needs of each person. For example, users can be provided with their own perspective on data stored in the information repository such that the data are viewed along different axes, such as topic, date, user, inputs, task, etc. In a similar fashion, the resource discovery service can provide a unique view over services they can use (such as preferred or in-house services) and the opportunity to attribute metadata to third-party services, such as trust or reliability.

The processes for each phase in the DDP could be formalized as a 'production workflow'. myGrid can then be used to support the collection, integration and exploration of all this data and knowledge during the DDP by enacting these workflows. This brings the interesting concept of people as services and the use of workflows to facilitate the collaboration of the sub-teams in the DDP. For example, a person representing a team in a DDP could be contacted and asked for their views on the feasibility of a protein to be a drug target during the enactment of a workflow in the target identification stage. Furthermore, both people and computational resources can be listed in a service registry to provide the pharma with a store of its assets and experience that can be utilized in drug development projects. When these resources are described with the appropriate metadata then a project manager for a DDP can choose the best people and resources for a particular project. Investigations are underway as part of the myGrid project to enable this functionality.

Provenance

Perhaps one of the most important issues in the DDP is the management of provenance. Provenance records need to be kept about who, what, why, when, where and how experiments were run and how data were derived. The DDP will involve legions of similar experiments that need to be organized and exploited. Provenance provides a connection between new data, information, and the methods used to create, or discover, those data. It is metadata that helps people to better understand the data available and

whether this represents valuable knowledge that can be exploited. In the DDP, important decisions are made on the best knowledge available at the time, and provenance has an important role in the making of these decisions. Understanding where data comes from can have an impact on how much trust is placed in that data, and therefore the reliability of the knowledge that derives from it.

Provenance information can provide an explanation of a specific DDP [22]; for instance, where did the current information about this specific potential drug come from, which experiments have been performed, by whom, when and why. The cumulative provenance information from a pharma's DDPs, both past and present, represents an experience base. The decisions in current DDPs can be guided by the successes and failures of the past.

In the myGrid case, where *in silico* experiments are described by workflows, some provenance information can be automatically gathered. The results of a workflow can be linked with the inputs, the workflow description itself, and the services that were called as part of the workflow. Importantly, the data, both intermediate and final, produced from a workflow can be co-ordinated with each other and the services from which they were generated. Such co-ordination is vital for validation and re-use of those data. These services, which could be lower level workflows, can also provide their provenance information. At the level of atomic services, it might be appropriate to record, for example, which BLAST service was run (its version, location, configuration), the metadata about the service invocation (size of database, number of hits, etc), and any relevant user account information (especially for a commercial service). The workflow provenance can also include contextual information: when the workflow was run, on whose behalf, using which workflow engine, and as part of which project, phase, or specific DDP. The provenance data that can be automatically gathered provides a set of foundational metadata. This can be augmented by users providing richer descriptions of the reasons why an *in silico* experiment was performed, specific ontological concepts associated with the inputs and outputs (e.g. this gene id was expressed in the studied patients with symptoms X but not in the control group) [24], and their notes and observations (this result should be disregarded because more recent knowledge shows that an assumption was invalid). It should be remembered that all objects within a repository have some kind of provenance.

Provenance information is not restricted to recording the workflow instance that produced a data item, and how that was used in other workflows. The tools that import data into myGrid, for example from a laboratory management system, or a user copying relevant data from the web, has provenance. Furthermore, the challenge is not just how to record provenance data, but also how to present and exploit it. One key is to identify the concepts that link provenance information so that there is an effective way of answering, 'Have we done something like this before, and what happened?'. Provenance data of all types can be reasonably associated with many different kinds of data according to the information model. Another key is to understand the relationship of provenance information to DDP decision points. Are there effective ways of summarizing the quality of information presented to decision makers, with detailed provenance records available if they choose to 'drill down'?

The combination of provenance information and the experiments, users and investigations for which it is recorded offers personalization of the scientific process. Data and knowledge about those data can be visualized and explored along many axes. For instance, a scientist could look along an axis of all experiments about a certain drug target receding back in time with all the contextual information surrounding that experiment axis. Indeed, all the surrounding contextual information is itself organized along different axis. Figure 4 illustrates how concepts can be used to relate the provenance information for one experiment. Combining the related information from numerous experiments will create Webs of Science [24]. These provide a mechanism for exploring and organising the data and knowledge provided by virtual organizations.

Summary

Grid technology enables the formation of virtual organizations composed of people, data and services, instrumentation and computational resources. myGrid offers a collection of services that provide a technical solution to the management of the heterogeneity and complexity of bioinformatics experiments. The scientific process is, however, a human or sociological process and as such a technical solution such as myGrid cannot be a panacea to every aspect of the DDP. myGrid does, however, gather together a rich, semantically enabled collection of services upon which to build applications for the non-human core of the scientific process. This marshalling of distributed resources, together with semantically enabled services provided by myGrid's upper level middleware, provides an environment for managing the scientific process. In the context of the DDP, this is the organization of an enormous amount of data and the organization's collective knowledge and all the stages, experiments and decisions in the DDP. Through its development of semantic Grid technology, myGrid provides a model for enabling knowledge-based organization of the scientific process through provenance.

Acknowledgements

The authors would like to acknowledge the assistance of the whole myGrid consortium. This work is supported by the UK e-Science programme EPSRC GR/R67743.

References

1 Foster, I. and Kesselman, C. (1998) *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann
2 Hodgman, C. (2001) An information-flow model of the pharmaceutical industry. *Drug Discov. Today* 6, 1256–1258
3 Stevens, R.D. *et al*. (2003) myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19, i302–i304
4 Goble, C. *et al*. (2003). The myGrid Project: services, architecture and demonstrator. In *Proceedings UK e-Science All Hands Meeting 2003 – September 2003*. (http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/128.pdf.)
5 Booth, D. *et al*. (2003) Web Services Architecture. *W3C* http://www.w3.org/TR/ws-arch/
6 Foster, I. *et al*. (2002). *The Physiology Of The Grid: An Open Grid Services Architecture For Distributed Systems Integration*. Technical Report of the Global Grid Forum.
7 Alpdemir, M.N. *et al*. (2003) OGSA-DQP: A service-based distributed query processor for the Grid. In *Proceedings UK e-Science All Hands Meeting 2003 - September 2003*. (http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/114.pdf.)
8 Oinn, T. *et al*. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* (in press)
9 van der Aalst, W. (2003) Don't go with the flow: web services composition standards exposed. *IEEE Intell. Syst.* 18, 72–76
10 Rice, P. *et al*. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277
11 Senger, M. *et al*. (2003) SoapLab – a unified Sesame door to analysis tools. In *Proceedings UK e-Science All Hands Meeting 2003 - September 2003*
12 Wang, L. *et al*. (2002) XEMBL: distributing EMBL data in XML format. *Bioinformatics* 18, 1147–1148
13 Miyazaki, S. and Sugawara, H. (2000) Development of DDBJ-XML and its application to a database of cDNA. In *Genome Informatics* Dunker, A.K., Konagaya, A., Miyano, S., akagi, T. (Eds) pp. 380–381, Universal Academy Press, Inc (Tokyo)
14 Kawashima, S. *et al*. (2003) KEGG API. (http://www.genome.ad.jp/kegg/soap/)
15 Eckart, J.D. and Sobral, B.W. (2003) A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *OMICS* 7, 79–88
16 Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open-source biological web services proposal. *Brief. Bioinform.* 3, 331–341
17 Leymann, F. (2001) Web Services Flow Language (WSFL 1.0). (http://www-3.ibm.com/software/solutions/webservices/pdf/WSFL.pdf)
18 Addis, M. *et al*. (2003) Experiences with eScience workflow specification and enactment in bioinformatics. In *Proceedings UK e-Science All Hands Meeting 2003 – September 2003* (http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/108.pdf.)
19 Oinn, T. *et al*. (2004) Delivering Web service coordination capability to users. Accepted as short note and poster for WWW2004.
20 Wroe, C. *et al*. (2003) A suite of DAML+OIL ontologies to describe bioinformatics web services and data. *Int. J. Coop. Inf. Syst.* 12, 597–624
21 Krishna, A. *et al*. (2003) myGrid Notification Service. In *Proceedings UK e-Science All Hands Meeting 2003 – September 2003* (http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/110.pdf.)
22 Greenwood, M. *et al*. (2003) Provenance of e-Science Experiments – experience from Bioinformatics. In *Proceedings UK e-Science All Hands Meeting 2003 – September 2003*
23 Stevens, R. *et al*. (2003) Performing *in silico* experiments on the Grid: a users perspective. *Proceedings UK e-Science All Hands Meeting 2003 – September 2003*
24 Zhao, J. *et al*. (2003) Annotating, linking and browsing provenance logs for e-Science. 2nd International Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data, Florida.