**myGrid: personalised bioinformatics on the information grid**

Robert D. Stevens\(^1\), Alan J. Robinson\(^2\) and Carole A. Goble\(^1,\ast\)

\(^1\)Department of Computer Science, University of Manchester, Oxford Road, Manchester, UK, M13 9PL and \(^2\)European Bioinformatics Institute, EMBL Outstation—Hinxton, Wellcome Trust Genome Campus, Cambridge, UK, CB10 1SD

Received on January 6, 2003; accepted on February 20, 2003

**ABSTRACT**

**Motivation:** The myGrid project aims to exploit Grid technology, with an emphasis on the Information Grid, and provide middleware layers that make it appropriate for the needs of bioinformatics. myGrid is building high level services for data and application integration such as resource discovery, workflow enactment and distributed query processing. Additional services are provided to support the scientific method and best practice found at the bench but often neglected at the workstation, notably provenance management, change notification and personalisation.

**Results:** We give an overview of these services and their metadata. In particular, semantically rich metadata expressed using ontologies necessary to discover, select and compose services into dynamic workflows.

**Availability:** Software is available on request from the authors and information from http://www.mygrid.org.uk.

**Contact:** carole@cs.man.ac.uk

**INTRODUCTION**

myGrid is a project targeted at developing open source high-level middleware to support personalised *in silico* experiments in biology on a Grid. The Grid is proposed as the next generation infrastructure necessary to support and enable the collaboration of people and resources through highly capable computation and data management systems (Foster and Kesselman, 1998). A number of BioGrid projects are underway, including the Asia Pacific BioGrid Initiative (http://www.apbionet.org/abgi/) and the North Carolina BioGrid (http://www.ncbiorggrid.org/), the Canadian BioGrid (http://www.cbr.nrc.ca/), the EUROGRID project (http://www.eurogrid.org/) and the Biomedical Informatics Research Network (http://www.nbim.net/).

These primarily focus on the sharing of computational resources, large-scale data movement and replication for simulations, remote instrumentation steerage or high throughput sequence analysis. However, much bioinformatics requires support for a scientific process that has more modest computational needs, but has significant semantic complexity. myGrid is building services for integration such as resource discovery, workflow enactment and distributed query processing. Additional services are needed to support the scientific method and best practice found at the bench but often neglected at the workstation, notably provenance management, change notification and personalisation. The target users of myGrid are tool and service providers who build applications for a community of biologists. Early prototypes of myGrid services were developed and tested with use cases based on the functional analysis of clusters of proteins identified in a microarray study of genes showing circadian rhythms in *Drosophila melanogaster* (Claridge-Chang et al., 2001).

Following this, a distributed system has been developed to meet the requirements of researchers studying the genetics of Graves’ disease.

**METHODS**

Rather than a data grid or computational grid, myGrid is a service grid. The myGrid middleware typewrap employs a service based architecture, firstly prototyped with XML-based Web Services (http://www.webservices.org/), but with a migration path to the ‘Open Grid Services Architecture’ (OGSA) (Foster et al., 2002, http://www.globus.org/ogsa/). This service model is uniform; the networked biological resources are services, as are the myGrid components. Figure 1 illustrates the myGrid architecture stack. In this section, we present the primary myGrid services in three categories:

1. **Services for forming experiments**

   myGrid regards *in silico* experiments as distributed queries and workflows. Data and parameters are taken as input to an analysis or database service; then output from these is taken, perhaps after interaction with the user, as input to further tools or database queries.
Bioinformatics services: Services such as databank retrieval and analysis tools need to be wrapped and offered in a form that accommodates their distribution and variety of data formats. Prototypes of bioinformatics Web Services are available from the myGrid web site for NCBI BLAST (Altschul et al., 1997), WU BLAST (W. Gish, personal communication; http://blast.wustl.edu/), the complete EMBOSS application suite of over eighty analysis tools (Rice et al., 2000), MEDLINE and SRS (Etzold et al., 1996).

Workflow Enactment: Once discovered or built, a workflow needs to be run by the workflow enactment engine that will call the bioinformatics services. myGrid uses the ‘Web Service Flow Language’ (WSFL) to define the type and order of service invocations.

Distributed database queries: The OGSA-DAI project (http://www.ogsadai.org/) and myGrid project are together building a distributed query processing system that will enable a user to specify queries across a set of Grid-enabled information repositories in a high level language (initially OQL). Complex queries on large data repositories may result in potentially high response times, but the system can address this through parallelisation (Smith et al., 2002).

2. Services for discovery and metadata management
A bioinformatician requires a great deal of background knowledge in order to build workflows or distributed queries efficiently and effectively from the appropriate data sources and analytical tools around the network. Some of this burden can be relieved by describing in a formal manner that is interpretable both by humans and computer applications: (i) the services that process objects and (ii) the objects themselves (Baker et al., 1999). Both need a continuum of descriptions: metadata about their origins, quality of service, etc. structural descriptions of their data types or method signatures; and semantic descriptions that cover their concept (e.g. an enzyme or alignment algorithm). 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 using (multiple) ontologies with DAML+OIL to produce semantically rich services from which workflows may be built, enacted, annotated and re-used (Wroe et al., 2003). In this multi-level service model, classes of services (e.g. an alignment service) and instances of services (e.g. a WU-BLAST service at the EBI) may be distinguished and described formally.

myGrid has developed a federated and extended version of the UDDI registry (http://www.uddi.org/) called UDDI-M. UDDI-M registers services together with metadata about their location, ownership, version, cost, quality of service, security, etc. As data and services are annotated using concepts drawn from ontologies, they can be then associated using those concepts with the COHSE hypermedia system (Carr et al., 2001, http://cohse.semanticweb.org). Thus when a user wants to find bioservices and compose them together, the syntactic and semantic types of the available services and objects can be found and checked for compatibility. The registry may also incorporate third party annotations enabling users to personalise the choice of services. myGrid’s registry and ontologies for bioinformatics services are being developed in collaboration with the BioMOBY project (Wilkinson and Links, 2002) and the Interoperable Informatics Infrastructure Consortium (http://www.i3c.org/).

3. Services for supporting e-Science
myGrid aids users in finding appropriate resources, offering alternatives to busy resources and guiding users in the composition of resources into workflows. Available user interfaces to myGrid services include its own Gateway service and Talisman (Oinn, 2003). In addition, myGrid offers:

Notification: A workflow may need to be re-run when new or updated data and analytical software become available. myGrid has a notification service to mediate an asynchronous interaction between services. Servers 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.

Personalisation: The myGrid Information Repository (mIR) stores: XML data generated by experiments with its metadata and ontology terms; annotations of information held in the mIR or external repositories; and provenance data. Since an organisation would typically have a single mIR, it is important that different users can be provided with appropriate views of its information. These views are enforced using the security features of the DBMS on which the mIR is built and may include security rules on the permitted modification and deletion of the contents of a mIR.

Provenance: Biologists routinely record the provenance of their bench experiments in lab books and this should be true for computational experiments too. To build an audit trail during the running of a workflow, myGrid services record automatically in the mIR as much provenance information as is available about data, services and results. As well as being important for auditing, this stored provenance information 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 an exemplar of how information Grid technology can be harnessed and enhanced to accommodate the needs of biologists and a wider range of 'e-Scientists' than was the original target audience of the Grid.

ACKNOWLEDGEMENTS

This work is supported by the UK e-Science programme EPSRC GR/R67743, and DARPA DAML subcontract PY-1149, Stanford University.

The authors would like to acknowledge the myGrid team: Matthew Addis, Nedim Alpdemir, Rich Cawley, David De Roure, Alvaro Fernandes, Justin Ferris, Rob Gaizauskas, Kevin Glover, Chris Greenhalgh, Mark Greenwood, Karon Mee, Peter Li, Xiaojian Liu, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Tom Oinn, Norman Paton, Steve Pettifer, Milena Radenkovic, Angus Roberts, Tom Rodden, Martin Senger, Nick Sharman, Paul Watson and Chris Wroe.

REFERENCES


