The myGrid project: services, architecture and demonstrator

Professor Carole Goble, Chris Wroe, Robert Stevens and the myGrid consortium

http://www.mygrid.org.uk
• EPSRC UK e-Science pilot project
• Open Source Upper Middleware for Bioinformatics
• (Web) Service-based architecture -> OGSA Grid services
• Prototype v1 Release Oct 2003, some services available now.
• Targeted at Tool Developers, Bioinformaticians and Service Providers
AHM meta-talk for a meta-paper

Experiences with e-Science workflow specification and enactment in bioinformatics

myGrid Notification Service

Provenance of e-Science Experiments - experience from Bioinformatics

The myGrid project: services, architecture and demonstrator

The NEReSC Core Grid Middleware

Semantic and Personalised Service Discovery

AMBIT: Acquiring Medical and Biological Information from Text

Service-Based Distributed Query Processing on the Grid

Performing in silico Experiments on the Grid: A Users Perspective

Soaplab – a Unified Sesame Door to Data Analysis Tools
Data-intensive bioinformatics

[source: GlaxoSmithKline]
Graves disease

- Autoimmune disease of the thyroid in which the immune system of an individual attacks cells in the thyroid gland resulting in hyperthyroidism
- Weight loss, trembling, muscle weakness, increased pulse rate, increased sweating and heat intolerance, goitre, exophtalmos
Biology working together with…

Grave’s Disease caused by the stimulation of the thyrotrophin receptor by thyroid-stimulating autoantibodies secreted by lymphocytes of the immune system.

What is the molecular basis for this autoimmune response?

What genes might be associated with Graves’ Disease?

Affymetrix microarray studies

Affymetrix data mining tool
- Probe IDs
- ESTs
- Gene ID
- NCBI

What genes are expressed in patient samples but not in controls, and vice versa?

Candidate gene pool
**Bioinformatics**

**Candidate gene pool**

**Annotation Pipeline**
- What is known about my candidate gene?
  - Medline
  - GO
  - EMBL
  - OMIM
  - BLAST
  - DQP

**Genotype Assay Design System**
- Select a SNP from candidate gene. Is this SNP associated with Disease?
  - Gene ID
  - Primer Design
  - Emboss Eprimer application in SoapLab
  - 
  - SNP
  - Restriction Fragment
  - Length Polymorphism experiment
  - Selection of restriction enzyme
  - Emboss Restrict in SoapLab
  - Talisman

**3D Protein Structure**
- What is the structure of the protein product encoded by my candidate gene?
  - PDB
  - Query PDB & display protein structure using Rasmol
  - AMBIT
  - Swiss-Prot
  - Interpro

**Query**

**Information Sources**
- Medline
- GO
- EMBL
- OMIM
- BLAST
- DQP
- PDB
- Interpro
- Swiss-Prot
- AMBIT
Workflows are *in silico* experiments

What is known about my candidate gene?

- Medline
- EMBL
- GO
- OMIM
- BLAST
- DQP

**Annotation Pipeline**

**Experimental orchestration**
- Exploratory
- Hypothesis driven
- Not prescriptive
- Methodology free
- Ad hoc

[http://cvs.mygrid.org.uk/scufl/NucleotideSeqAnnotationPipelineWithGoTerms/](http://cvs.mygrid.org.uk/scufl/NucleotideSeqAnnotationPipelineWithGoTerms/)
Experiment life cycle

Resource & service discovery
- Repository creation
- Workflow creation
- Database query formation

Discovering and reusing experiments and resources
- Workflow discovery & refinement
- Resource & service discovery
- Repository creation
- Provenance

Forming experiments
- Executing experiments
- Workflow enactment
- Distributed Query processing
- Job execution
- Provenance generation
- Single sign-on authentician
- Event notification

Providing services & experiments
- Service registration
- Workflow deposition
- Metadata Annotation
- Third party registration

Managing experiments
- Information repository
- Metadata management
- Provenance management
- Workflow evolution
- Event notification

Personalisation
- Personalised registries
- Personalised workflows
- Info repository views
- Personalised annotations
- Personalised metadata
- Security

Personalised registries
- Personalised workflows
- Info repository views
- Personalised annotations
- Personalised metadata
- Security

Security
- Forming experiments
- Personalisation
- Executing experiments
- Providing services & experiments
- Managing experiments
- Discovering and reusing experiments and resources
- Resource & service discovery
Bio in silico experiments service types

- Making in silico experiments
  - workflow
  - distributed database query processing.

- Managing experimental outcomes
  - information management
  - managing metadata

- Scientific method
  - provenance management
  - change notification
  - personalisation

- Sharing experiments
  - semantic services for discovering services and workflows, and managing metadata
  - third party service registries and federated personalised views over those registries,
  - ontologies and ontology management.

- The base services that tools that will constitute the experiments
  - third party services such databases, computational analyses, simulations ....
  - specialised services such as AMBIT text extraction.
Investigation = set of experiments + metadata

- **Experimental design components**
- **Experimental instances** that are records of enacted experiments
- **Experimental glue** that groups and links design and instance components
- Life Science IDs
- URIs
- RDF
**myGrid in a nutshell**

**External Services:** AMBIT, SoapLab, EMBOSS...

**External Applications:** workbench, portal, Talisman, Taverna

**Core services**

- **Semantic grid capabilities**
  - knowledge-based technologies,
  - semantic-based service,
  - workflow & data discovery,
  - match making
  - linking investigation components.

- **High level services for e-Science experimental management**
  - provenance, change notification,
  - personalisation,
  - investigation and experiment holdings management

- **High level services for data intensive integration**
  - workflow & distributed query processing

- **A “second generation” open service-based Grid project,**
  - a **test bed** for the OGSI, OGSA and OGSA-DAI base services

**Services**

**Applications**

- **Bioinformaticians**
  - Service Providers
  - Tool Providers
myGrid Service Stack

- Work Bench
- Taverna workflow environment
- Web Portal
- Talisman application
- Gateway

- Registries
- Service and Workflow Discovery
  - Ontology Mgt
  - Metadata Mgt
- Personalisation
  - Provenance mgt
  - Event Notification

- FreeFluo Workflow enactment engine
- myGrid Information Repository
- OGSA Distributed Query Processor

- Web Service & Grid communication fabric

- Soaplab
- Bio Services
- AMBIT Text Extraction Service
- EMBOSS

Applications
- Core services
- External services

Service Providers
- Bioinformaticians
- Tool Providers
Putting the services together

- Service
- Registry
- Knowledge Service
- FreeFluo Workflow enactment engine
- Distributed Query Processor
- Job Execution
- Notification Service
- Provenance browser
- User Proxy
- AMBIT Information Extraction Service
- mIR
- Service & workflow browser
- Find Component
- Registry View
- Service Publication
- syntactic registration
- Semantic registration
- Match maker
A work bench for demonstrating services

myView on the mIR

Workflow

Metadata about workflow

NetBeans

note about workflow

Customizer Dialog
This annotation pipeline workflow retrieves the EMBL accession number, Medline ID, OMIM ID and GO IDs from the molecular function, biological process and cellular component ontologies for a given Affymetrix probe set ID. The results for probe set set 2d, 1001 set are:

string []

probeSetId

EmbAccNumber

string

getEmblAccessionNumber

string

us

queryByEmblNumber

queryBy HGNC

ok
Notification service

- A new gene with changed expression in Graves’ Disease added to mIR
- User registers interest in notification topics
- Informs the user via a notification client in the workbench that new data has been added to the mIR.
- Notifications presented to the user with a client in the workbench environment.
Semantic discovery – services & workflows

- Services and workflows described using semantic web technologies and ontologies
- Selection by the types of inputs they use, outputs they produce, the bioinformatics tasks they perform…
- DAML+OIL → OWL
- RDF-based UDDI registry
- Multiple & 3rd party registries
- Multiple & 3rd party metadata
The mIR holds the experimental components

- We need to discover which workflows have been published that can operate on data of this specific semantic type (an Affymetrix probe set identifier)
- Some might be in mIR, some might be in global registry
- mIR holds all experimental components
- Multiple mIRs
-Built on RDMS & OGSA-DAI
- Plans: Federated architecture, LSIDs and RDF
Create and run a workflow

• If an appropriate workflow does not exist, a new one can be created in the Taverna editor
• Workflow & outputs stored in mIR
• Freefluo workflow enactment engine
• WSFL & Scufl
• Joint development with HGMP and EBI

http://sourceforge.net/projects/taverna/)
Provenance logging and reusing

- FreeFluo provides a detailed provenance record stored in the mIR describing what was done, with what services and when
- Can be viewed within the workbench
- XML document
- Every mIR object have (dublin core) provenance properties

**Workflow Provenance Data**

**Workflow overall summary**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>workflowInstanceID</td>
</tr>
<tr>
<td>Status</td>
<td>COMPLETE</td>
</tr>
<tr>
<td>Start Time</td>
<td>2003-06-11 00:14:06</td>
</tr>
<tr>
<td>End Time</td>
<td>2003-06-11 03:14:08</td>
</tr>
<tr>
<td>User</td>
<td><a href="mailto:user1@cs.man.ac.uk">user1@cs.man.ac.uk</a>/fotherleven</td>
</tr>
</tbody>
</table>

**Invoked Services Details**

<table>
<thead>
<tr>
<th>SOAPLAB SERVICE</th>
<th><a href="http://industry.ebi.ac.uk/soap/soaplab/nucleic_primers?primer=3">http://industry.ebi.ac.uk/soap/soaplab/nucleic_primers?primer=3</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>COMPLETE</td>
</tr>
<tr>
<td>Start Time</td>
<td>2003-06-11 00:14:06:968</td>
</tr>
<tr>
<td>End Time</td>
<td>2003-06-11 00:14:08:960</td>
</tr>
<tr>
<td>Soaplab status</td>
<td>0</td>
</tr>
</tbody>
</table>

Provenance is not just workflow

Derivation paths ~ workflows, queries
Annotations ~ notes
Evolution paths ~ workflow → workflow
• Wrap CORBA, Perl etc to look like web services, to become Grid services (eventually)

• SoapLab
  – A soap-based programmatic interface to command-line applications
  – ~300 different classes of services
  – Swiss-Prot, EMBOSS, Medline…

• 3rd parties
  – JEMBOSS, PathPort, bioMoby
Talisman application: using individual services

http://www.ebi.ac.uk/collab/mygrid/service1/talisman/index.html
The annotation pipeline to identify Genes of Interest

Annotation Pipeline

What is known about my candidate gene?

Look at contents of work bench

User notified of new Affy data

Run a workflow over new Affy data
  - Launch workflow wizard
  - Discover appropriate workflow
  - Enact workflow
  - Monitor workflow

Look at provenance

Select and view results
Status and plans

- Reflecting on what we have
  - All the components have an implementation in various states of maturity and functionality, some of which are downloadable already: Freefluo, Taverna, Soaplab.
  - Field evaluations with Uni. Newcastle Grave Disease geneticists and GSK with seeded data
- Expanding the user base
  - Use cases in Sleeping Cow
- Each component has plans, e.g.
  - More sophisticated model of provenance and other experimental data holdings, to store much more heavily linked metadata about provenance that will enable us to create views of the mIR along many axes.
  - The myGrid Information Repository to be significantly revised.
  - Review & Systematisation of type management
- Migration strategy to OGSA
Summary

- myGrid offers service based middleware components
- Open source and freely downloadable
- Open Grid Service Architecture-compliant
- Allows the scientist to be at the centre of the Grid -- Personalisation
- Generic middleware that suits the creation of bioinformatics applications
- Inclusion of rich semantics to facilitate the scientific process
- Available from http://www.mygrid.org.uk
Our Biology colleagues

Simon Pearce

Claire Jennings

Institute of Human Genetics
School of Clinical Medical Sciences
University of Newcastle
UK
The techy dudes

Matthew Addis, Nedim Alpdemir, Rich Cawley, [Vijay Dialani], Alvaro Fernandes, Justin Ferris, Rob Gaiauskas, Kevin Glover, Carole Goble (director), Chris Greenhalgh, Mark Greenwood, Ananth Krishna, Peter Li, Xiaojian Liu, Darren Marvin, Karon Mee, Simon Miles, Luc Moreau, Juri Papay, Norman Paton, Steve Pettifer, Milena Radenkovic, Peter Rice, [Angus Roberts], Alan Robinson, Martin Senger, Nick Sharman, Paul Watson, Anil Wipat & Chris Wroe.