Exploring Williams-Beuren Syndrome

Professor Carole Goble
http://www.mygrid.org.uk

myGrid is a UK e-Science project involving five UK universities, the European Bioinformatics Institute and many industrial collaborators. The myGrid project aims to exploit the growing interest in Grid technology, with an emphasis on the Information Grid, and provide middleware layers that make it appropriate for the immediate needs of bioinformatics. Specifically, my Grid is building high level services for data and application resource integration such as resource discovery, workflow enactment and distributed query processing. However, these services merely enable experiments to be formed and executed. Additional services are needed to support the e-based scientific method and best practice found at the bench but often neglected at the workstation, notably provenance management, change notification and personalisation.

In this presentation I will outline the objectives of myGrid, the services we have developed and show how myGrid components have been used to assist in the characterisation of unknown genes implicated in Williams-Beuren Syndrome
Acknowledgements

myGrid is an EPSRC funded UK eScience Program Pilot Project

Particular thanks to the other members of the Taverna project, http://taverna.sf.net

Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
Roadmap

- **myGrid in a nutshell**
- **Gene characterisation in Williams-Beuren Syndrome.**
- **Semantic Aspects**
  - Information model
  - Service discovery
  - Data Management - LSID
  - Metadata management for provenance – RDF
- **Lessons learnt and opportunities**

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
Experiment life cycle

Forming experiments

Personalisation

Managing lifecycle, provenance and results of experiments

Executing and monitoring experiments

Sharing services & experiments

Discovering and reusing experiments and resources

Forming experiments

Personalisation

Managing lifecycle, provenance and results of experiments

Executing and monitoring experiments

Sharing services & experiments

Discovering and reusing experiments and resources

Experiment life cycle

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
In a nutshell

- Bioinformatics toolkit
- Open (Web) Services
  - myGrid components
  - External domain services
  - No control or influence over service providers
- Open to third party metadata
- Open extensible architecture
  - Assemble your own components
  - Designed to work together
  - Toolkit
  - Axis/Apache based
  - RDF and DAML+OIL/OWL
  - Jena, OilEd, Instance Store & FaCT

Openness

- open source
- open world of services
- open extensible technology
- open to wider eScience context
- open to user feedback
- open to third party metadata

Collection of components for assembly
Pick and mix
Williams-Beuren Syndrome

- Microdeletion of 155 Mbases on Chromosome 7
- Hannah Tipney, May Tassabehji, Andy Brass, St Mary’s Hospital, Manchester, UK
- Characterise an unknown gene
- Annotation pipelines and Gene expression analysis
- Services from USA, Japan, various sites in UK

Kind of the same as what Robert said - we have found some interesting things were are going to follow up; but if I told you I'd have to kill you ; )

The results are biologically interesting mainly because they are correct! The problem I am trying to solve is mapping the hugely complex WBS phenotype to the genotype, but the genotype is incomplete due to the gaps so this kinda makes the whole process a bit tricky. So, the workflow is helping in my long and drawout efforts to produce a COMPLETE map of the WBS deleted region - once we have this finished map we will at least know exactly what WBS people are missing and so can start linking specific genes to aspects of the larger phenotype. And then the really interesting biology begins....!

Hannah’s aim is to generate a good, correct genetic map for the WBS critical region. So, she finds new bits of DNA for the region; then does all the standard bioinformatics.

Once a biogist has a genetic map they feel much more confident about knowing what is going on and designing experiment etc.

Williams-Beuren Syndrome (WBS) is a rare, sporadically occurring microdeletion disorder characterised by a unique set of physical and behavioural features [10]. WBS is caused by a 1.5Mb deletion [12] located in chromosome band 7q11.23 [3]. WBS is a complex, multisystem genetic disorder with an intricate phenotype. This encompasses a variety of symptoms [3, 11], including an unusual unequal cognitive profile combining expressive language, very weak visuospatial skills and poor fine motor ability [14]. The region commonly deleted in WBS is flanked by highly repetitive regions, 320-500 kb in length [13] containing both pseudogenes and genes. Most WBS individuals have a deletion of 1.5Mb, encompassing 24 genes [17] (see Figure 1). The deletion causes the haploinsufficiency for genes in the region which subsequently results in the WBS phenotypes [3]. A smaller region within the common WBS deleted region containing the genes whose absence are critical to the WBS phenotype has been identified [11, 18] (see Figure 1).

Many maps of the region have been published [2, 13, 19, 12], each with an increasing level of resolution and density...
Williams-Beuren Syndrome microdeletions reside on chromosome 7q11.23. Patients with deletions fall into two categories. Those with classic WBS (* indicates the common deletion) and those with SVAS but not WBS, caused by hemizygous deletion of the elastin gene. A physical map of the region composed of genomic clones is shown with a gap in the critical region. The myGrid software was used to continue the contig and identify more genes at this locus.
**Filling a genomic gap**

Two major steps:
- Extend into the gap: Similarity searches; RepeatMasker, BLAST
- Characterise the new sequence: NIX, Interpro, etc…

- Numerous web-based services (i.e. BLAST, RepeatMasker)
- Cutting and pasting between screens
- Large number of steps
- Frequently repeated – info now rapidly added to public databases
- Don’t always get results
- Time consuming
- Huge amount of interrelated data is produced – handled in lab book and files saved to local hard drive
- Mundane
- Much knowledge remains undocumented
- Bioinformatician does the analysis

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
Bioinformatics analyses typically involve visiting many data resources and analytical tools.
Most services aren’t ours
High level architecture

- Provenance and Data browser i.e. Haystack
- LSID Authority
- mIR
- Store Service
- Event Notification Service
- Freefluo Workflow Engine
- Taverna Workbench
- View Service
- UDDI
- Semantic Discovery & Registration
- Web services, local tools, User interaction etc.
WBS task

- Wrap services as web services
- Register them
- Build a workflow using the services
- Evolve the workflow
- Run it over and over again in case data has changed
- Record results & provenance
- Inspect and compare results & provenance
- Event notification, portal, 3rd party annotation…

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu.
User Results

Benchmark: Two iterations of workflows (1 day run)
  – Reduced gap by 267,693 bp at its centromeric end
  – Correctly located all seven known genes in this region
  – Identified 33 of the 36 known exons residing in this location

Manually: takes two days (+) including analysis
Now: takes 30 mins to produce results and half a day for analysis.
  • Less boring. Less prone to mistakes.
  • Once notification installed won’t even have to initiate it.
Where is the semantics
Information Model v2

- Resources and Identifiers
- People, teams and organizations
- Representing the e-science process
- Experimental methods for e-science
- Scientific data and the life-science identifier
  - Types
  - Identifier Types
  - Values and Documents
- Provenance information
- Annotation and Argumentation

Describing data sets etc
E-Science workshop part of Link-Uo meeting → May.
Semantic discovery

- The User does the choosing of services
- A common ontology is used to annotate and query any myGrid object including services.
- Ontology is built using DAML+OIL and reasoning
- Deployed as a static RDF graph
- Discover workflows and services described in the registry via Taverna.
- Look for all workflows that accept an input of semantic type nucleotide sequence.

Most of the domain services are not ours
Unreliable and changeable without notice
Some wrapped by us
Soaplab
Some are foreign WSDL
WSDL in the wild sucks.
Darn hard to describe
Labelling data items in databases
   Semantic typing for controlling inputs and outputs
   Use by distributed query processing
Linking & browsing XML-based myGrid information components
   COHSE
Work to link with the Life Science Identifier (I3C)
Generate BioMOBY Central service classification
Observations
Services

- Practically all the services are remote and third party
- Services are changeable and unreliable
- Redundant services are essential
- WSDL in the wild is poor
- Automated annotation

Registration – the ability to register services and workflows within Taverna so that others in the organisation know they exist.
Can you guess what it is yet?

```xml
<?xml version='1.0' encoding='UTF-8'?>
<definitions name='Blast'>
  <message name='execute0In'>
    <part name='accession' type='xsd:string'/>
  </message>
  <message name='execute0Out'>
    <part name='Result' type='xsd:string'/>
  </message>
  <portType name='Blast'>
    <operation name='execute' parameterOrder='accession'>
      <documentation>Execute Blast</documentation>
      <input name='execute0In' message='tns:execute0In'/>
      <output name='execute0Out' message='tns:execute0Out'/>
    </operation>
  </portType>
  <soap:binding style='rpc'
    transport='http://schemas.xmlsoap.org/soap/http'/>
  <soap:operation name='execute'>
    <soap:operation soapAction='execute' style='rpc'/>
    <input name='execute0In'>
      <soap:body use='encoded'/>
    </input>
    <output name='execute0Out'>
      <soap:body use='encoded'/>
    </output>
  </soap:operation>
</definitions>
```
Service then becomes an organisational layer. We want to share descriptions, but not much else.

It’s a unit of publishing rather than a unit of functionality.

I think it was all the fields requiring a controlled vocabulary / ontology.
Control flow, iteration and data flow
Data sets and nested flows
Configurable failure handling
Incorporated Life Science Id resolution
Provenance and status reporting
Type and data management
Plug-ins
User notification
Data entry wizard

Libraries of SHIM services
Libraries of workflows

To retrieve newly submitted human genomic sequences that extend into the gap. Similarity searches are made against a range of GenBank databases using the BLAST programme BLASTN [1]. Repeat Masker2 was used to search against RepBase Update 6.3 [8] to avoid spurious multiple hits against repetitive sequences.

Any high-scoring matches from human chromosome 7 are submitted to Genscan3, which is used to find any gene(s) residing on those new fragments. The manual protocol for these analyses uses the NIX programme4, that collects together a wide range of gene prediction and DNA characterisation tools. We also aim to use this programme within myGrid. To characterise any gene, surrounding regions and any putative gene product, the genomic DNA is analysed for a full range of motifs and features and translated in all six frames. The most suitable reading frame is used in a similarity search against protein databases and submitted to a standard collection of characterisation tools (see Figure 2 for details).

This workflow takes the last verified piece of sequence (<3000 bp) in the contig flanking a gapped region and produces a shortlist of sequences which may extend the contig into the gap region. The query sequence is masked using RepeatMasker to prevent spurious hits prior to being used by the NCBI BLASTN program to identify overlapping sequences. Only new or improved hits are relevant, and so the results are first translated into a simplified format using 'SIMPLIFIER', before being compared to the results of the previous run using the 'COMPARER' service. 'RETRIEVE' takes new hits and determines which are located on human chromosome 7 before returning those sequences in FASTA format. Intermediate results are kept in case the filtering operation has excluded relevant but miss-annotated sequences.
**Results management**

- Automated workflows produce lots of heterogeneous data
- These are just some of the results from one workflow run for Williams Disease

<table>
<thead>
<tr>
<th>type</th>
<th>size</th>
<th>type</th>
<th>created</th>
<th>modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>74KB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>3</td>
<td>74KB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>4</td>
<td>1.2MB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>5</td>
<td>1.2MB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>6</td>
<td>0KB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>7</td>
<td>0KB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>8</td>
<td>298KB</td>
<td>File</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>9</td>
<td>13KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>10</td>
<td>13KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>11</td>
<td>57KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>12</td>
<td>57KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>13</td>
<td>30KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>14</td>
<td>13KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>15</td>
<td>18KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>16</td>
<td>18KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>17</td>
<td>13KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>18</td>
<td>13KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>19</td>
<td>75KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>20</td>
<td>75KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>21</td>
<td>44KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>22</td>
<td>44KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>23</td>
<td>44KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>26</td>
<td>74KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>27</td>
<td>74KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>28</td>
<td>4KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>29</td>
<td>4KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>30</td>
<td>4KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>31</td>
<td>4KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>32</td>
<td>348KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>33</td>
<td>348KB</td>
<td>Text Document</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
<tr>
<td>34</td>
<td>9KB</td>
<td>Dir</td>
<td>2004-06-10</td>
<td>2004-06-10</td>
</tr>
</tbody>
</table>
Amplification

One input

Many outputs
Manually: Do analysis as perform experiment
Workflow: Do analysis at end of experiment
Therefore need good result co-ordination for back-tracking
Intermediate Results

- Workflows change the way the bioinformatican works
- Before: analyse results as go along
- After: all results in one go
- So linking intermediate results important
Life Science IDs

- LSID provides a uniform naming scheme.
- LSID Resolver guarantees to resolve to same data object.
- LSID Authority dishes them out. Also returns metadata of object.
- Used throughout myGrid as an object naming device.
- myGrid Repository acts as an LSID Authority
- LSID allows universal access to results for collaboration, as well as for review.
- RDF+LSID explains the context of results, and provides guidance for further investigations.

I3C / IBM / EBI proposal for a Life Science Identifier

Pioneered by myGrid

Each data item used/produced by the workflow is assigned an lsid automatically and the data published to a repository which acts as an LSID authority.

Data items can then be browsed in Internet Explorer or Haystack. Haystack provides additional project wide metadata.

Shouldn’t be exposed to the user!
Process Provenance

---

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
Who, what, where, why, when, (w)how?
The tracability of knowledge as it evolves and as it is derived.
Identity – the Life Sciences ID
The Lab Book. Methods in papers.
Immutable Metadata
Migration – travels with its data but may not be stored with it.
Aggregates as data aggregates
Private vs Shared provenance records.
Ownership => success -> being sued?
Credit.
IBM’s BioHaystack

GenBank record

Portion of the Web of provenance

Managing collection of sequences for review

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
Observations

- Managed the transition from generic middleware development to practical day to day useful services
- Real users (plural) fundamental to that
- End to end support for an entire scenario
- Bury the semantics
- Show stoppers for practical adoption are not technical showstoppers
  - Can I incorporate my favourite service?
  - Can I manage the results?
- By tapping into (defacto) standards and communities we can leverage others results and tools – LSID, Haystack, Pedro.
myGrid is an EPSRC funded UK eScience Program Pilot Project

Particular thanks to the other members of the Taverna project, http://taverna.sf.net
myGrid People

Core

Users
- Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK
- Hannah Tipney, May Tassabehji, Andy Brass, St Mary's Hospital, Manchester, UK

Postgraduates
- Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, John Dickman, Keith Flanagan, Antoon Goderis, Tracy Craddock, Alastair Hampshire

Industrial
- Dennis Quan, Sean Martin, Michael Niemi, Syd Chapman (IBM)
- Robin McEntire (GSK)

Collaborators
- Keith Decker

GGF11 Semantic Grid Applications Workshop,
Hilton Hawaiian Village Beach Resort & Spa, Honolulu, Thursday June 10, 2004
http://www.mygrid.org.uk