

# CONTROLLING THE CHAOS: DEVELOPING POST-GENOMIC GRID INFRASTRUCTURES

DR RICHARD SINNOTT, DR MICHA BAYER<sup>†</sup>

*National e-Science Centre // Bioinformatics Research Centre,  
University of Glasgow, Glasgow G12 8QQ, Scotland  
Email: ros@dcs.gla.ac.uk*

“Why does Scotland have one of the highest rates of heart attacks in Europe? Are there genetic factors which contribute to this statistic”? The analysis and exploration of a broad array of life science data sets are needed to answer such questions. The Grid provides, at least conceptually, one way in which these kinds of data sets can be linked and analyzed. The life science domain places specific requirements on the Grid infrastructure needed to answer such questions. In this paper we describe these requirements and outline how they are being addressed in the DTI funded BRIDGES project.

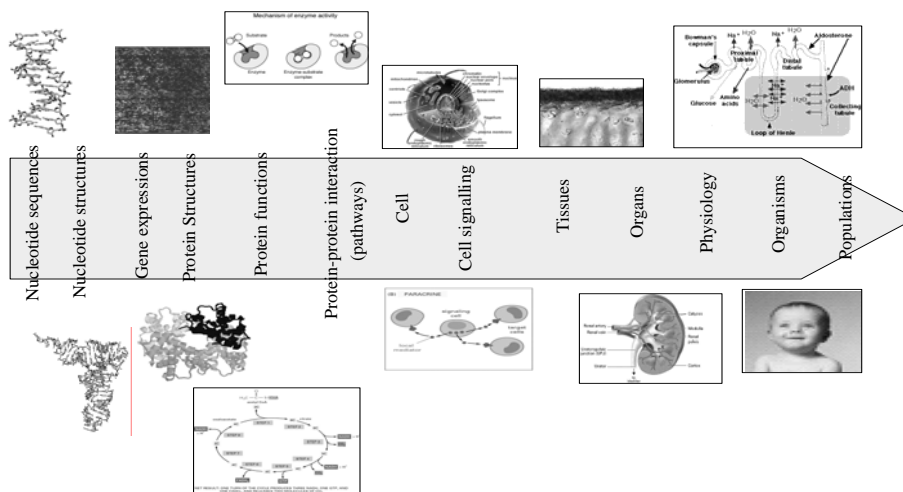
## 1. INTRODUCTION

The life science community is experiencing a period of unprecedented change, challenge and opportunity. With the completion of the sequencing of the human genome (and ever increasing numbers of other genomes), the opportunities of *in-silico* scientific research offer a new horizon of possibilities: from rapid targeted drug discovery, identification of genetic factors to disease causes and epidemiological studies through to complete biological understanding of complete organisms and tailored genetic treatments supporting e-Health solutions. The possibilities abound! Fundamental to the realization of this vision is the infrastructure needed to use and understand the vast array of data sets associated with such research as depicted in Figure 1. These data sets are growing exponentially, have radically different characteristics, are often maintained by completely different groups and bodies, and importantly are perpetually evolving. In this context, the development of an infrastructure that allows to access, use, and

---

<sup>†</sup> Work supported by DTI grant to BRIDGES project.

analyze such changing and growing amounts of data is both technically challenging, offers huge benefits to the scientific community and is potentially extremely viable commercially.



**Fig. 1:** Spectrum of Life Science Data

To support the vision of e-Health it is clear that computational infrastructures must address (at least) the following requirements:

- tools that allow simplified access to and usage of the potentially complex data structures that comprise life science data sets;
- provide access to large scale computational resources needed to process and search the life science data sets, e.g. when comparing genomes;
- ensure that appropriate security mechanisms are in place to deal with the data sets and infrastructure upon which they exist;
- make this infrastructure easy to use and ideally targeted towards the needs of the specific scientific groups.

In the following sections we shall see how the BRIDGES project<sup>1</sup> is realizing these requirements through the development of a state of the art Grid infrastructure.

## 2. BIOMEDICAL RESEARCH INFORMATICS DELIVERED BY GRID ENABLED SERVICES

The Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) project<sup>1</sup> has been funded by the UK Department of Trade and Industry to develop a computational infrastructure to support the needs of the Wellcome Trust funded (£4.34M) Cardiovascular Functional Genomics (CFG) project<sup>2</sup>. The CFG consortium is investigating possible genetic causes of hypertension, one of the main causes of cardiovascular mortality. This consortium which involves five UK and one Dutch site (depicted in Figure 2) is pursuing a strategy combining studies on rodent models of disease (mouse and rat) contemporaneously with studies of patients and population DNA collections.

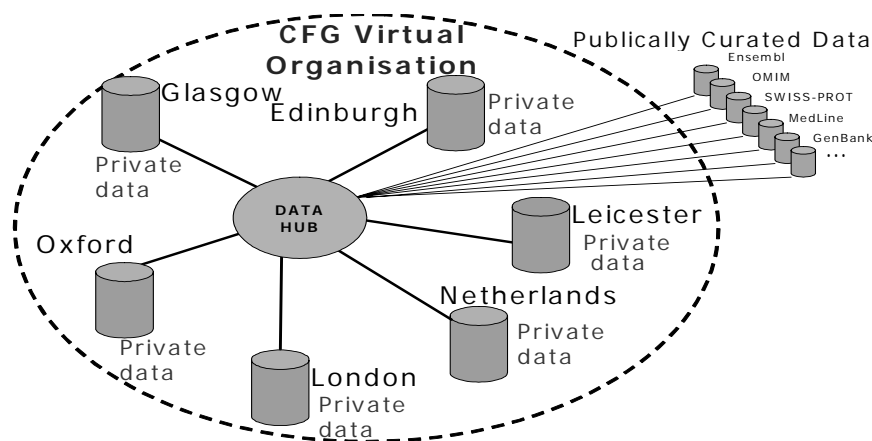


Fig. 2: Data Distribution and Security of CFG Partners

Currently many of the activities that the CFG scientists undertake in performing their research are done in a time consuming and largely non-automated manner. This is typified through “internet hopping” between numerous life science data sources. For example, a scientist might run a microarray experiment and identify a gene (or more likely set of genes) being differentially expressed. This gene is then used as the basis for querying a remote data source (e.g. MGI in Jackson<sup>3</sup>). Information retrieved from this query might include a link to another remote data source, e.g. on who has published a paper on this particular gene in MedLine<sup>4</sup> or PubMed<sup>5</sup>. Information from these repositories might include links to Ensembl<sup>6</sup>

where further information on this gene, e.g. its start and end position in a given chromosome can be established. Such sequences of navigations typify the research undertaken by scientists.

## **2.1. Simplified Access To and Targeted Usage of Life Science Data Sets**

A key component of the architecture in Figure 2 is the Data Hub. This represents both a local data repository, together with data made available via externally linked data sets. These data sets exist in different heterogeneous, remote locations with differing security requirements. Some data resources are held publicly (e.g. genome databases such as Ensembl<sup>6</sup>, gene function databases such as OMIM<sup>7</sup> and relevant publications databases such as MedLine<sup>4</sup>); whilst others are for usage only by specific CFG project partners (e.g. microarray data sets<sup>8</sup> or quantitative trait loci (QTL) data sets<sup>9</sup>).

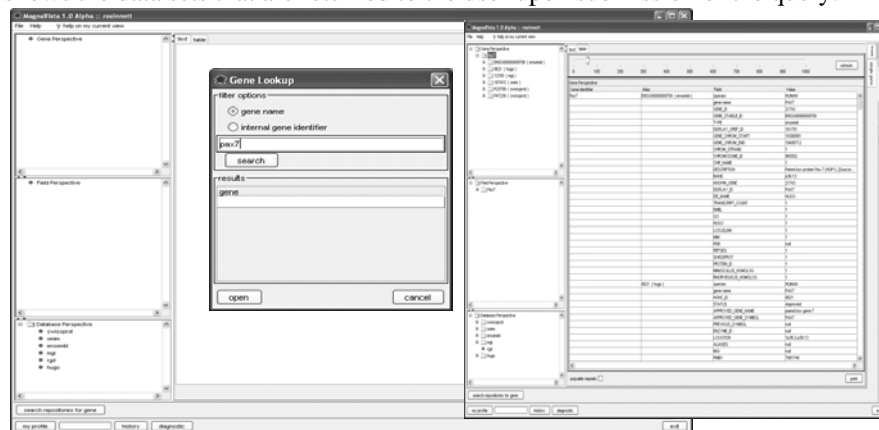
Currently the public data sets are accessible via two different technologies: IBM's Information Integrator (IBM-II) – formerly known as DiscoveryLink<sup>10</sup> (and soon to be known as Masala), and the Open Grid Service Architecture – Data Access and Integration (OGSA-DAI) technology<sup>37</sup>. IBM-II technology has been developed to meet the challenge of integrating and analyzing large quantities of diverse scientific data from a variety of life sciences domains offering single-query access to existing databases, applications and search engines. This is achieved through wrappers which use the data source's own client-server mechanism to interact with the sources in their native dialect. Through IBM-II access to a broad array of heterogeneous data sets can be achieved, e.g. in relational databases, XML databases, Excel spreadsheets, flat files etc.

In a similar vein, the OGSA-DAI technology provides a generic data access and integration mechanism overcoming issues related to the heterogeneity of technologies and data sets as well as the remoteness of data sets themselves. This technology is being continually refined and extended to be compliant with on-going Grid standardization efforts. At the time of writing, initial experiments are on-going in performing performance benchmarks against these two solutions for access to and usage of life science data.

To support the life science community it is essential that applications are developed that allow them simplified access to life science data sets as well as to personalize their environments. The personalization might well include the data

sources that are of most interest to the scientists and the information that they are most interested in from those data sources.

To support such personalization the BRIDGES project has developed the application MagnaVista<sup>1</sup>. This application provides a completely configurable environment through which the scientists can navigate to and access a broad array of life science data sets of relevance to their research. The basic user interface to MagnaVista is depicted in Figure 3. Here the user can include the genes that they are most interested in (central pop up window). The lower left corner of Figure 3 lists the remote data sources that are accessible (SWISS-PROT<sup>11</sup>, MGI<sup>3</sup>, Ensembl<sup>6</sup> (rat, mouse, human DBs), RGD<sup>12</sup>, OMIM<sup>7</sup>). The pop up window to the right of Fig. 3 shows the data sets that are returned to the user upon submission of the query.



**Fig. 3:** MagnaVista Basic Usage for Gene Query

Thus rather than the user manually hopping to each of these remote resources, a single query is used to deliver collections of data associated with the genes of interest. To support the specific targeted data needs of the scientists, the MagnaVista application can be personalized in various ways. It currently supports user based selection of specific (remote) databases that should be interrogated; user based selection of the various data sets (fields) that should be returned from those databases; storage of specific genes of interest, as well as personalization of the look and feel of the application itself.

## 2.2. Tools Supporting Cognitive Aspects of Complex Data Sets

Life science data sets are notoriously complex, requiring a great deal of expertise to understand and utilise. Tools that facilitate cognitive understanding of these data sets, e.g. through visualisation are essential. One such relation that is especially important when dealing with translational studies between different genomes is *synteny*. Synteny is the condition of two or more genes being located on the same chromosome. Of particular importance is *conserved synteny* which may be defined as the condition where a syntenic group of genes from one species have orthologues and/or homologues in another species, i.e. similar sets of genes where the similarity itself can be ascertained through a combination of approaches such as protein sequence similarity, structure, function etc.

The analysis of conserved synteny between the different organisms (e.g. mouse, rat and human), in combination with quantitative trait loci (QTL) data<sup>9</sup> and microarray experiments<sup>8</sup>, is one of the main methods used by the CFG scientists in investigating hypertension. Their aim is to discover genes responsible for hypertension in rat or mouse organisms and translate these findings into knowledge about the mechanisms for hypertension in human. It should be noted that knowledge of syntenic relationships and of known QTLs between organisms provides supporting, but not necessarily guaranteed, evidence about the location and functional role of candidate genes causing hypertension between species. In displaying conserved synteny, two (or more!) chromosomes need to be shown simultaneously. SyntenyVista was developed for this purpose as shown in Fig 4.

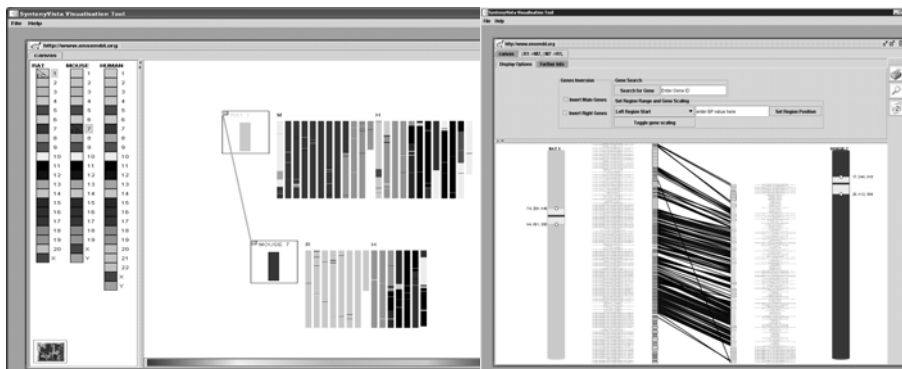


Fig. 4: Grid Enabled Syntenic Visualization Tool

The three columns of shaded boxes on the left of Figure 4 are the sets of chromosomes for the rat, mouse and human genomes. Users of SyntenyVista are able to drag these individual colour coded chromosomes onto the pallet where the QTL information is depicted. This represents the level of syntenic relations between rat, mouse and human chromosomes. In this example, the rat 1 and mouse 7 chromosomes are being visualized since they share a high degree of synteny. Users are then able to show detailed information on the relationships between the many thousands of genes in these two chromosomes as depicted on the right of Figure 4. Further information is also available to the users when scrolling through these chromosomes. For example, users are able to pan out to visualize the complete chromosomes, pan in to see individual relationships between specific genes, find where specific genes start and end on the chromosome, and gain access to detailed information associated with these genes.

### **2.3. Services for Simplified Usage of Large Scale Computational Resources**

In their pursuit of novel genes and understanding their associated function life scientists often require access to large scale compute facilities to analyse their data sets, e.g. in performing large scale sequence comparisons or cross-correlations between large biological data sources. The Basic Local Alignment Search Tool (BLAST)<sup>13</sup> has been developed to perform this function. Numerous versions of BLAST currently exist which are targeted towards different sequence data sets and offer various levels of performance and accuracy metrics. BLAST involves sequence similarity searches, often on a very large scale, with query sequences being compared to several million target sequences to compute alignments of nucleic acid or protein sequences with the goal of finding the  $n$  closest matches in a target data set. BLAST takes a heuristic (rule-of-thumb) approach to a computationally highly intensive problem and is one of the fastest sequence comparison algorithms available.

There are a number of public sites<sup>14,15</sup> that provide users with web based access to BLAST, and these generally use dedicated clusters to provide the service. However, the growth of sequence data over the last decade has been exponential<sup>16</sup>, and consequently searches take increasingly longer. Given that most biologists use these public sites rather than running the computation on their own machines (BLAST is freely available as a command line tool), the load on the purpose built clusters has increased dramatically and now significant queuing times are becoming

increasingly common. A typical use of BLAST will usually involve searching against the *nt* database from NCBI<sup>17</sup> - a data set that contains several of the main sequence repositories and is currently ~3 GB in size (with over 2,2M target sequences).

Usage of BLAST on large scale HPC resources is often non-trivial to achieve and typically requires knowledge of scripting languages (for decomposing the input data sets and recomposing/merging the results data) and local job schedulers. Users should not have to learn the often complex options associated with job submission to job schedulers such as Condor<sup>18</sup> or OpenPBS<sup>19</sup>. In addition, one of the primary benefits of Grid technology is the ability to dynamically select and use a variety of heterogeneous resources is essential. This in turn requires that meta-schedulers are available that can dynamically schedule jobs across a variety of heterogeneous resources utilising a variety of local job schedulers. The BRIDGES Grid BLAST service which provides such a simplified BLAST based job submission system, enabling access to and usage a collection of HPC facilities is shown on the left of Fig. 5.

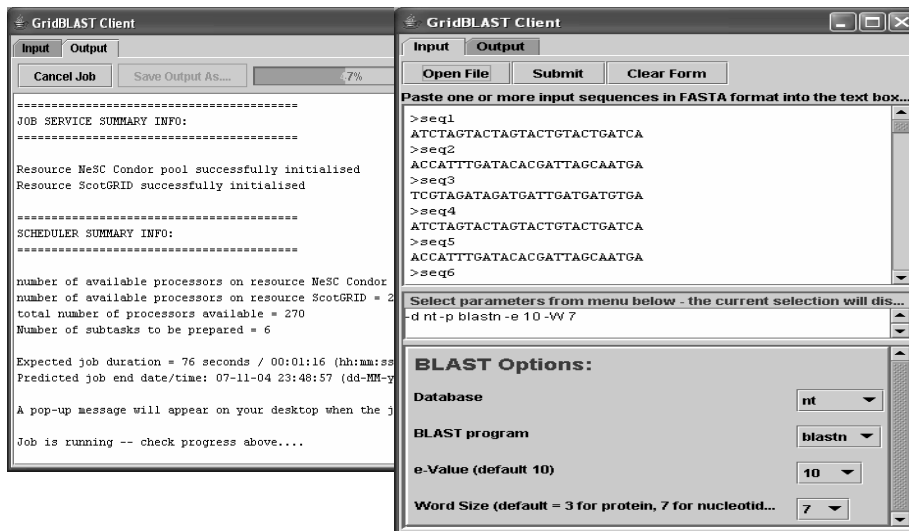


Fig. 5: Grid enabled BLAST service and Monitoring its Usage

An explicit requirement in designing this service was that it should be extensible. This has been achieved through XML based resource configuration files which easily allow new sets of resources to be added and subsequently used. Currently this Globus toolkit version 3<sup>20</sup> service provides access to the ScotGrid computational resource<sup>21</sup> and a Condor pool at the University of Glasgow. The ScotGrid resource is *the* e-Science resource at the University of Glasgow and represents a consolidation of resources across a variety of research groups and departments. It consists of the equivalent of 255 1GHz processors (using hyper-threading) and 15TB disk space comprised of IBM xSeries, Blade server, FASrT500 and Dell and Cisco technologies. It uses the Maui scheduling software<sup>22</sup> which implements the scheduling policies for the OpenPBS<sup>19</sup> batch submission system.

This service provides intelligent default settings for a variety of BLAST services. When used, the service checks what resources are available, where the jobs are best run and subsequently provides a prediction of how long the complete BLAST job will take to complete. In addition, monitoring of the status of the various sub-jobs is undertaken and staging of the various input and output files onto the compute resources is provided. This is indicated on the right of Figure 5. At the time of writing the BLAST service is being extended to make use of the UK e-Science National Grid Service<sup>41</sup>.

## **2.4. Supporting a Scalable and Fine Grained Security Infrastructure**

The widespread acceptance and uptake of Grid technology can only be achieved if it can be ensured that the security mechanisms needed to support Grid based collaborations are at least as strong as local security mechanisms. The predominant way in which security is currently addressed in the Grid community is through Public Key Infrastructures (PKI)<sup>23</sup> to support *authentication*. Whilst PKIs address user identity issues, authentication does not provide fine grained control over what users are allowed to do on remote resources (*authorization*). The Grid community has put forward numerous software proposals for authorization infrastructures<sup>24-26</sup>. It is clear that for the foreseeable future a collection of solutions will be the norm.

Key concepts associated with authorization are a Policy Enforcement Point (PEP) and a Policy Decision Point (PDP). The PEP ensures that all requests to access the target are authorized through checking with the PDP. The PDP's authorization decision policy is often represented through collections of rules

(policies). The different authorization infrastructures associated with Grid technology have put forward their own mechanisms for realizing PEPs and PDPs. Recently however the Grid community has put forward a generic API – the SAML AuthZ API<sup>27</sup>. This is an enhanced profile of the OASIS Security Assertion Markup Language v1.1<sup>28</sup>.

The OASIS SAML AuthZ specification defines a message exchange between a policy enforcement point (PEP) and a policy decision point (PDP) consisting of an *AuthorizationDecisionQuery* flowing from the PEP to the PDP, with an assertion returned containing some number of *AuthorizationDecisionStatements*. The *AuthorizationDecisionQuery* itself consists of: a *Subject* element containing a *NameIdentifier* specifying the initiator identity; a *Resource* element specifying the resource to which the request to be authorized is being made, and one or more *Action* elements specifying the actions being requested on the resources.

The GGF SAML profile specifies a *SimpleAuthorizationDecisionStatement* (essentially a granted/denied Boolean) and an *ExtendedAuthorizationDecisionQuery* that allows the PEP to specify whether the simple or full authorization decision is to be returned. In addition the GGF query supports both the pull and push modes of operation for the PDP to obtain attribute certificates, and has added a *SubjectAttributeReferenceAdvice* element to allow the PEP to inform the PDP where it may obtain the subject's attribute certificates from.

Through this SAML AuthZ API, a generic PEP can be achieved which can be associated with arbitrary (GT3.3+) Grid services\*. Thus rather than developers having to explicitly engineer a PEP on a per application basis, the information contained within the deployment descriptor file (.wsdd) when the service is deployed within the container, is used. Authorization checks on users attempting to invoke "*methods*" associated with this service are then made using the information in the .wsdd file and the contents of the LDAP repository (PDP) together with the DN of the user themselves. Note that this "method" authorization basis extends current security mechanisms such as GSI which work on a per service/container basis. This generic solution can be applied to numerous infrastructures used to realize PDPs such as PERMIS.

---

\* It has been stated by the Globus team (V. Welch, J. Schopf) that this API will be supported in future versions of the Globus toolkit.

PERMIS provides a Role Based Access Control (RBAC) infrastructure. RBAC models have been designed to make access control manageable and scalable<sup>29</sup>. PERMIS provides a standards-based Java API that allows developers of resource gateways to enquire if a particular access to the resource should be allowed. PERMIS provides tools that allow creation of XML based policies defining rules, specifying which access control decisions are to be made for given resources, e.g. definitions of subjects that can be assigned roles; definitions of Source of Authority (SOA) - trusted to assign roles to subjects; definitions of roles and their hierarchical relationships; definitions of what roles can be assigned to which subjects; definitions of targets that are governed by the policy, and the conditions under which a subject can be granted access.

Both PERMIS and the Globus toolkit version 3.3 (GT3.3) have been extended to support the SAML AuthZ API. PERMIS tools such as the Policy Editor and Privilege Allocator have been applied to create XML based policies which allow restricted access to the compute and data resources to the various CFG scientist roles.

The work on BRIDGES has applied this authorization infrastructure to restrict (authorize) access to specific data sets within the federated repository and to the specific compute resources that are accessible via the BLAST service. The current resource specific policy that is supported is based upon three key roles depending upon whether the user has a valid UK e-Science certificate; and/or whether the user has a local account on the HPC facility (ScotGrid) at Glasgow. If neither of these conditions is true, then the user may only perform a BLAST job on the freely available Condor pool at Glasgow University. This has been demonstrated to work, however the scalability of such low level policies is an issue that must be resolved.

## **2.5. Portal Technologies and Simplified Delivery Mechanisms**

There are various possibilities available for hosting the services to be made available to the CFG scientists. Given that user friendliness is a key aspect, development of a project portal was made. This portal provides a personalizable environment that the scientist is offered to explore all of the (Grid related) software, data resources and general information associated with the BRIDGES, and hence the CFG projects.

Arguably the most mature portal technology on the market and the market leader is IBM WebSphere Portal Server<sup>30</sup>, which has been used to develop the BRIDGES portal, although we note that several other solutions were also investigated, including GridSphere<sup>31</sup> and the Commodity Grid toolkit<sup>32</sup>. WebSphere Portal Server runs as another layer on top of the highly developed WebSphere Application Server. Since this provides a fully functional enterprise Java hosting environment it is possible to deploy a Java based Grid service instance within the same virtual machine container.

The BRIDGES portal itself provides an integrated and personalizable environment through which the scientists have access to the various Grid services that they need. This includes the MagnaVista service, the SyntenVista service, the Grid BLAST service and other services allowing the scientists to store and share a variety of bioinformatics data sets, including Quantitative Trait Loci (QTL) and microarray data sets (based upon the MIAME compliant services). Depending upon their role within the project, the personalization of the portal to the scientist is based upon secured (authorized) profiles accessible via the GGF SAML AuthZ API.

Simplified delivery mechanisms are crucial to ensure the success of Grid based technologies to the wider community. It is infeasible to expect non-computer scientists to have to deal with software deployment aspects related to the set up and configuration of the complex infrastructures associated with Grid technology. To address this issue, one mechanism that has been successfully explored within the BRIDGES project is Sun's WebStart technology<sup>33</sup>. Through this technology, users require only a browser to gain access to the various services. The portal provides *launch* buttons which when selected by the end user, check whether WebStart technology exist on the remote (end user) system. If this is not the case the user is prompted if they want to install this and if so, it is automatically installed along with the application itself. WebStart also allows easy handling of changes and updates to the Grid services available from the portal itself through providing checks on the latest version of the applications available on the end user systems.

### 3. CONCLUSIONS

The BRIDGES project began in October 2003 and investigated a wide variety of Grid technologies applicable to the life science domain. The current implementation status has provided a proof of concept prototype. Grid technologies *do* allow for simplified access to and usage of a broad set of post-genomic data sets – bringing

the data to the scientist! Services to support the analysis and visualization of these large life science data sets, efficiently utilizing HPC facilities have also been realized, taking into consideration appropriate security mechanisms deemed applicable. In short, it works! The work has not been without issues. The stability of the Grid middleware such as the Globus toolkit and the associated documentation remains below an appropriate level to easily produce Grid based systems. Compromises had to be made between the architecture and design, and the final systems that have been implemented due to for example, operating system dependencies of the middleware.

The work is evolving based upon feedback from usage of the infrastructure by the CFG scientists. Close liaison with the scientific community is essential to ensure that we are developing the “right software” and accessing the right data sets. Given that the CFG project are primarily interested in functional genomic based data resources, e.g. in supporting their microarray experiments, a bioinformatics workbench that allows the scientists to take up/down regulated gene names from microarray experiments and garner further information are of special interest. We note that the data sets accessible via the Data Hub are not a fixed set. Other resources can easily be added. However one of the challenges in this area is the issue in gaining access to these remote resources. For example, few of these resources provide the necessary programmatic access needed, i.e. to query their database directly. Instead, they often only offer compressed files that can be downloaded. As a result, the Data Hub includes a local warehouse for downloaded data sets. Currently federated access to Ensembl (rat, mouse, human) and MGI is supported, with the other data sets having to be locally warehoused. This then requires that local (downloaded) data sets are kept up to date with the remote data sets.

In our experience it is often non-trivial to establish a local repository that uses these data sets, even with the local downloading of the data. Thus for example, the data providers often do not provide schemas or data models for the data themselves. Rather they simply provide large compressed text files that can be ftp'ed. It requires a significant amount of effort to understand how this data can be input into a database, i.e. in working out what associated the data model/schema is. To address this numerous funding councils in the UK (MRC, BBSRC, NERC, JISC, DTI, Wellcome Trust) have funded the Joint Data Standards Survey project<sup>34</sup>. This is investigating the technical, social, political and often ethical issues that are currently

prohibiting the effective sharing of life science data sets, with the idea being that policies can be formulated to facilitate life science data sharing.

Underpinning data sharing is of course agreeing upon (standardizing) data models and technologies used to access these resources. The Grid community<sup>35</sup> is currently defining standards<sup>36</sup> for access to data on the Grid. The OGSA-DAI project<sup>37</sup> in particular has helped to shape this work and has produced Grid based implementations. A report comparing OGSA-DAI and IBM Information Integrator is currently under production.

Access to a broad range of life science data sets is essential if the vision of a future e-Health infrastructure is to be achieved. As described, the BRIDGES project has focused on the left hand side of Figure 1: functional genomics. Extending this to incorporate a wider variety of data resources and incorporating further life science applications is an on-going process. Two new projects at Glasgow in particular will allow for this process to be explored. The 4-year Scottish Bioinformatics Research Network (SBRN) project<sup>38</sup> will allow to both expand the number of applications available within BRIDGES, and to bring the existing prototypes to a more robust, production quality level. The 3-year Virtual Organisations for Trials and Epidemiological Studies (VOTES) project<sup>39</sup> is exploring how Grid based technologies can be applied in the clinical trials domain. Clinical trials require up to date information on patterns of disease/frequency of clinical procedures. VOTES will explore three aspects of clinical trials: recruitment of patients; data collection on those patients; and tools that facilitate the management of a clinical trial. Since the information that underpins clinical trials is located in a range of highly secure sites such as GP/doctor databases, hospital databases, clinical registries, death registries etc., exploring the applicability of the Grid in this domain will require extremely rigorous security and ethical considerations to be visibly supported. Linkage of such data sets with other genomics related data sets *a la* BRIDGES is necessary if e-Health is to be supported however, and to answer the opening question in the abstract!

### **3.1. Acknowledgements**

This work was supported by a grant from the Department of Trade and Industry. The authors would also like to thank members of the BRIDGES and CFG team including Prof. David Gilbert, Prof Malcolm Atkinson, Dr Dave Berry, Dr Ela Hunt and Dr Neil Hanlon. Drs Hanlon and Hunt are also acknowledged for their contribution to the original SyntenyVista software. Magnus Ferrier is acknowledged

for his contribution to the MagnaVista software and Derek Houghton for his work in developing the data repository. Acknowledgements are also given to the IBM collaborators on BRIDGES including Dr Andy Knox, Dr Colin Henderson and Dr David White. The CFG project is supported by a grant from the Wellcome Trust foundation.

#### 4. REFERENCES

1. BioMedical Research Informatics Delivered by Grid Enabled Services (BRIDGES), [www.brc.dcs.gla.ac.uk/projects/bridges](http://www.brc.dcs.gla.ac.uk/projects/bridges)
2. Cardiovascular Functional Genomics project, <http://www.brc.dcs.gla.ac.uk/projects/cfg/>
3. Mouse Genome Informatics (MGI), <http://www.informatics.jax.org/>
4. US National Library of Medicine, <http://www.nlm.nih.gov/>
5. PubMed Central Home, <http://www.pubmedcentral.nih.gov/>
6. EMBL-EBI European Bioinformatics Institute, <http://www.ebi.ac.uk/>
7. NCBI Online Mendelian Inheritance in Man, <http://www.ncbi.nlm.nih.gov/OMIM/>
8. Minimal Information About a Microarray Experiment (MIAME) <http://www.mged.org/Workgroups/MIAME/miame.html>
9. An Overview of Methods for Quantitative Trait Loci (QTL) Mapping, Lab of Statistical Genetics, Hallym University [http://bric.postech.ac.kr/webzine/content/review/indivi/2002/Aug/1\\_08\\_index.html](http://bric.postech.ac.kr/webzine/content/review/indivi/2002/Aug/1_08_index.html)
10. IBM Information Integrator, [http://www3.ibm.com/solutions/lifesciences/solutions/Information\\_Integrator.html](http://www3.ibm.com/solutions/lifesciences/solutions/Information_Integrator.html)
11. SWISS-PROT, <http://us.expasy.org/sprot/>
12. Rat Genome Database, <http://rgd.mcw.edu/>
13. Basic Local Alignment Search Tool (BLAST), <http://www.ncbi.nlm.nih.gov/Tools/>
14. EBI BLAST website, <http://www.ebi.ac.uk/blastall/index.html>
15. NCBI BLAST website, <http://www.ncbi.nlm.nih.gov/BLAST/>
16. GenBank statistics web page, <http://www.ncbi.nih.gov/Genbank/genbankstats.html>
17. NCBI Nucleotide database, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>
18. Condor website, <http://www.cs.wisc.edu/condor/>
19. Open Portable Batch System, [www.openpbs.org](http://www.openpbs.org)
20. Globus toolkit, [www.globus.org/toolkit](http://www.globus.org/toolkit)
21. ScotGrid, [www.scotgrid.ac.uk](http://www.scotgrid.ac.uk)
22. Maui Job Scheduler, <http://www.supercluster.org/maui/>

23. C Adams and S Lloyd (1999), Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations, Macmillan Technical Publishing.
24. Privilege and Role Management Infrastructure Standards Validation project [www.permis.org](http://www.permis.org)
25. L Pearlman, et al., A Community Authorisation Service for Group Collaboration, in Proceedings of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks. 2002.
26. M Thompson, et al., Certificate-Based Access Control for Widely Distributed Resources, in Proc 8th Usenix Security Symposium. 1999: Washington, D.C.
27. V. Welch, F. Siebenlist, D. Chadwick, S. Meder, L. Pearlman, Use of SAML for OGSA Authorization, June 2004, <https://forge.gridforum.org/projects/ogsa-Authz>
28. OASIS. Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) v1.1, 2 September 2003, <http://www.oasis-open.org/committees/security/>
29. D. Chadwick and A. Otenko. The PERMIS X.509 role based privilege management infrastructure, in Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies, Monterey, California, USA. 2002.
30. WebSphere Portal software, <http://www-306.ibm.com/software/genservers/portal/>
31. GridSphere Portal, [www.gridisphere.org](http://www.gridisphere.org)
32. Commodity Grid toolkit, [www-unix.globus.org/cog](http://www-unix.globus.org/cog)
33. Sun WebStart Technology, <http://java.sun.com/products/javawebstart/>
34. Joint Data Standards Survey Project, [www.nesc.ac.uk/hub/projects/jdss](http://www.nesc.ac.uk/hub/projects/jdss)
35. Data Access and Integration Services working group, <https://forge.gridforum.org/projects/dais-wg>
36. Grid Data Service Specification, [http://forge.gridforum.org/docman2/ViewCategory.php?group\\_id=49&category\\_id=517](http://forge.gridforum.org/docman2/ViewCategory.php?group_id=49&category_id=517)
37. OGSA-DAI project, [www.ogsadai.org.uk](http://www.ogsadai.org.uk)
38. Scottish Bioinformatics Research Network (SBRN), [www.nesc.ac.uk/hub/projects/sbrn](http://www.nesc.ac.uk/hub/projects/sbrn)
39. Virtual Organizations for Trials and Epidemiological Studies (VOTES) project, [www.nesc.ac.uk/hub/projects/votes](http://www.nesc.ac.uk/hub/projects/votes)
40. R. Sinnott, Grid Based Clinical Trials Scenarios, presentation given at Global Grid Forum Life Science Research Group, Brussels, September 2004, <http://www.nesc.ac.uk/talks/ros/ClinicalTrialsOutlineScenariosv2.pdf>
41. UK e-Science National Grid Services, [www.ngs.ac.uk](http://www.ngs.ac.uk)