

Development of a Grid Infrastructure for Functional Genomics

Dr R. Sinnott¹, Dr M. Bayer¹, D. Houghton¹, Dr D. Berry², M. Ferrier²

¹National e-Science Centre, University of Glasgow, G12 8QQ,
{ros@dcs.gla.ac.uk, bayermm@dcs.gla.ac.uk, derekh@dcs.gla.ac.uk}

²National e-Science Centre, University of Edinburgh, EH8 9AA
{daveb@nesc.ac.uk, magnus@nesc.ac.uk}

Abstract. The BRIDGES project is incrementally developing and exploring database integration over six geographically distributed research sites with the framework of a Wellcome Trust biomedical research project (the Cardiovascular Functional Genomics project) to provide a sophisticated infrastructure for bioinformaticians. Grid technologies are being used to facilitate this integration. Key issues to be investigated in BRIDGES are data integration and data federation, security, user friendliness, access to large scale computational facilities and incorporation of existing bioinformatics software solutions, both for visualisation as well as analysis of genomic data sets. This paper outlines the initial experiences in applying Grid technologies and outlines the on-going designs put forward to address these issues.

1 Introduction

Hypertension affects a quarter of the adult population in western societies and is the major cause of cardiovascular mortalities. It is believed that hypertension is caused by a combination of factors including both genetic and environmental influences. The Wellcome Trust has funded a large (£4.34M) collaborative project (Cardiovascular Functional Genomics - 'CFG' [1]) to investigate the causes of hypertension. This five year project involves five UK and one Dutch site (depicted in Fig. 1). It is pursuing a strategy combining studies on rodent models of disease (mouse and rat) contemporaneously with studies of patients and population DNA collections. The project is a prime example of the large-scale computational problems associated with modern biology, with requirements to combine vast arrays of heterogeneous information about three species, human, mouse and rat.

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 often requiring navigation to many different data resources web sites and following multiple links to potentially relevant information. Similarly, in their pursuit of novel genes and understanding their associated function the 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 Biomedical Research Informatics Delivered by Grid Enables Services (BRIDGES) project [2] has recently been funded by the UK Department of Trade and Industry to directly address the needs of the CFG scientists and provide a thorough investigation of relevant technologies for this purpose. Specifically, BRIDGES will investigate the application of Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) [3] and IBM's Information Integrator product [4] to deal with federation of distributed biomedical data. Evaluation and benchmarking of these technologies is an important component of the BRIDGES work. In addition security is extremely important for the scientists. The scientific data itself may have different

characteristics. We consider three primary data kinds based upon their security characteristics:

- **Public data:** including public genome databases such as Ensembl [29] and GenBank [35], gene function databases such as LocusLink [36] and OMIM [37] and relevant publications databases such as PubMed [39] and MedLine [40];
- **CFG specific data:** that is to be shared between the CFG consortia only, or subsets of the CFG consortia;
- **Private data:** including potentially patient records and animal experiment data. This data has strict requirements on its access and usage which the Grid infrastructure must adhere to.

To meet these security requirements, Grid technology is being employed to establish a CFG *virtual organisation*. Virtual organisations provide a framework through which the rules associated with the participants and resources are agreed and enforced – especially those specific to security requirements. The distribution of CFG partners and the data security needs are depicted in Fig. 1. A central component to this virtual organisation is the notion of a Data Hub which is described in detail in section 2.

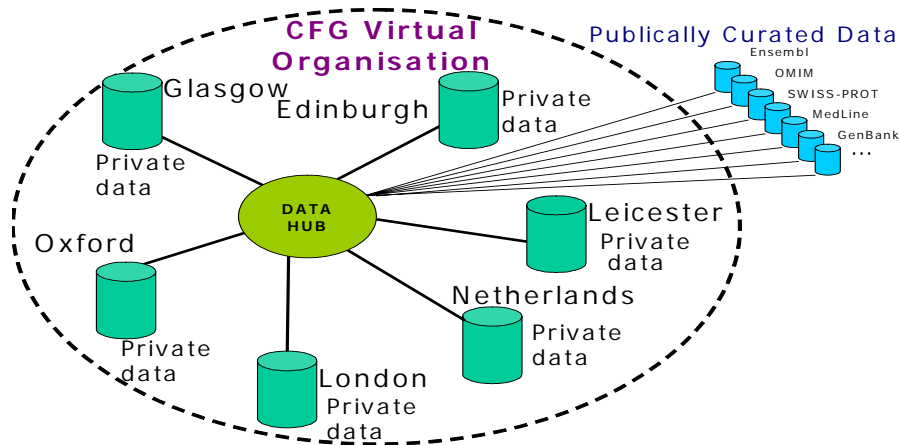


Fig. 1. Data Distribution and Security of CFG Partners

The Grid infrastructure to be deployed by BRIDGES should address all of the security concerns and interlinking of the different data sets in as transparent, and user friendly a manner as possible. The overall architecture and Grid related technologies themselves used to ensure this are discussed in section 2 (Overall Bridges Architecture), section 3 (Data Access and Integration), section 4 (Security Considerations) and section 5 (Portal Technologies). Issues and experiences in Grid enabling bioinformatics tools that the scientists use for visualisation of genomic data sets (and between data sets), as well as for analysis of these data sets on high throughput computational farms are discussed in section 6 (Grid Service Development). Finally in section 7, we draw conclusions and outline plans for the future, and provide acknowledgements.

2 Overall Bridges Architecture

The architecture of the Bridges infrastructure is depicted in Fig. 2. A key component of this architecture is the Data Hub which represents both a local, DB2 based, data repository, and data made available via externally linked data sets (through Information Integrator federated views as described in section 2.2). These data sets exist in different remote locations with differing security requirements. Some data resources are held publicly whilst others are for usage only by specific CFG project partners, or in some instances, only by the local scientists. It is especially important that local security issues are considered. Hence this architecture assumes the existence of multiple different institutional firewalls. The Data Hub itself makes use of two key technologies: OGSA-DAI and IBM's Information Integrator.

2.1 Grid Enabled Data Access and Integration Solutions

The Grid community is currently developing appropriate specifications for data access and integration in a Grid environment through the Data Access and Integration Service working group [20] at fora such as the Global Grid Forum [21]. Much of this work is driven by results from the OGSA-DAI project [3] and the recently funded follow up project, Data Access and Integration 2 (DAIT) [3]. OGSA_DAI/DAIT is a collaborative programme of work involving the Universities of Edinburgh, Manchester and Newcastle, the National e-Science Centre, with industrial participation by IBM and Oracle. Their principal objective is to produce open source database access and integration middleware which meets the needs of the UK e-Science community for developing Grid and Grid related applications. Its scope includes the definition and development of generic Grid data services providing access to and integration of data held in relational database management systems, as well as semi-structured data held in XML repositories.

OGSA-DAI have focused upon making these data resources available within an OGSA compliant architecture. The OGSA-DAI Grid services themselves provide the basic operations that can be used to perform sophisticated operations such as data federation and distributed queries within a Grid environment, hiding concerns such as database driver technology, data formatting techniques and delivery mechanisms from clients. This is achieved by the provision of a Grid-enabled middleware reference implementation of the components required to access and control data sources and resources.

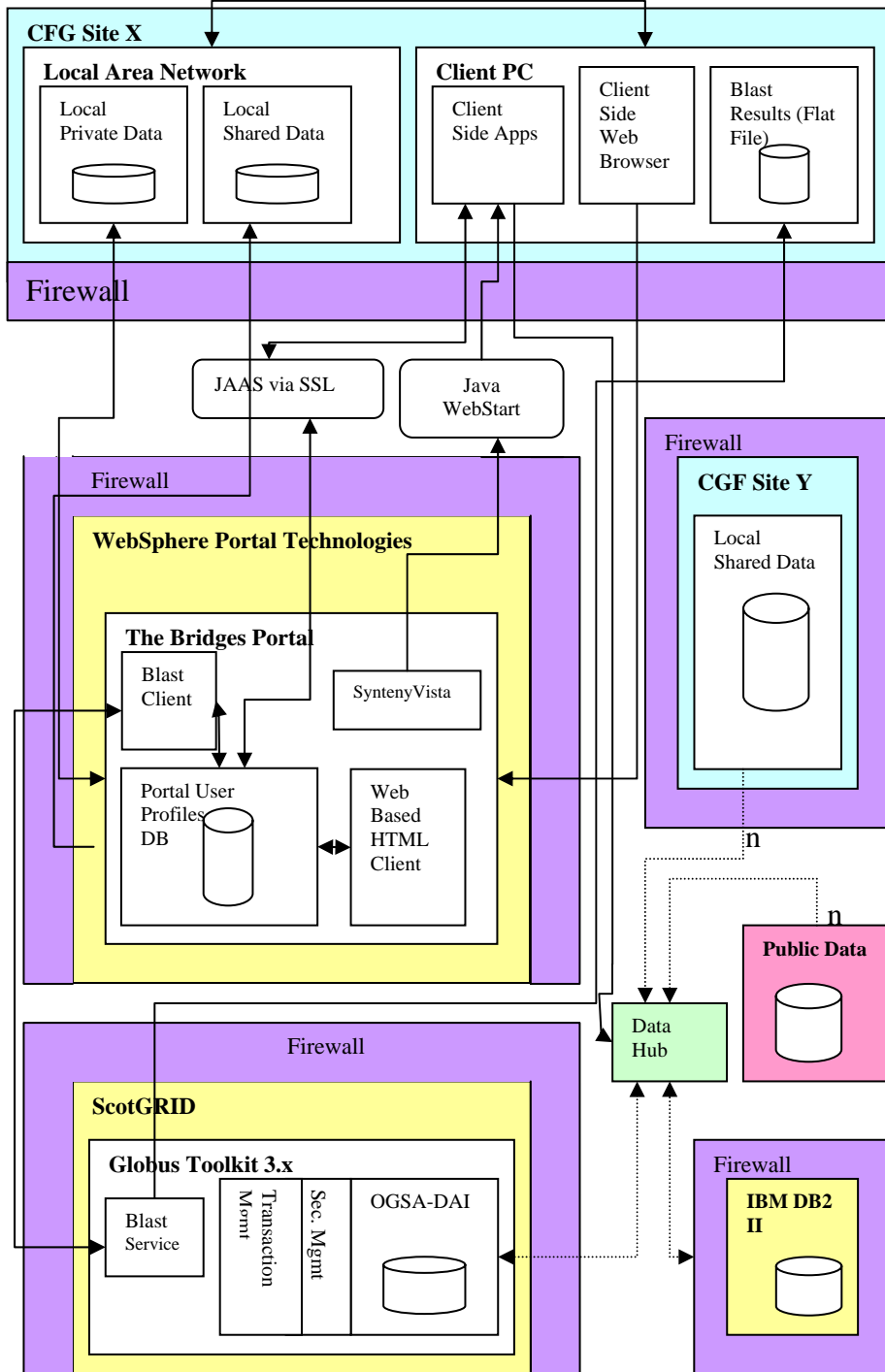


Fig. 2. Overall Bridges Architecture

OGSA-DAI itself can be considered as a number of co-operating Grid services. These Grid services provide a middleware layer for accessing the potentially remote systems that actually hold the data, i.e. the relational databases, XML databases or, as planned for the near future, flat file structures. Clients requiring data held within such databases

access the data via the OGSA-DAI Grid services. The precise functionality realised by OGSA-DAI is described in detail in the Grid Data Service Specification [22]. A typical scenario describing how this functionality might be applied to find, access and use (remote) data sets involves a persistent DAI Service Group Registry (made available via a Grid services hosting container such as Apache Tomcat [23]) offering persistent service factories (used for creating services to access and use specific data resources). Clients would contact the DAI Service Group Registry to find out what data sets are available, and once a required data source was found, create an instance of the Grid data service (via the appropriate factory) that would give access to this resource. The client can then issue queries (submit *Perform* operations via XML documents) to this Grid data service which extracts the queries and submits them to the appropriate databases, e.g. as SQL queries, before results are finally returned in XML documents. Extensions to this scenario to have multiple Grid data services supporting multiple, parallel queries executing through a given client query are possible.

2.2 Commercial Data Access and Integration Solutions

Information Integrator – which was previously known as DiscoveryLink - has been developed to meet the challenge of integrating and analyzing large quantities of diverse scientific data from a variety of life sciences domains. IBM Information Integrator offers single-query access to existing databases, applications and search engines. The Information Integrator solution includes the combined resources of Information Integrator middleware and IBM Life Sciences services. Using this software, IBM Life Sciences services can create new components that allow specialized databases—for proteomics, genomics, combinatorial chemistry, or high-throughput screening—to be accessed and integrated quickly and easily. This is depicted in Fig. 3.

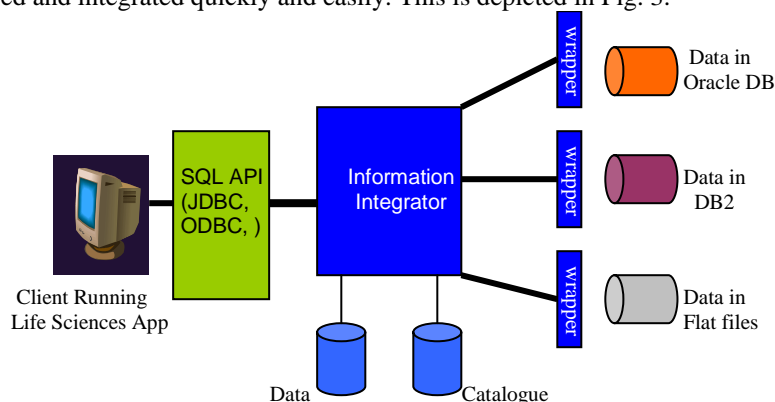


Fig. 3. IBM Information Integrator Data Access and Integration

At the far right of Fig. 3 are the data sources. Information Integrator talks to the sources using wrappers, which use the data source's own client-server mechanism to interact with the sources in their native dialect. Information Integrator has a local catalogue in which it stores information (metadata) about the data accessible (both local data, if any, and data at the backend data sources). Applications of Information Integrator manipulate data using any supported SQL API, for example, ODBC or JDBC are supported, as well as embedded SQL. Thus an Information Integrator application looks like any normal database application.

OGSA-DAI and Information Integrator have broadly similar aims: to connect distributed heterogeneous data resources. It is important that these two offerings are compared in a realistic life science environment. BRIDGES investigations will provide valuable information on the benefits of both of these solutions which will inform the wider Grid and life science communities. Currently the Information Integrator product for example requires programmatic access to different data repositories. This is not always the case - indeed it is normal for the life science public resources to *not* offer programmatic APIs where for example SQL based queries can be issued. Instead, these resources will generally offer only a web based front end for query submission, or make available their databases as compressed downloadable files. Similarly, open issues are being discovered with the current OGSA-DAI implementation, e.g. the ability to query resources offered as flat files, e.g. bioinformatics resources such as SWISS-PROT, and perform distributed joins across multiple remote databases.

The focus of OGSA_DAI and Information Integrator is primarily upon access and integration of data and not specifically upon security concerns. Security in the context of the Grid is an area that is currently receiving much attention since it is a crucial factor in the wider uptake of the Grid. There are numerous standards under development addressing aspects of security [10,24]. Within BRIDGES we are considering two security aspects: authentication and authorisation. These are of course considered in conjunction with existing best practice security, e.g. firewalls.

3 Security Considerations

Authentication is widely recognised as being only a starting point in establishing the security of a given Grid based system [3]. Authentication allows establishment of the identity of Grid users. The UK e-Science community has established a public key infrastructure (PKI) based upon X.509 certificates [2] for authentication which are issued through a central Certificate Authority (CA) at Rutherford Appleton Laboratories (RAL) in the UK [13]. These certificates are used to maintain a strong binding between a user's name and their public key when accessing remote Grid resources. It is recognised [3] however, that this approach to certification is unlikely to scale to a wider community, e.g. once Grid technologies are rolled out to the (life sciences) masses.

In the context of the Grid, X.509 based certificates are used to support the establishment and management of virtual organisations (VOs). Currently, however, existing solutions to establishing VOs do not adequately address the security needs of VO members. A fundamental requirement in establishing a VO is to ensure that efficient access control is achieved. Access control is usually done by comparing the authenticated name of an entity to a name in an Access Control List. In the UK e-Science Level 2 Grid [25] for example, which is based upon Globus toolkit version 2 [17], statically maintained "gridmap files" are used to limit who may or may not gain access to remote resources. This is achieved through so called Grid *gatekeepers*. This approach lacks scalability, manageability and does not meet the needs of dynamicity inherent to VOs. It is also limited in the level of granularity of the security model, e.g. in supporting fine grained authorisation to shared resources by potentially, dynamically changing collaborating VO members. We note that the CFG VO is unlikely to be especially dynamic however.

To improve this situation, authentication should be augmented with authorisation capabilities, which in this context can be considered as what Grid users are allowed to do on a given Grid end-system. This "what users are allowed to do" can also be interpreted as the privileges users have been allocated on those end-systems. The X.509

standard [2] has standardised the certificates of a privilege management infrastructure (PMI). A PMI can be considered as being related to authorisation in much the same way as a PKI is related to authentication. Consequently, there are many similar concepts in PKIs and PMIs. An outline of these concepts and their relationship are discussed in detail in [6].

A key concept from PMI are attribute certificates (ACs) which, in much the same manner as public key certificates in PKI, maintain a strong binding between a user's name and one or more privilege attributes. The entity that digitally signs a public key certificate is called a Certification Authority (CA) whilst the entity that signs an AC is called an Attribute Authority (AA). The root of trust of a PKI is sometimes called the root CA – which in terms of the UK e-Science community is given by the Grid Support centre at RAL [13]. The root of trust of the PMI is called the source of authority (SOA). CAs may have subordinate CAs whom they trust and to which they delegate the powers of authentication and certification. Similarly, SOAs may delegate their powers of authorisation to subordinate AAs. If a user needs to have their signing key revoked, a CA will issue a certificate revocation list. Similarly, if a user needs to have authorisation permissions revoked, an AA will issue an attribute certificate revocation list (ACRL). Typically, a given users' access rights are held as access control lists (ACLs) within each target resource. In an X.509 PMI, the access rights are held within the privilege attributes of ACs that are issued to users. A given privilege attribute within an AC will describe one or more of the user's access rights. A target resource will then read a user's AC to see if they are allowed to perform the action being requested.

The Privilege and Role Management Infrastructure Standards Validation (PERMIS) [15,16] is a role based authorisation infrastructure that realises a PMI – indeed the PERMIS project built and validated the world's first X.509 attribute certificate based authorisation infrastructure. Role Based Access Control (RBAC) models have been designed to make access control manageable and scalable [8]. To cater for RBAC solution for many applications, the PERMIS access control module has been developed [6,7]. It is a standards-based Java API that allows developers of resource gateways (gatekeepers) to enquire if a particular access to the resource should be allowed. PERMIS RBAC uses XML based policies defining rules, specifying which access control decisions are to be made for given VO resources [9]. These rules comprise:

- 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;
- the conditions under which a subject can be granted access.

The roles are assigned to subjects by issuing them with a standard X.509 Attribute Certificate [2]. The PERMIS team are currently working closely with the Globus team to design a standard Security Assertion Markup Language (SAML) [17] interface to any authorisation infrastructure. This will allow Grid applications to plug and play any authorisation infrastructure. As a result, the BRIDGES project has agreed to work with the PERMIS team and provide a rigorous investigation of security authorisation in a Grid biomedical context. Currently PERMIS has been extended with the SAML API and work is on-going to extend Globus Toolkit version 3 with a similar API [21]. The expected date for release of the extended GT3 is April 2004. Currently the BRIDGES

team is involved in defining suitable XML based policies suitable for the security authorisation requirements of the CFG project consortia, and identifying policy decision and enforcement points to be used when accessing the Grid services and associated CFG specific data sets. We are also involved in discussions in how remote CFG sites might make available their data sets in as secure a manner as possible, without compromising their local security requirements, e.g. through restricted and controlled opening of firewalls. Avenues incorporating existing solutions like *ssh/scp* are being explored.

In addition to authentication and authorisation security aspects, a key requirement of the CFG scientists is related to privacy. Privacy relates to the use of data, in the context of consent established by the data owner. There is little prior art in privacy Grid science, although there is useful UK background in privacy including hospital systems [18]. Web based standards such as P3P [19] may contribute to only a small fraction of the necessary security mechanisms. The area of privacy of biomedical data will be investigated as work on BRIDGES evolves.

4 Portal Technologies

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 should provide a personalisable 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. Portals in general offer several key advantages as a hosting and delivery mechanism for Grid services. They are:

- Highly flexible and extensible solutions
- Support content and information delivery suitable to users role
- Standardized look and feel across application suite
- Single entry point for services, data resources

Arguably the most mature portal technology on the market and the market leader is IBM WebSphere Portal Server, which has been used to develop the BRIDGES portal, although we note that several other solutions were also investigated including GridSphere [26] and the Commodity Grid toolkit [27]. 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 is shown in Fig. 4.

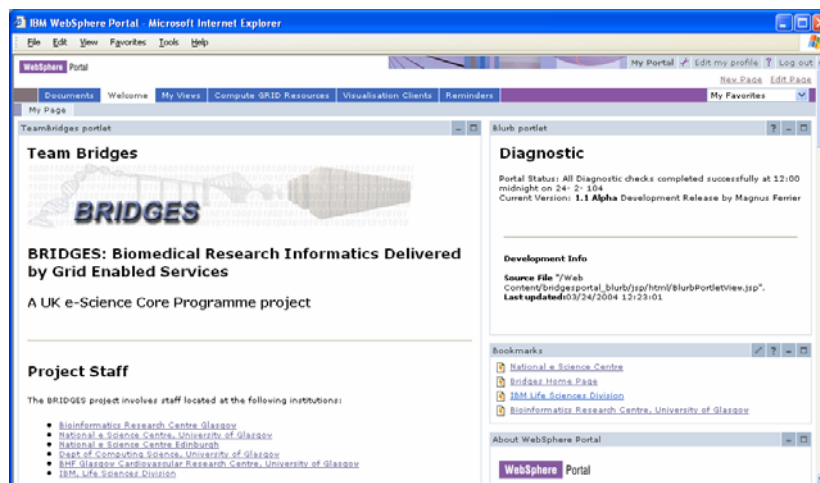


Fig. 4. BRIDGES Portal

This portal provides an integrated and personalisable environment through which the scientists have access to the various Grid services that they need. This will include Grid data services for the various information repositories of interest to the scientists; Grid services for visualisation of genomic data sets, and Grid services for analysis of genomic data sets. Currently the portal supports Grid enabled visualisation and analysis tools. Work is on-going in development of the Grid data services making use of the Data Hub and hence remote federated data sets.

Integral to the portal is security. The scientists have been issued (by the UK e-Science Certification Authority) with X.509 certificates which are embedded in their browsers. Depending upon the role of the portal user (e.g. scientist, systems administrator, principal investigator etc) the X.509 certificate is used to limit what services the portal user sees and subsequently is allowed to invoke. Of course only certificates for scientists involved with the CFG virtual organisation are recognised, hence non-authorized access to the portal and the services/data sets available there is not possible. A generic mapping between certificates used for authentication and certificates used for fine grained authorisation is currently under development.

5 Grid Service Development

At the time of writing several trial-Grid services have been engineered and made available through the BRIDGES portal. We describe each of these in turn.

5.1 Grid Enabled Synteny Visualisation Services

Synteny is the condition of two or more genes being located on the same chromosome. Of particular interest to the CFG scientists is conserved synteny which may be defined as the condition where a syntenic group of genes from one species have orthologues (similar genes, where the similarity itself can be ascertained through a combination of approaches such as protein sequence similarity, structure, function etc) in another species.

The analysis of conserved synteny between the different organisms (e.g. mouse, rat and human), in combination with quantitative trait loci (QTL) data [28] and microarray experiments, 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. 5. Originally SyntenyVista was developed under the assumption that the relevant chromosome data sets were locally available, e.g. as files on the same machine where SyntenyVista itself was running. This has numerous limitations. Firstly, the user must manually download the relevant data sets, and possibly the complete databases from public resources offering syntenic information, e.g. Ensembl [29]. Secondly, each time the user wants to visualise syntenic data sets, they need to manually check, e.g. by visiting the public data repository, whether a newer more up to date version of their syntenic data set is available. Grid technology offers a mechanism to overcome these restrictions, through automatically accessing and pulling down relevant remote data sets as and when needed.

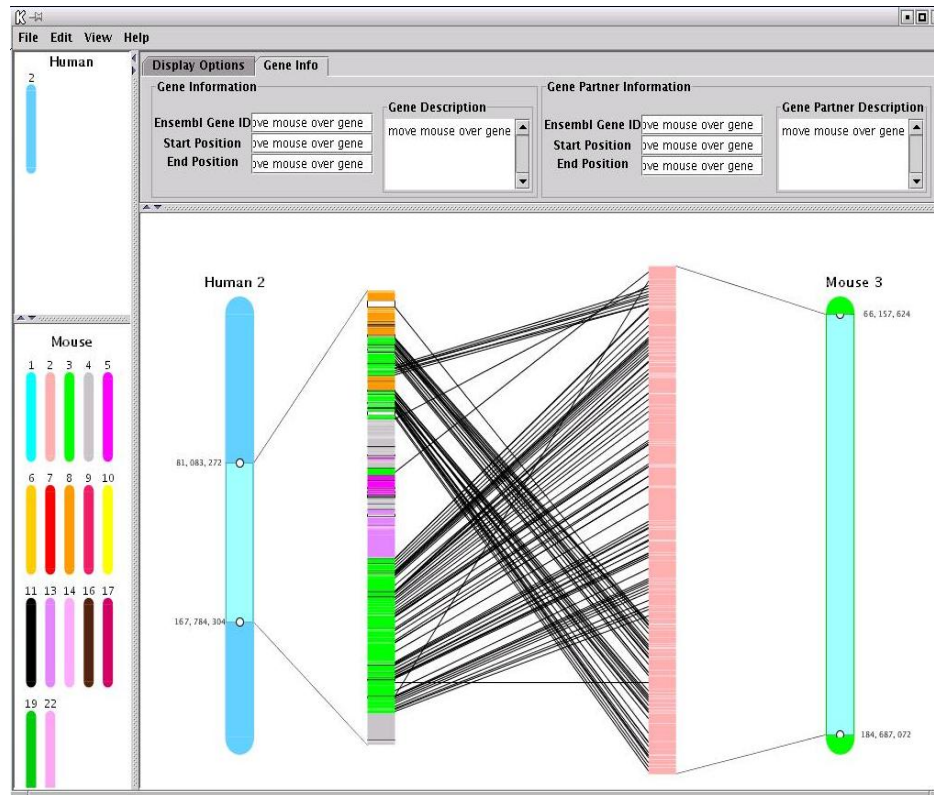


Fig. 5. Grid Enabled Syntenic Visualisation Tool

To achieve this SyntenyVista has been augmented with OGSA-DAI capabilities. Specifically, SyntenyVista is now able to access remote syntenic data sources and pull

down (cache) them as needed. The portal delivery mechanism for SytenyVista itself is via Sun's WebStart [30]. The Java WebStart technology simplifies deployment of Java applications and enables launch of full-featured applications through single clicks from web browsers. Upon launching SytenyVista from the portal, automatic checks are done to ensure that the latest version of the software is available, and if not, automatically downloaded and installed.

The Grid enabled version of SytenyVista automatically checks on syntenic data sets that might have been cached already. When these are available, they are loaded (onto the pallet on the left hand side of Fig. 5). Users are then able to drag these onto the main window to see where syteny might be present between different chromosomes. When these data sets are not cached locally, remote resources accessible via OGSA-DAI are accessed and the data pulled down. Currently the Grid data services for finding syntenic data sets are under development, and as such the current implementation uses OGSA-DAI directly to connect to a single syntenic data source (Ensembl).

5.2 Grid Enabling BLAST Services

Bioinformaticians typically need to be able to find similarities between different genomic or protein sequences. The Basic Local Alignment Search Tool (BLAST) [31] 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. Typically full scale BLAST jobs across whole genomes is a highly compute intensive activity. As a result, large scale compute farms are often required.

The ScotGrid computational resource at the University of Glasgow offers precisely such a high throughput compute facility [32]. It 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 is used by varieties of scientists across the university and internationally including particle physicists, electronic engineers, computer scientists and bioinformaticians. ScotGrid itself offers a Beowolf Linux cluster with the equivalent of 330 1GHz processors and 15TB disk space comprised of IBM xSeries, Blade server, FAStT500 and Dell and Cisco technologies. It uses the Maui scheduling software [33] which itself is based on OpenPBS [34].

To provide Grid enabled BLAST services accessing and using ScotGrid, it was required that the BLAST software was ported into the Grid environment, i.e. made available as a GT3 based Grid service. The original version of BLAST we used was implemented in C hence this required writing appropriate Java wrappers and exploiting specific GT3 APIs. The current prototype for GT3 BLAST job submission is based on the GT3 core only, and involves a simple wrapper to OpenPBS commands, due to difficulties with the full GT3 installation.

BLAST takes in a search sequence and assumes a target sequence database. This requires that the source and target data sets were staged onto the ScotGrid infrastructure, and the results pulled off once the job itself had completed. It should be noted that in the current implementation, we simply keep a copy of the relevant target database on ScotGrid, however it is expected that this solution will be modified once we deal with more security dependent data sets. Job monitoring services based on usage of the OpenPBS API have also been implemented. These Grid services and monitoring services are all available through the BRIDGES portal and provide a valuable resource to the CFG scientists.

6 Conclusions and Acknowledgements

The BRIDGES project began in October 2003 and is engaged in the evaluation of a wide variety of Grid technologies applied to the life science domain. Focus thus far has been oriented towards, mainstream implementations of Grid technology such as Globus toolkit, and its recent web service based version (GT3). GT3 usage has not been without issues however, and often workaround solutions have been necessary. For example, full installation of GT3 with GRAM for job submission/management capabilities was found to be especially problematic due to undocumented operating systems dependencies. Further, since we often required software to run on Windows OS flavours, e.g. for running WebSphere portal software, compromises had to be made between the architecture and design, and the final systems that have been implemented.

The current implementation has provided a proof of concept prototype. The next phase of the work will look in more detail at the key requirements of the CFG scientists. For example, the scientists are especially interested in microarray analysis. Tools and workflows that allow the scientists to take up/down regulated gene names from microarray experiments and garner further information are of special interest. This in turn requires that our Data Hub can connect to relevant data sites to pull down specific information on the genes themselves. The current focus of our Data Hub is on genome databases such as Ensembl [29], GenBank [35]; gene function databases such as LocusLink [36], OMIM [37], transcriptome databases such as Unigene [38], and relevant publications databases such as PubMed [39] and MedLine [40]. Currently, few of these resources provide the necessary programmatic access needed. Instead, they often only offer compressed files that can be downloaded. As a result, a local warehouse is being established and populated. This will make use of both IBM's Information Integrator technology and other Grid services, e.g. for replica location management [41]. This will support automated, dynamic updating of the local repository with remote, public data sets, as and when these change.

There exist numerous other visualisation and analysis tools that the CFG scientists would like to be supported and offered through the BRIDGES portal. These include sequence visualisation tools and multiple alignment tools. Work is currently on-going in Grid enabling these. Similarly, the scientists are keen to use numerous analysis and sequence comparison tools such as clustalw [42] and Smith-Waterman [43]. Their Grid enablement is also under way. We note that not all scientific compute demands are satisfied by farms such as ScotGrid, and often a large SMP machines are required. Where this is the case, large SMP resources offered through the UK e-Science Grid such as Blue Dwarf at Edinburgh University will be made.

Finally a role based authorisation infrastructure meeting the needs of the CFG scientists and their associated is under development. This will give valuable insight into the much needed security concerns raised by rolling out the Grid to the wider (life science) community.

6.1 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 CFG team including Prof. David Gilbert, Prof Malcolm Atkinson, Dr Ela Hunt and Dr Neil Hanlon. Drs Hanlon and Hunt are also acknowledged for their contribution to the original SyntenyVista software. Acknowledgements are also given to the IBM collaborators on BRIDGES, notably Drs David White, Andy Knox and Jean-Christophe Mestres. The CFG project is supported by a grant from the Wellcome Trust foundation.

References

1. Cardiovascular Functional Genomics project, <http://www.brc.dcs.gla.ac.uk/projects/cfg/>
2. BioMedical Research Informatics Delivered by Grid Enabled Services (BRIDGES), www.brc.dcs.gla.ac.uk/projects/bridges
3. Open Grid Service Architecture – Data Access and Integration project (OGSA-DAI), www.ogsadai.org.uk
4. IBM Information Integrator, http://www3.ibm.com/solutions/lifesciences/solutions/Information_Integrator.html
5. E-Science Security Roadmap: Technical Recommendations v0.5, UK e-Science Security Task Form, draft executive summary v0.51
6. ITU-T Rec. X.509 (2000) | ISO/IEC 9594-8 The Directory: Authentication Framework
7. C Adams and S Lloyd (1999), Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations, Macmillan Technical Publishing.
8. Adams, C., Lloyd, S. (1999). "Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations", Macmillan Technical Publishing, 1999
9. Austin, T. "PKI, A Wiley Tech Brief", John Wiley and Son, ISBN: 0-471-35380-9, 2000
10. Grid Security, <https://forge.gridforum.org/projects/sec>
11. 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.
12. M Thompson, et al., Certificate-Based Access Control for Widely Distributed Resources, in Proc 8th Usenix Security Symposium. 1999: Washington, D.C.
13. VOMS Architecture, European Datagrid Authorization Working group, 5.9.2002.
14. Steven Newhouse, Virtual Organisation Management, The London E-Science centre, <http://www.lesc.ic.ac.uk/projects/oscar-g.html>
15. 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.
16. Privilege and Role Management Infrastructure Standards Validation project www.permis.org
17. P Hallem-Baker, E Maler, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML), OASIS, SAML 1.0 Specification. 31 May 2002. <http://www.oasis-open.org/committees/security/#documents>
18. I. Denley and S.W. Smith, Privacy in clinical information systems in secondary care. British Medical Journal, 1999. 318: p. 1328-1331.
19. Platform for Privacy Preferences (P3P) Project, W3C, <http://www.w3.org/P3P/>
20. Data Access and Integration Services working group <https://forge.gridforum.org/projects/dais-wg>
21. Global Grid Forum, www.ggf.org
22. Grid Data Service Specification, https://forge.gridforum.org/docman2/ViewCategory.php?group_id=49&category_id=517
23. Apache web site, www.apache.org
24. Web Security Standards, http://www.oasis-open.org/committees/documents.php?wg_abbrev=wss
25. UK e-Science Engineering Task Force, www.grid-support.ac.uk/etf
26. GridSphere Portal, www.gridsphere.org
27. Commodity Grid toolkit, www-unix.globus.org/cog
28. 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
29. EMBL-EBI European Bioinformatics Institute, <http://www.ebi.ac.uk/ensembl/>
30. Sun WebStart Technology, <http://java.sun.com/products/javawebstart/>
31. Basic Local Alignment Search Tool (BLAST), <http://www.ncbi.nlm.nih.gov/Tools/>
32. ScotGrid, www.scotgrid.ac.uk
33. Maui Job Scheduler, <http://www.supercluster.org/maui/>
34. Open Portable Batch System, www.openpbs.org

35. NCBI GenBank, <http://www.ncbi.nlm.nih.gov/Genbank/>
36. NCBI LocusLink, <http://www.ncbi.nlm.nih.gov/LocusLink/>
37. NCBI Online Mendelian Inheritance in Man, <http://www.ncbi.nlm.nih.gov/OMIM/>
38. NCBI Unigene, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>
39. PubMed Central Home, <http://www.pubmedcentral.nih.gov/>
40. US National Library of Medicine, <http://www.nlm.nih.gov/>
41. Replica Location Service (RLS), www.globus.org/rls
42. EMBL-EBI European Bioinformatics Institute clustalw, <http://www.ebi.ac.uk/clustalw/>
43. EMBL-EBI European Bioinformatics Institute MPSrch, <http://www.ebi.ac.uk/MPSrch/>