Grid Services Supporting the Usage of Secure Federated, Distributed Biomedical Data

Richard Sinnott[1], Malcolm Atkinson[2], Micha Bayer[1], Dave Berry[2], Anna Dominiczak[3], Magnus Ferrier[2], David Gilbert[4], Neil Hanlon[3,4], Derek Houghton[1], Ela Hunt[5], David White[6]

[1]National e-Science Centre, e-Science Hub, Kelvin Building, University of Glasgow, Glasgow G12 8QQ
[2]National e-Science Centre, e-Science Institute, 15 South College Street, Edinburgh EH8 9AA
[3]BHF Blood Pressure Group Division of Cardiovascular & Medical Sciences, Western Infirmary, Church Street, Glasgow, G11 6NT
[4]Bioinformatics Research Centre, University of Glasgow, Glasgow G12 8QQ
[5]Department of Computing Science, University of Glasgow, Glasgow G12 8QQ
[6]IBM Life Sciences UK, 1 New Square, Bedfont Lakes, Feltham, Middlesex, TW14 8HB

Abstract

The BRIDGES project is a UK e-Science project that provides grid based support for biomedical research into the genetics of hypertension – the Cardiovascular Functional Genomics Project (CFG). Its main goal is to provide an effective environment for CFG, and biomedical research in general, including access to integrated data, analysis and visualization, with appropriate authorisation and privacy, as well as grid based computational tools and resources. It also aims to provide an improved understanding of the requirements of academic biomedical research virtual organizations and to evaluate the utility of existing data federation tools.

1. Introduction

The Wellcome Trust has funded a large collaborative project (Cardiovascular Functional Genomics - 'CFG' [1]) over five years that involves five UK and one Dutch site to investigate the genetic factors in hypertension. This project exemplifies the large-scale computational problems of modern biology, with requirements to combine information about three species: human, mouse and rat.

The BRIDGES (Biomedical Research Informatics Delivered by Grid Enabled Services) project directly addresses the needs of the CFG scientists using Grid based technology. Specifically, BRIDGES is investigating the application of the Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) system [2] and IBM's Information Integrator product [3] to deal with federation of distributed biomedical data. In addition, it is addressing the security requirements, which is important for the scientists. The scientific data itself mainly consists of public domain data from a variety of sources but also includes various types of data generated by the research consortia, some of which is for group use only, and some of which is shared within the consortium.

2. BRIDGES Architecture

The system architecture employed by BRIDGES is shown in Figure 1. The key component in this 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). 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..

3. Data Integration

The OGSA-DAI software 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.

IBM's Information Integrator offers single-query access to existing databases, applications and search engines as depicted in Figure 2. 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.

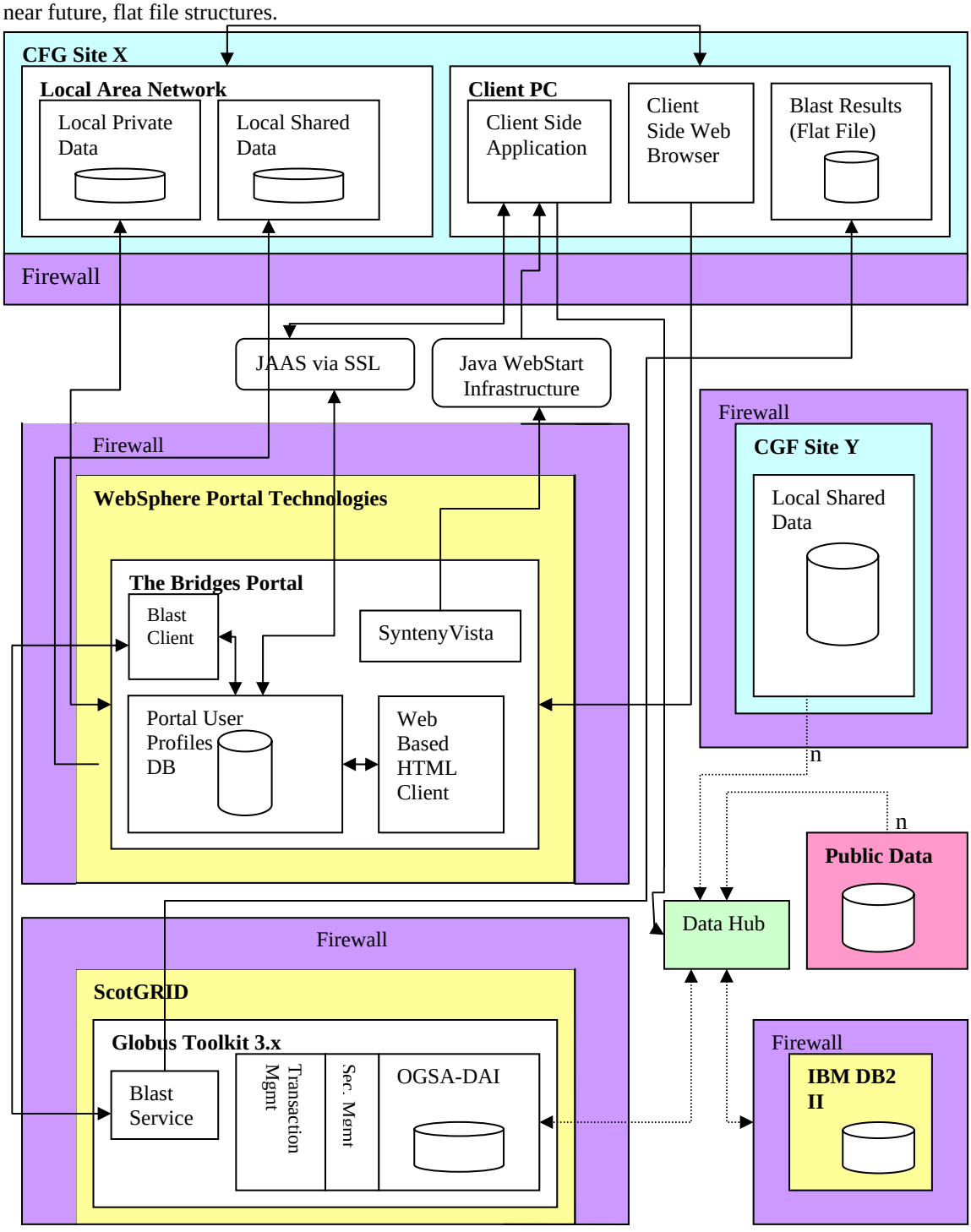

Figure 1: Overall Bridges Architecture

Currently both OGSA-DAI and Information Integrator require programmatic access to data sources. This, however, is not always available - indeed it is normally the case that the life science public resources do not offer direct 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. SWISS-PROT, and perform distributed joins across multiple remote databases.

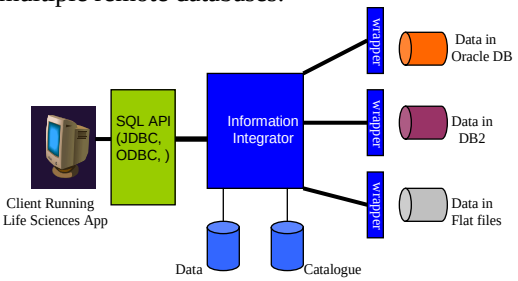

Figure 2: IBM Information Integrator Data Access and Integration

4. Security

Within BRIDGES we mainly consider two security aspects: authentication and authorisation.

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 for authentication which are issued through a central Certificate Authority (CA) at Rutherford Appleton Laboratories (RAL) [5].

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. 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 has standardised the certificates of a privilege management infrastructure (PMI).

The Privilege and Role Management Infrastructure Standards Validation (PERMIS) [6,7] 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. The roles are assigned to subjects by issuing them with a standard X.509 Attribute Certificate. The PERMIS team are currently working closely with the Globus team to design a standard Security Assertion Markup Language (SAML) [8] 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 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 when accessing the Grid services and associated CFG specific data sets.

5. Portal Technology

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, a web-based project portal was developed. This portal provides 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. One of the more mature portal technologies on the market, and the market leader, is IBM WebSphere Portal Server which has been used to develop the BRIDGES portal (Figure 3).

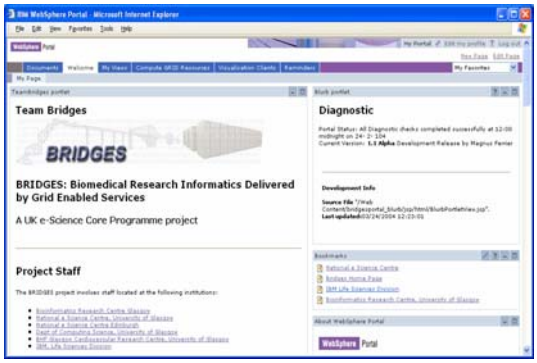

Figure 3: BRIDGES Portal

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.

6. Grid Services

6.1 SyntenyVista Visualisation Tool

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 in another species. SyntenyVista (Figure 4) was developed for this purpose. 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, and therefore SyntenyVista has been augmented with OGSA-DAI capabilities.

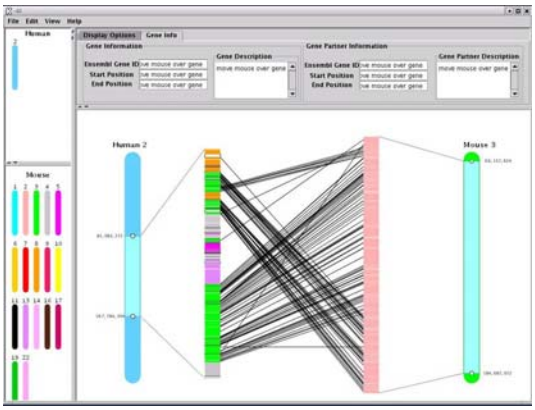

Figure 4: Grid Enabled SyntenyVista tool

The portal delivery mechanism for the Grid-enabled version of SyntenyVista is via Sun's WebStart [9]. This version of SyntenyVista 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 Figure 4). When these data sets are not cached locally, remote resources accessible via OGSA-DAI are accessed and the data pulled down.

6.2 Grid-Based BLAST Service

Biologists often need to be able to detect similarities between different genomic sequences. The Basic Local Alignment Search Tool (BLAST) [10] has been developed to perform this function. Numerous versions of BLAST exist, targeted towards different sequence data sets and offering various levels of performance and accuracy metrics. Typically, full scale BLAST jobs across whole genomes are 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 such a high throughput compute facility [11]. 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. To provide Grid-enabled BLAST services accessing ScotGrid, the BLAST software was ported into the Grid environment, i.e. made available as a GT3 based Grid service. The current prototype for GT3 BLAST job submission is based on the GT3 core only, and involves a simple wrapper to OpenPBS commands.

6.3 MagnaVista

MagnaVista is a tool for the integrated textual display of data from the federated data sources (see above). It provides a convenient way of viewing all the available data relating to a single gene, in a variety of formats including sorted tables and trees.

7. Acknowledgements

This work was supported by a grant from the Department of Trade and Industry. Acknowledgements are also given to the IBM collaborators on BRIDGES, notably Dr's Andy Knox, Colin Henderson and Jean-Christophe Mestres. The CFG project is supported by a grant from the Wellcome Trust foundation.

8. References

[1] Cardiovascular Functional Genomics project: http://www.brc.dcs.gla.ac.uk/projects/cfg/
[2] Open Grid Service Architecture – Data Access and Integration project (OGSA-DAI): www.ogsadai.org.uk
[3] IBM Information Integrator: http://www3.ibm.com/solutions/lifesciences/solutions/Information Integrator.html
[4] BioMedical Research Informatics Delivered by Grid Enabled Services (BRIDGES): www.brc.dcs.gla.ac.uk/projects/bridges
[5] VOMS Architecture, European Datagrid Authorization Working group, 5 September 2002.
[6] 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.
[7] Privilege and Role Management Infrastructure Standards Validation project: www.permis.org
[8] P Hallem-Baker and 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
[9] Sun WebStart Technology: http://java.sun.com/products/javawebstart/
[10] Basic Local Alignment Search Tool (BLAST): http://www.ncbi.nlm.nih.gov/Tools/
[11] ScotGrid: www.scotgrid.ac.uk