University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
SINNOTT, RICHARD;BAYLISS, CHRISTOPHER

Title:
Towards data grids for microarray expression profiles

Date:
2006

Citation:
Sinnott, R., & Bayliss, C. (2006). Towards data grids for microarray expression profiles. In Proceedings of the Third International Life Science Grid Workshop (LSGRID) , Yokohama, Japan.

Publication Status:
Published

Persistent Link:
https://hdl.handle.net/11343/28844

Towards Data Grids for Microarray Expression Profiles

Richard Sinnott and Christopher Bayliss

University of Glasgow, National e-Science Centre, Glasgow, G12 8QQ, UK
r.sinnott@nesc.gla.ac.uk c.bayliss@nesc.gla.ac.uk

Abstract. The UK DTI funded Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) project developed a Grid infrastructure through which research into the genetic causes of hypertension could be supported by scientists within the large Wellcome Trust funded Cardiovascular Functional Genomics project. The BRIDGES project had a focus on developing a compute Grid and a data Grid infrastructure with security at its heart. Building on the work within BRIDGES, the BBSRC funded Grid enabled Microarray Expression Profile Search (GEMEPS) project plans to provide an enhanced data Grid infrastructure to support richer queries needed for the discovery and analysis of microarray data sets, also based upon a fine-grained security infrastructure. This paper outlines the experiences gained within BRIDGES and outlines the status of the GEMEPS project, the open challenges that remain and plans for the future.

1 Introduction

Post-genomic in-silico life science research is now a reality with a whole vista of potential benefits within grasp from personalised e-Health, drug discovery, to understanding of complete organisms and their genetic make-up. However, supporting such research poses new challenges that must be addressed before the possible benefits of the post-genome era can be realised.

Arguably the greatest challenge in supporting this research is dealing with the exponential data growth with ever increasing numbers of genomes being sequenced, proteomes being populated, pathways being discovered, and increased numbers of automated systems for expediting the production of these data sets and numerous others. The data challenge is made more difficult due to combinations of factors including: the breadth of data across numerous different research domains, numerous species and organisms; the possibility for erroneous or contradictory data, and knowledge and assumptions based upon these potentially erroneous data sets; the largely independent myriad collections of data owners and data providers along with their own idiosynchracies in how they wish to make available their data sets available, and in turn the standards and schemas associated with their respective data sets; the likelihood of system change and evolution based on new insights and scientific discoveries. In all of this, the willingness of the scientific community to adopt appropriate standards to facilitate

data sharing and reuse must be recognised. Technologies that are too *difficult* or standards that require too much effort to support will not gain widespread acceptance and take up. It is clearly the case that data sharing considerations also need to be cognisant of the often cultural, social, ethical, political research processes and concerns of the scientific community they are to support.

Grid technology and the ideas behind the Grid in overcoming issues of distribution and infrastructure heterogeneity offers some possible solutions to address some of the challenges associated with creating, managing and using life science data sets, however technology alone is insufficient and must be guided by the wider scientific community needs and experiences. This includes community standardisation efforts in how to annotate data so that it can subsequently be found, accessed, integrated and analysed. The scientific community needs to be made aware of what it means to provide controlled access to their research data and the potential ramifications thereof. Biologists tend not to be computer scientists and are unfamiliar with advanced Grid data access or security solutions. As such any solutions that are put forward in this domain have to be intuitive and allay their potential fears on compromises of their research data, or potential exploitation by competitors or third parties. New developments such as gene identification, gene function and development of new targeted drugs offers enormous opportunities for researchers both financially and from research recognition. As such, they need to be completely satisfied that any new technological solutions will fit into the way in which they wish to work, and importantly protect their research results and data from compromise.

Such research is however rarely undertaken by a single site. Multi-site collaboration drawing on expertise in a range of disciplines is a common model for collaboration. In such circumstances the need to securely collaborate and share results, data, services and processes more generally requires technology that facilitates the research process. In Grid parlance, the establishment, control and enforcement of virtual organisations (VOs) offers one suitable model by which researchers can effectively co-ordinate their efforts.

Within this context, the UK Department of Trade and Industry funded Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) project [1] developed a compute and data Grid infrastructure through which many of the challenges associated with suporting life science research were explored. Specifically BRIDGES developed an infrastructure for the Wellcome Trust funded Cardiovascular Functional Genomics (CFG) project [2] which was exploring the genetic factors involved in hypertension which effects 25% of western society and is a major cause of cardiovascular mortality. The experiences in BRIDGES have been documented in [3–7]. This paper summarises some of the experiences with BRIDGES especially with regard to data Grids and security, and outlines plans and the status of a complementary follow on project funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) in the UK: the Grid Enabled Microarray Expression Profile Search (GEMEPS) project [8].

2 Data Grids within BRIDGES

The BRIDGES project began in 2003 and successfully completed at the end of December 2005. It involved the National e-Science Centre at the University of Glasgow and Edinburgh, with industrial participation from IBM. BRIDGES was specifically targeted to develop a Grid infrastructure meeting the needs and requirements of the CFG project.

The overall CFG partner distribution and associated data distribution identified at the project inception is depicted in Figure 1. At the heart of the CFG Virtual Organisation supported by BRIDGES was a data hub through which simplified user-oriented access to a range of genomic data sets was made available. The data hub was thus a data Grid. This data hub was realised through two different technologies: the commercial data integration technology solution, IBM DiscoveryLink - later remarketed as IBM Information Integrator [9]; the second based on the Grid communities open source Open Grid Service Architecture Data Access and Integration (OGSA-DAI) software [10]. An evaluation and comparison of these technologies including their performance and overall usability in the functional genomics domain was made and is documented in [11].

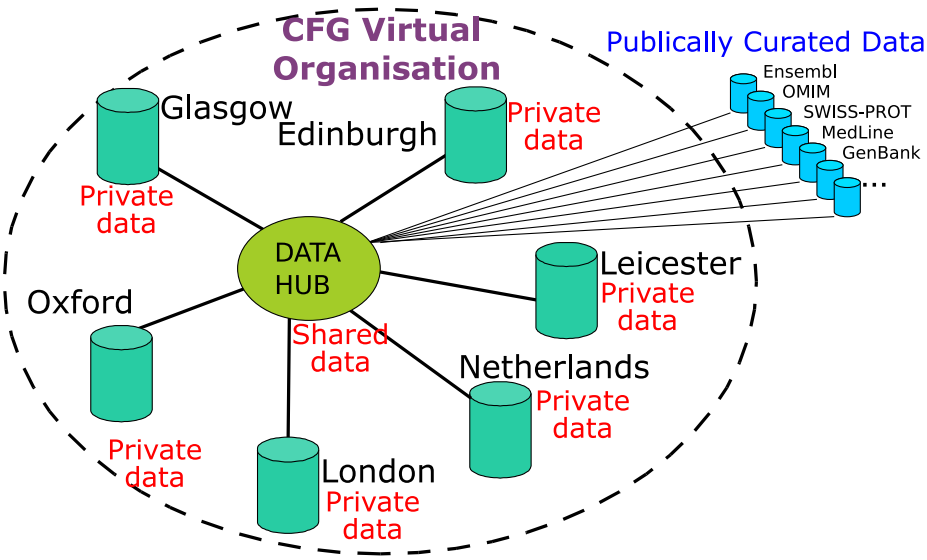

Fig. 1. Data Distribution and Security of CFG Partners

The scientists were primarily interested in translational based research focused specifically on rodent and mammalian organisms. Through breeding rodents to be hypertensive and running microarray experiments based upon these

hypertensive rodent sets, analysis between healthy and hypertensive rodents and subsequently to other mammalian species was undertaken. The fundamental questions that were searched for by the scientists were in understanding which genes caused or involved in the cause of hypertension, and once identified how drugs could be targeted to those specific genes.

In supporting this translational research, the scientists required a single unified view of a range of public genomic databases including: Ensembl (rat, mouse, human databases) [12]; Mouse Genome Informatics (MGI) [13]; Online Mammalian Inheritance in Man (OMIM) [14], Human Genome Organisation (HUGO) [15], Rat Genome Database (RGD) [16] and the Gene Ontology (GO) data base [17]. The typical scenario supported within BRIDGES and desired by the CFG scientists was based upon the scientists running a microarrary experiment at their local institutional microarray facility, e.g. the Sir Henry Wellcome Functional Genomic Facility (SHWFGF) at the University of Glasgow. The predominant microarray chips that were used were based upon versions of the Affymetrix chip sets [18]. The output of these experiments included amongst other things, collections of data identifying genes and their levels of expression. Based on this, the scientists would then use the BRIDGES data Grid infrastructure to return a variety of information on specific genes of interest. Specifically to support this BRIDGES developed various client side tools (MagnaVista and latterly GeneVista) through which queries could be formulated and result sets returned from the data Grid. MagnaVista was a Java application built to access the IBM DiscoveryLink version of the data Grid, and GeneVista (Figure 2) which was targeted towards the OGSA-DAI version of the data Grid. Both of these applications were based upon a similar use case model, namely that they would take a gene name, e.g. from a microarray experiment, and return all information associated with that gene.

The information returned was dependent upon the remote schemas associated with the public genome databases and would typically include references to published journals papers in MedLine or PubMed, protein sequences from ensembl, accession numbers in other databases, and a variety of other information. We note that client side tools were designed to be adaptable to the interests and needs of the scientists. We also note that the scientists provided continued feedback on the usability and HCI aspects of the tools. Hence GeneVista was designed to be "google-like" in its interface as demonstrated on the top of Figure 2 with results returned on the bottom of Figure 2.

The BRIDGES data Grid was, at the project inception, planned to support an array of different kinds of data across the CFG partner sites with different security classifications as depicted in Figure 1. These included: public data from the public genome resources mentioned above; processed public data that has additional annotation or indexing to support the analyses needed by CFG; sensitive data about individuals in the cohorts of patients or data derived from animal experiments; special experimental data such as quantitative trait loci (QTL) or microarray data; personal research data specific to a researcher as a result of experiments or analyses that that researcher is performing; team research data

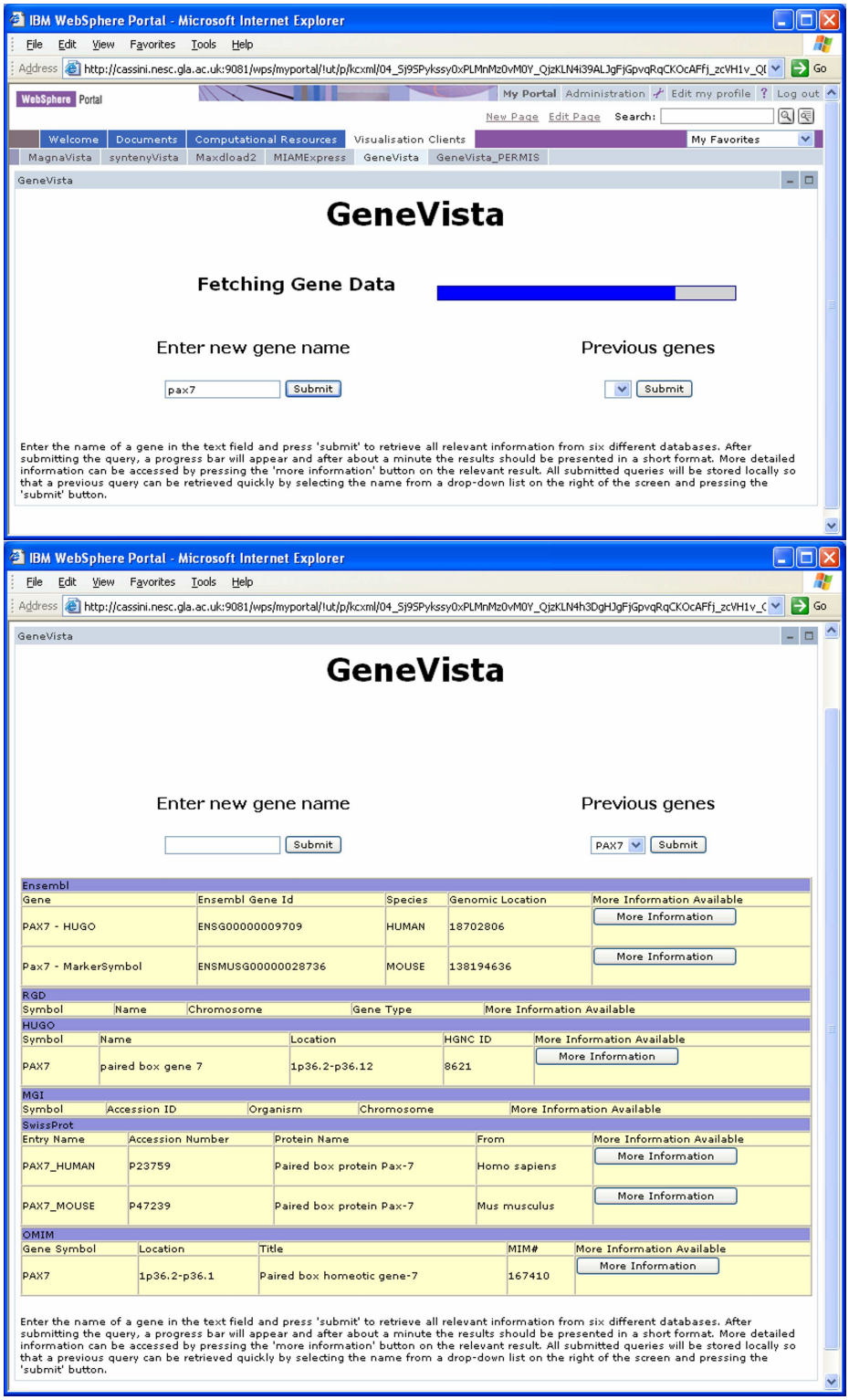

Fig. 2. GeneVista interface and returned data displayed

shared by the team members at a site or within a group at a site; consortium research data produced by one site or a combination of sites that has been made available for the whole consortium; personalisation data and metadate collected and used to improve the tools pertinent to individual users.

Initially, the BRIDGES data Grid development work was focused on providing a Grid infrastructure that allowed single unified access to public genomic resources. In this model the client tool (MagnaVista/GeneVista) was used to issue a query to the DB2 database at the heart of the data hub. Through DiscoveryLink wrappers or OGSA-DAI interfaces to remote resources accessible from the DB2 repository, these queries were subsequently federated to the appropriate remote resources and the result sets joined together before being displayed by the tool to the client. Whilst simplistic from a basic data Grid design perspective, it was soon recognised that this was a fraught process for numerous reasons. Firstly, the programmatic access needed for the Grid data integration technologies was only available for the ensembl and MGI databases. For the other genomic data resources alternative solutions were required including downloading the data (often with no schema being provided), parsing the flat files and developing solutions to trigger the population of these files into the DB2 database. Secondly, even with this limited selection of remote data resources, changes in a remote database schema or the renaming of that database resulted in queries failing. Even the renaming of a table column from "Gene Name" to "Gene Identifier" or to "Gene Reference" resulted in failed queries since the fail of a single query from the federated query set meant that no join of the result sets was possible. Thirdly, at the time of development neither DiscoveryLink or OGSA-DAI supported flat file data access and integration in a manner which directly supported CFG. More accurately, DiscoveryLink supported flat files however a requirement was that these files had the same set of permissions as the local DB2 installation. Since this was never true for files on remote sites other solutions were required. A fourth problem that arose was the lack of unique terms by which these tables could subsequently be joined. Naming of genes and associated data sets by the different public genome repositories is not closely co-ordinated with different naming conventions used and changes and updates to databases and their schemata done independently. These issues are described in much more detail in [11] along with the challenges and solutions that were adopted to handle changes in the remote database schemas.

Given the difficulties with supporting live access to remote public repositories, the BRIDGES project also developed solutions which allowed the scientists to support access to shared secure data sets. Specifically solutions were developed which allowed for upload of quantitative trait loci (QTL) (as shown in Figure 3) and microarray data sets. The microarray data tools both supported the Minimal Information about a Microarray Experiment (MIAME) guidelines [19]. These tools were MIAMExpress [20] and MaxDLoad [21]. Metadata capture and validation needed to support the further reuse of QTL data sets was made and is shown in Figure 3. The security infrastructure itself was based upon the Privilege and Role Management Infrastructure Validation Service (PERMIS)

technology [22], which provided a role based authorisation infrastructure where fine grained data access was supported. The experiences in developing and applying security infrastructures based on PERMIS are described in [23].

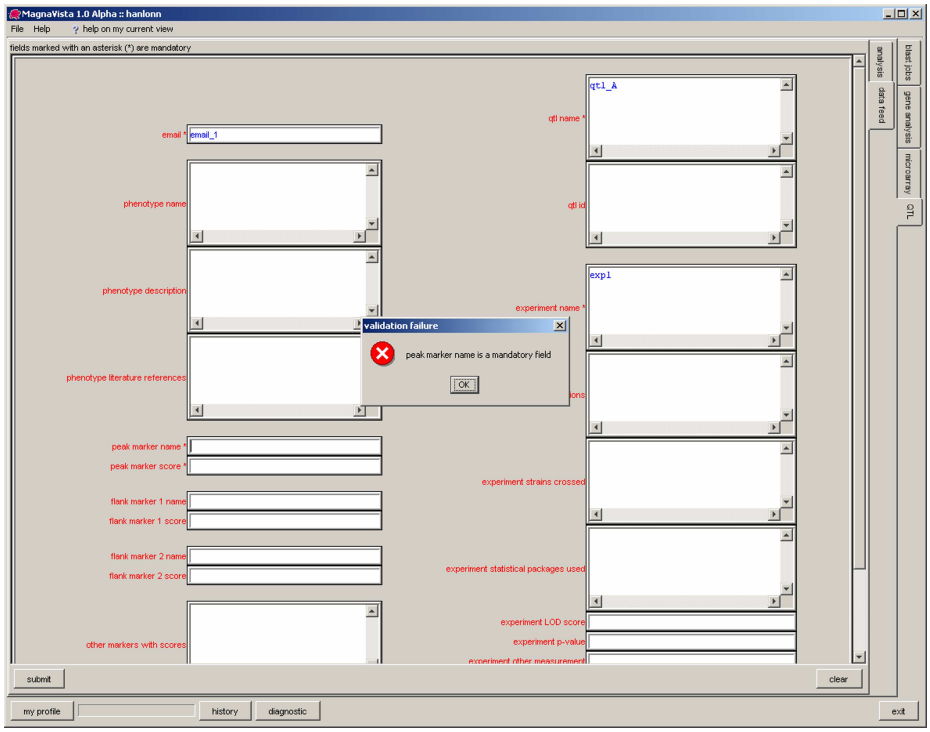

Fig. 3. Quantitive Trait Loci (QTL) upload facility with metadata validation

Despite the development of this security oriented data Grid infrastructure, it was largely the case that the scientists did not fully exploit it. There were many possible reasons for this. The data Grid providing access to the public genome repositories needed perpetual maintenance. Within the lifetime of the project, numerous evolutions and changes of data repositories occurred. Each evolution required extensions and refinements to the data Grid to accomodate schema changes for example. Another cultural reason is also that the scientists themselves have their own way of undertaking their research. Whilst internet hopping is not ideal and the Grid allows multiple genome repositories to be queried in one fell swoop, the scientists themselves are adept at using search tools such as Google for finding information associated with for example their genes of interest. It was also the case that the scientists were more comfortable and trusting in dealing with data sites such as *ensembl* directly, rather than through a Grid middleware layer. Knowing and trusting that the data is up to

date and from the actual live repository, rather than possible via a downloaded version of the data which had been inserted into a local DB2 repository was especially important for the scientists.

One of the key experiences in developing this infrastructure was the cultural barriers scientists had with how their own data sets might be shared. It was largely the case that the scientists were unwilling to share their data sets with one another. This fact is an important consideration which research councils in the UK have recognised and taking steps to address in defining their data sharing policies. Scientists are both collaborators and competitors. Being the first researcher to identify the genes which cause or are indirectly involved in causing hypertension can be both financially beneficial from subsequent grants and industrial interest, as well as gaining international recognition from peers in the scientific community. Whilst it is the case that leading journals are now requesting that data upon which papers they publish are based, this information is often of limited use. Firstly these data sets are often published potentially years after the experimental data was produced and with the rate of scientific insights, this often means that the data sets and results are superceded. Secondly the data sets themselves are often not informative enough for others to be able to repeat or verify the experiment, e.g. all necessary metadata describing the experiment, how the samples were prepared, how the data was normalised etc needs to be given.

The large scale challenges in building live data Grids in the life science domain requires an on-going and continued effort. Grid based virtual organisations which allow scientists to collaborate does have merits and is a model which should be taken forward. The BBSRC funded GEMEPS project is based upon a collaboration between NeSC and the SHWFGF at the University of Glasgow, the RIKEN Institute in Japan [24] and the Computational Biology Service Unit in Cornell University [25].

3 Data Grids within GEMEPS

GEMEPS is based upon the premise that scientists recognise that it is to their advantage to collaborate. Academics and researchers will always need to refer to and publish in journals and leading publications in their respective fields, however targeted real time access to research data between collaborators and institutes needs to occur to expedite the knowledge discovery process. Experiences from BRIDGES have shown that the scientists and their supporting IT staff, have to be fully informed and in control of the security infrastructures by which they make their data sets available and to whom.

Whereas BRIDGES had a focus on a range of functional genomics related data sets where scientists where interested in retrieving a variety of information on a specific gene or genes, GEMEPS aims to develop a Grid infrastructure for discovery, access, integration and analysis of microarray data sets. Through the GEMEPS infrastructure scientists should be able to support scenarios such as:

- who has run a microarray experiment and generated similar results to mine;

- show me the results from a particular collaborator;
- show me the conditions and analysis associated with experimental results similar to mine;
- show me all results for a particular phenotype, or for a given cell type or given pathogen;
- show me all results for a particular microarray chip set;

There are several large scale repositories that exist specifically for storage of microarray data sets. Some of these include Gene Expression Omnibus (GEO) at NCBI [26], ArrayExpress [27] and CIBEX [28]. As well as storing microarray data sets, these repositories also provide various kinds of services through which the repositories themselves might be searched or mined. These repositories typically require data sets to be MIAME compliant.

The stated goal of MIAME is to *outline the minimum information required to interpret unambiguously and potentially reproduce and verify an array based gene expression monitoring experiment* [19]. Whilst the details of particular experiments themselves may be different, it is the intention of MIAME to define a core that is common to most experiments. It should be noted that MIAME is not a formal specification, but rather a set of guidelines which concentrate on the content of information. It is not in itself a data format but provides a conceptual structure for capturing the metadata associated with microarray experiment descriptions. A MIAME description will typically describe the design of the *array platform* and of the *gene expression experiment*. The array design specification consists of the description of the common features of the array as the whole, and the description of each array design elements, e.g. each spot. The gene expression experiment description includes a description of the overall experimental design; the samples used; how extracts were prepared; which hybridisation procedures were followed and ultimately what data was measured and how it was analysed and normalised.

MIAME compliance is not prescriptive in the sense that all or a given subset of the various sections that might be associated with a given experiment must be given. These sections are usually provided in free text format, along with recommendations requiring *maximum use of controlled vocabularies or external ontologies*. MIAME recognises that few controlled vocabularies have been fully developed, hence it encourages users to provide their own qualifiers and values identifying the source of the terminology. Of those that are available, the Microarray Gene Expression Data Society (MGED) [29] is one of the more established ontologies for microarray experiment description. Several data formats have been defined and applied across different sites and with different user communities. These include: MAGE-ML [30], SOFTtext [31], MINiML [32] and SOFTmatrix [33].

MAGE-ML is part of the MGED family of standards and is MIAME compliant and XML based. Libraries for handling MAGE-ML exist for Java, C# and Perl with a python version in development. Many major repositories, such as GEO, ArrayExpress and CIBEX support results being deposited in MAGE-ML as well as supplying data in that format.

SOFTtext is a simple text based format designed by GEO. Unlike MAGE-ML, SOFTtext is not XML based using instead keywords for describing platform, sample and results. It has fewer fields than MAGE-ML yet is still MIAME compliant. GEO supports submissions in this format and makes results availiable in it as well. Since SOFTtext is based around a simple format it is easy to parse and use.

MIAME Notation in Markup Language (MINiML) is an XML based format used by GEO and is equivalent to SOFT. The NCBI accepts data deposited in MINiML format and makes records available in this format. MINiML can be considered an XML equivalent to SOFTtext as it provides the same properties, however in XML form. NCBI has made a schema for MINiML avaliable allowing a validating parser to confirm that a MINiML file is well formed. This is a distict advantage over SOFTtext where there is no formal definition of how the files should be formatted. As with the other SOFT formats MINiML is MIAME compliant yet has fewer fields than MAGE-ML. The relative simplicity of MINiML when compared to MAGE-ML has direct advantages for usability and associated learning curve.

SOFTmatrix is a new format based on a spreadsheet. Like SOFTtext it was developed by the NCBI based on MIAME. The format uses Microsofts Excel .xsl files as a base and consists of a simple template. Given the extensive use of Excel in processing microarray results by the biological community, using it as a form of exchange format was arguably inevitable. It should be noted that the .xsl format is proprietary and its format is not officially published in the public domain. As a result, long term usage may be a potential issue due to potential licensing issues.

As seen a multitude of on-going efforts in how to describe and annotate the data and metadata associated with microarray experiments and results exist. It is within this context that the GEMEPS project is developing a security oriented Grid infrastructure for microarray experiment profiles.

3.1 GEMEPS Data Grid Architecture

The overall architecture of the GEMEPS data Grid is depicted in Figure 4 with shaded areas depicting what has been implemented thus far.

In the course of the GEMEPS project, an evaluation of existing state of the art technology for gene expression profiles was undertaken. The Cell Montage (CM) software [34] provides a solution that supports searches of gene expression repositories typically where the expression data sets are kept in files and directories. CM supports two different kinds of expression profile search. A typical microarray experiment will result in several thousand genes and their associated levels of expression being recorded. CM establishes the similarity of two gene expression profiles by comparing the order of genes ranked by their expression. (This is based upon the Spearman rank correlation co-efficient). Although this is a simple measure it has been observed that it is sufficient to characterize cell types across different microarray platforms. CM also supports scenarios where

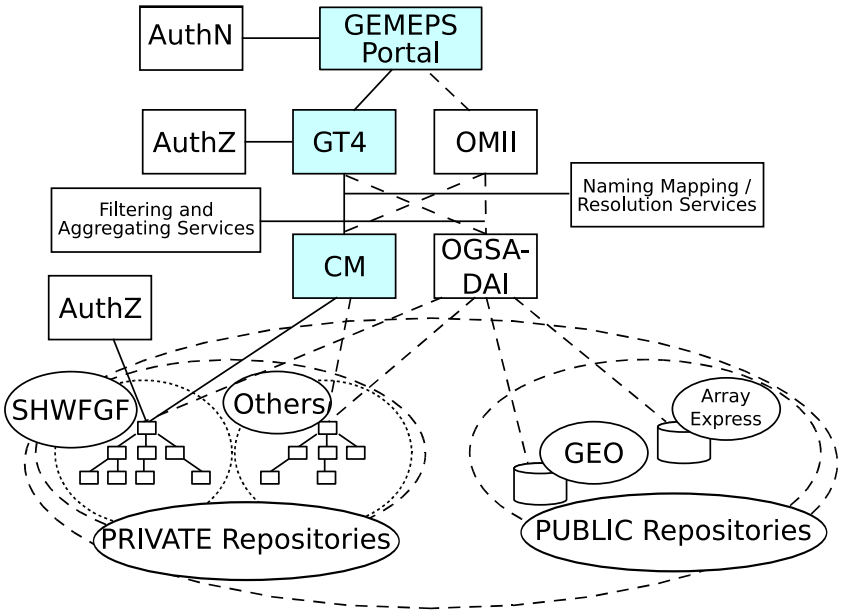

Fig. 4. Overall GEMEPS Data Grid Architecture

the gene names and their expression values are used for searching. The benefits of this method are that potentially important values are maintained, e.g. where differences in expression values are significant. However, due to the inherent limitations in the accuracy and reliability of microarray experiments, a more accurate assessment is often based upon the relative expression orderings and not on the values of the expression *per se*.

The GEMEPS project has Grid-enabled the CM technology using Globus toolkit version 4 [35] and made this available within the GridSphere portal environment [36]. The shaded areas in Figure 4 represents the current implementation status of GEMEPS. At the time of writing we have focused predominantly on microarray data sets within the University of Glasgow, however we plan in the near future to extend the infrastructure to the other partners and to incorporate public repositories. It is also planned that a case study will be undertaken using the Open Middleware Infrastructure Initiative (OMII) middleware stack [37] including the Open Grid Service Architecture Data Access and Integration (OGSA-DAI) [10] components. This will allow experiments across a range of databases and repositories to be incorporated. Some of the specific challenges of this domain that are being tackled within GEMEPS include naming and name resolution of genomic informations and security, which we outline here.

Name Mapping and Resolution Services One of the primary challenges that must be overcome in supporting this work is naming resolution of gene

identifiers and the associated experimental and array informations. Being able to compare the results of different experiments fundamentally depends at the very least upon being able to assert a relation between the gene names or platform specific information between the experiments. Unfortunately repositories and individual sites typically use different naming conventions such as *entrez* and *unigene*. Accession numbers have also been introduced as a mechanism to uniquely identify genes and establish correspondences between information stored in different or in some case the same repository. For example, the NCBIs GEO data set is available in both MINiML and SOFT formats but the two are not equivalent. There are many more SOFT files than MINiML but not all of the entries are available in one format or another.

As a result the GEMEPS architecture is developing services that allow for correspondences to be established between gene names. This has included a detailed exploration of the Life Science Identifier (LSID) initiative [38]. LSIDs are designed as a Uniform Resource Name (URN) based identifier which itself is a form of Uniform Resource Identifier. LSIDs themselves are written in the form: `urn:lsid:<authority>:<database>:<object>:<version>` where `<authority>` is the name of the authority who issued the LSID, `<database>` is the name of the authoritys database the LSID is stored in and `<object>:<version>` identifies the object within the database and its revision.

LSIDs are intended to serve as persistent identifiers allowing them to be used without later being reassigned. They allow to map to exactly the same set of bytes *permanently*. This means that an LSID, once assigned, is permanently attached to a specific encoding of its data which cannot be updated or corrected. An immediate advantage of this is that makes LSIDs usable as references. LSIDs also support attaching metadata, in a variety of forms, allowing an automated parser to discover for instance, synonyms, creation information and alternate versions of the LSID. The versioning field at the end of the LSID is optional but can be used to differentiate between revisions of the object or different representations as well. When there is a mapping from an existing datasets accession number to an LSID it is possible for previous accession systems to generate an LSID for their data making any program that uses LSID able to access a wider range of data. No standard mechanism for performing this transform is defined however, hence this makes the use of automatically generated LSIDs by a program risky until a recognised authority formally assigns them.

The LSID specification suggests using an LSID proxy, e.g. `lsid.biopathways.org`, to resolve LSIDs. The biopathways resolver provides LSIDs for many existing data sets such as the NCBI databases, ArrayExpress and SwissProt for example. However relying on a sole point of access is dangerous as in the event of its failure, all of the data sets accessed through the proxy will become unavailable. A model with independent authorities is more robust as the loss of one authority results in a smaller loss. Conversely, having a great many authorities ensures that, at any given time, some of the authorities will be unavailable. Whilst there is no mechanism for reserving LSIDs, there are mechanisms for requesting that valid LSIDs exist. At the time of writing, it is unclear whether

LSIDs will solve the problems arising in uniquely identifying information in the life science domain. For example, the closure of the Interoperable Informatics Infrastructure Consortium (I3C) means the loss of RDF metadata associated with LSIDs. References to this data still appear in examples and tutorials but the I3C itself website no longer exists. The only implementations of the LSID stack found are from the IBM LSID project on sourceforge. There are two implementations available one in Java the other Perl. The logs of the source repository reveal little activity with the majority of the code remaining untouched since 2004. To address this, other more pragmatic solutions based upon for example, local hash tables and schemas for cross referencing gene expression naming information are being considered within GEMEPS. Whilst suitable for demonstration and prototype production within the lifetime of GEMEPS, this will ultimately be a short term solution. A common standard and agreement adopted by the life science community is urgently required.

Security Components Security is crucial to the scientific community that we are planning to support. Typically security in the context of the Grid is split into three areas: Authentication whereby the provider of a given resource can verifying the identity of a particular user, most commonly through public key infrastructures and X.509 digital certificates; Authorization whereby a resource provider is able to assign and attest that an authenticated user has sufficient privileges to access their resource; Accounting whereby the unambiguous recording and logging of actions by specific users is made which can subsequently be used for auditing or potentially non-repudiation in the event of security breaches for example.

Within GEMEPS authentication is made directly through access to a portal using a specific username and password that has been allocated to VO members. Once logged-in the user is given access to various portlets depending upon their level of privilege in the portal. For example, whether they are able to access data sets at a remote VO site is dependent upon them having sufficient privileges to access that site. This in turn is dictated by the agreements that exist between the VO partners in describing what data they are willing to make available to one another. All users of the portal are given access to the public microarray repositories. The actual authorisation decision itself will be made by a combination of a policy enforcement point (PEP) and a policy decision point (PDP). The former is an API specific to the Grid services which is invoked when a user issues a query. The latter is the source of the actual policies itself. We are currently exploring a variety of different technologies for supporting such scenarios including PERMIS and lighter weight inhouse solutions based upon access matrices.

When a user issues a query via the portlet to a Grid service, this service will ensure that this user is privileged to issue that query. The PDP associated with the portal itself will provide the initial decision on whether this request is a valid one for this particular user and the associated data sets. Ultimately however, a local VO site will want to enforce its own access decisions. Without this, the basic model of a site giving secure federated access to their data is

broken. That is, they are simply delegating the access decisions to a remote party which is a model that will not gain widespread acceptance by the security focused life science community. Instead, sites may wish at any time to change their own policies on data access and usage and so maintain their own autonomy. As such, when a query is sent to a secure VO site, they will also wish to verify that firstly the request is from a valid member of the VO, and secondly that the query is in accordance with their own local security policy on data access and usage as agreed within the VO. It is our intention to work closely with the GEMEPS project VO partners to demonstrate how authorisation models can be defined and enforced on the Grid and supported within their own remote IT infrastructures.

4 Implementation Status of GEMEPS

At the time of writing the GEMEPS project has been on-going for 5 months and has a further 7 months remaining. Whilst it was our initial intention to build directly upon the work undertaken in BRIDGES, it was recognised that other solutions such as Cell Montage would address many of the issues we envisaged having to overcome ourselves. As such, we have refined our plans to incorporate these solutions whilst still maintaining elements from our original plan.

This rethinking of the project is in turn opening up numerous research and implementation avenues. For example, whilst the Cell Montage software allows to perform expression profile searches on a directory structure, it does not yet support searches of heterogeneous platforms and multi-site databases. Cell Montage is itself an independent application and not a database access, query and integration driven approach such as OGSA-DAI. To explore and contract these issues we plan to develop OGSA-DAI based solutions (accessible through both GT4 and OMII-UK Grid middleware). This dual evaluation will also allow us to make numerous comparisons including performance between the different technologies, and importantly how they impact upon the security infrastructure. Thus whilst we have Grid enabled Cell Montage to work over a local directory structure at Glasgow, having this application run remotely over a secure VO partner directory/file structure may well raise issues on security in itself. Remote partners may well be happy allowing a query of a database or specific tables within a database, but running an application across an internal directory may well require more detailed negotiations. We have seen already through numerous other projects how it is possible to build OGSA-DAI based solutions to access restricted views of remote database tables. In this case the user view of the data often corresponds with the role of the user as supported within the authorisation infrastructure.

From a performance perspective we have already seen that Cell Montage allows to run a gene expression comparison over several thousand experiments typically in the order of 0.5-1.5 seconds. We do not expect the full distributed Grid version of this application to be of that order (based upon experiences in other projects).

Once these different versions of the gene expression profiles are available we also plan on developing filtering and aggregation services. This will allow results to be merged based upon numerous different criteria. Thus a user might only wish to see result for a specific platform (such as a given Affymetrix chip set) and only the first 10, 100 or 1000 genes; or only results for a particular type of experiment. This in turn will impact upon how the queries themselves are formulated, and in the case of OGSA-DAI how the subsequent database JOINS are made and subsequently merged with the results from the Grid enabled Cell Montage software.

The basic alpha prototype of GEMEPS is shown in Figure 5. This shows how the Cell Montage software has been Grid enabled (using GT4), integrated into the GridSphere portal environment, and used to upload a gene expression profile (top of Figure 5) with the corresponding matching profile result set shown on the bottom of Figure 5.

5 Conclusions

Grid technology offers many potential benefits to the life science community. We believe that the primary challenges that must be overcome are in managing the explosion of data across the various life science research domains. The BRIDGES project built a data Grid for functional genomics to investigate the genetic causes of hypertension but to all intent and purposes this has become obsolete since the project ended (at the end of 2005). The data models and schemas upon which this infrastructure was built have evolved. These experiences will continue to prevail until standards are agreed and adopted by the wider community. Whilst we acknowledge that standardisation too early is a bad thing and as new insights and scientific discoveries are made, it is likely that updates and extensions to data models and services will naturally occur, we believe that putting in place guidelines and generic solutions can help in minimising the chaos. LSIDs for example tackled a problem which is common to all life science research: how to uniquely and permanently name things and subsequently avoid discrepancies arising. Sadly these efforts appear to have dissipated and the wider life science community will continue for the time being at least to develop short term solutions.

These technological impacts are a difficult but not insurmountable problem and solutions can be engineered to overcome the issues of lack of global name spaces. Another perhaps more challenging aspect associated with the success of the Grid in this domain is in supporting the needs of the scientific community. Middleware push has to be replaced by scientific pull. Experiences from BRIDGES have shown that the scientists can be a fickle community, yet they are the customers and whatever their feedback, it has to be taken seriously. Time will tell within the GEMEPS project if this community are willing and happy to use the infrastructure and securely share their data sets.

It is also our intention to use the experiences in the BRIDGES and on-going GEMEPS project to contribute to the recently started Scottish Bioinformatics

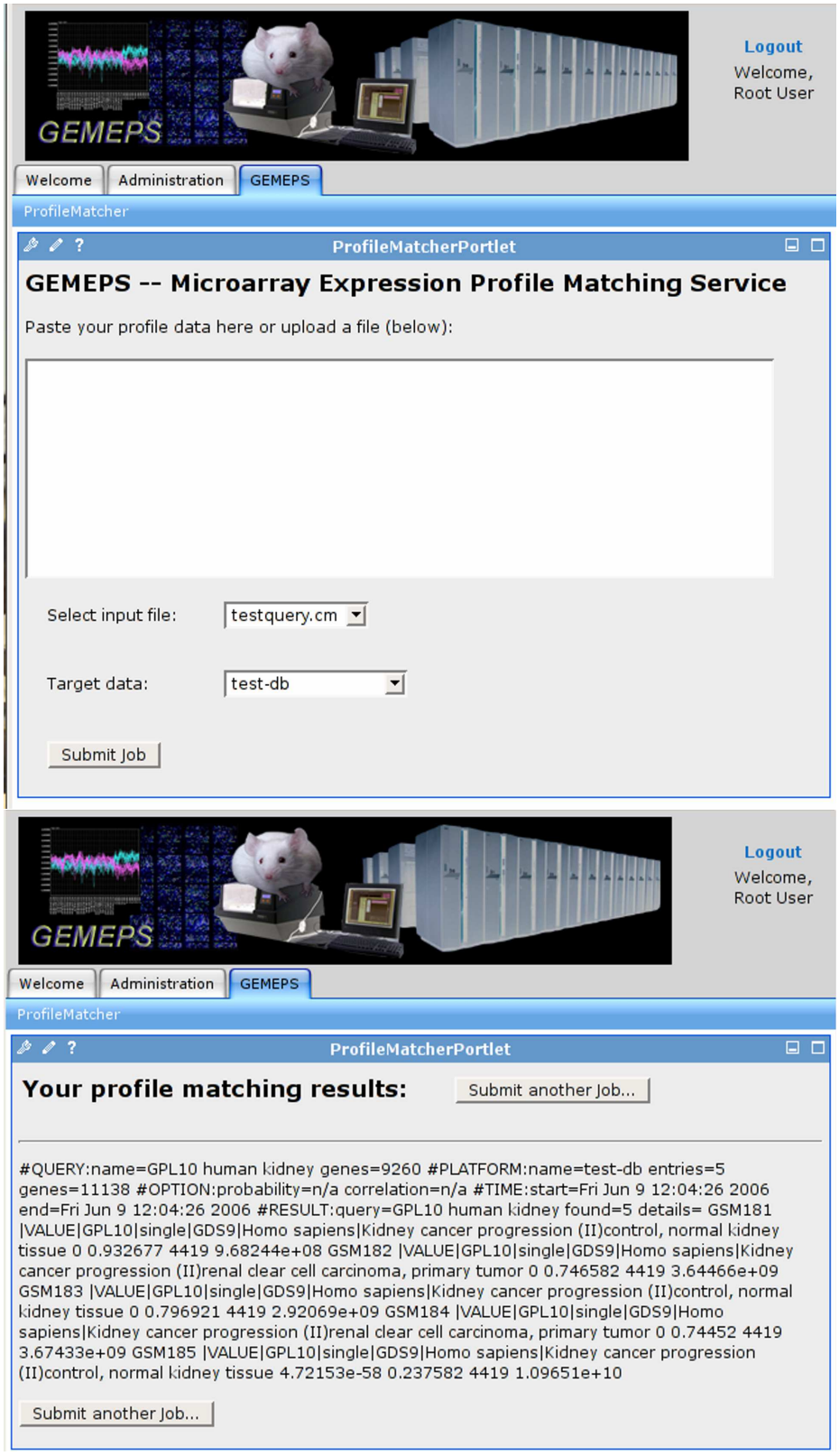

Fig. 5. Basic Interface and Usage of Alpha Version of GEMEPS Portal

Research Network [39], which has the intention to build a world class bioinformatics research infrastructure across Scotland.

5.1 Acknowledgements

The work described here was suported by grants from the UK Department of Trade and Industry (DTI) and the Biotechnology and Biological Sciences Research Council (BBSRC). We gratefully acknowledge their support. We also acknowledge inputs to the science we are trying to support from Dr Pawel Herzyk at the Sir Henry Wellcome Functional Genomics Facility at the University of Glasgow.

References

1. "Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) project." URL. `http://www.nesc.ac.uk/hub/projects/bridges`.
2. "Cardiovascular Functional Genomics (CFG) project." (URL). `http://www.brc.dcs.gla.ac.uk/projects/cfg`.
3. R. O. Sinnott, "Security focused federation of distributed biomedical data," in *Proceedings of UK e-Science All Hands Meeting*, (Nottingham, UK), September 2003.
4. R. O. Sinnott, M. Bayer, D. Houghton, D.Berry, and M. Ferrier, "Development of a grid infrastructure for functional genomics," in *Proceedings of Life Science Grid Conference (LSGrid 2004)*, June 2004.
5. R. O. Sinnott, M. Bayer, D. Berry, M. Atkinson, M. Ferrier, D. Gilbert, E. Hunt, and N. Hanlon, "Grid services supporting the usage of secure federated, distributed biomedical data," in *Proceedings of UK e-Science All Hands Meeting*, (Nottingham, UK), August - September 2004.
6. R. O. Sinnott and M. Bayer, "Controlling the chaos: Developing post-genomic grid infrastructures," in *Life Science Grid Conference (LSGrid2005)*, May 2005.
7. R. O. Sinnott, M. Bayer, J. Koetsier, and A. J. Stell, "Advanced security on grid-enabled biomedical services," in *Proceedings of UK e-Science All Hands Meeting*, (Nottingham, UK), September 2005.
8. "Grid Enabled Microarray Expression Profile Search (GEMEPS) project." URL. `http://www.nesc.ac.uk/hub/projects/gemeps`.
9. "IBM Information Integrator." URL. `http://www-306.ibm.com/software/data/`.
10. M. Antonioletti, M. Atkinson, R. Baxter, A. Borley, N. P. C. Hong, B. Collins, N. Hardman, A. C. Hume, A. Knox, M. Jackson, A. Krause, S. Laws, J. Magowan, N. W. Paton, D. Pearson, T. Sugden, P. Watson, and M. Westhead, "The design and implementation of Grid database services in OGSA-DAI," in *Concurrency and Computation: Practice and Experience*, vol. 17, pp. 357–376, Febuary 2005. `http://www.ogsadai.org.uk`.
11. R. Sinnott and D. Houghton, "Comparison of data access and integration technologies in the life science domain," in *Proceedings of UK e-Science All Hands Meeting*, (Nottingham, UK), September 2005.
12. E. Birney, D. Andrews, M. Caccamo, Y. Chen, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, S. Graf, M. Hammond, J. Herrero, K. Howe, V. Iyer, K. Jekosch,

A. Kahari, A. Kasprzyk, D. Keefe, F. Kokocinski, E. Kulesha, D. London, I. Longden, C. Melsopp, P. Meidl, B. Overduin, A. Parker, G. Proctor, A. Prlic, M. Rae, D. Rios, S. Redmond, M. Schuster, I. Sealy, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, A. Stabenau, J. Stalker, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and T. J. P. Hubbard, "Ensembl 2006," in *Nucl. Acids Res.*, vol. 34, pp. D556–561, 2006. http://www.ebi.ac.uk/ensembl/.

13. J. T. Eppig, C. J. Bult, J. A. Kadin, J. E. Richardson, J. A. Blake, and the Mouse Genome Database Group, "The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology," *Nucl. Acids Res.*, vol. 33, pp. D471–475, 2005. http://www.informatics.jax.org/.
14. M. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore and M. National Center for Biotechnology Information, National Library of Medicine (Bethesda, "Online Mendelian Inheritance in Man, OMIM (TM)," http://www.ncbi.nlm.nih.gov/OMIM/.
15. "Human Genome Organisation (HUGO)." URL. http://www.gene.ucl.ac.uk/hugo.
16. "Rat Genome Database (RGD)." http://rgd.mcw.edu/.
17. M. Ashburner, C. A. Bal, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," in *Nature Genetics*, vol. 25, pp. 25–29, May 2000. http://www.ebi.ac.uk/GO/.
18. "Affymetrix microarray technologies." URL. http://www.affymetrix.com.
19. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (miame)[mdash]toward standards for microarray data," in *Nature Genetics*, vol. 29, pp. 365 – 371, December 2001. http://www.mged.org/Workgroups/MIAME.
20. "MIAMexpress." URL. http://www.ebi.ac.uk/miamexpress/.
21. "MaxDLoad." URL. http://bioinf.man.ac.uk/microarray/maxd/maxdLoad/.
22. D. W. Chadwick and A. Otenko, "The permis x.509 role based privilege management infrastructure," in *SACMAT '02: Proceedings of the seventh ACM symposium on Access control models and technologies*, (New York, NY, USA), pp. 135–140, ACM Press, 2002.
23. R. O. Sinnott, M. M. Bayer, J. Koetsier, and A. J. Stell, "Grid infrastructures for secure access to and use of bioinformatics data: Experiences from the bridges project," in *1st International Conference on Availability, Reliability and Security (ARES06)*, (Vienna, Austria), April 2006.
24. "Riken genomic sciences centre bioinformatics group." Yokohama Institute, Yokohama, Japan, http://big.gsc.riken.jp/.
25. "Computational biology service unit." Cornell University, Ithaca, New York http://www.tc.cornell.edu/Research/CBSU/.
26. T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "NCBI GEO: mining millions of expression profiles–database and tools," *Nucl. Acids Res.*, vol. 33, no. 1, pp. D562–566, 2005. http://www.ncbi.nlm.nih.gov/geo/.
27. A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen,

P. Rocca-Serra, and S.-A. Sansone, "Arrayexpress: a public database of gene expression data at ebi," in *Nucleic Acids Research*, vol. 31, pp. 68 – 71, Oxford University Press, January 2003.
28. K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori, and Y. Tateno, "Cibex: center for information biology gene expression database," in *Comptes rendus biologies*, vol. 326, pp. 1079–1082, Elsevier, October-November 2002.
29. "Microarray gene expression data society (mged) ontology working group." `http://www.mged.org/ontology`.
30. "MicroArray and Gene Expression Markup Language (MAGE-ML)." URL. `http://www.mged.org/Workgroups/MAGE/mage-ml.html`.
31. "Simple Omnibus Format in Text (SOFTtext)." URL. `http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html`.
32. "MIAME Notation in Markup Language (MINiML)." URL. `http://www.ncbi.nlm.nih.gov/projects/geo/info/MINiML.html`.
33. "Simple Omnibus Format in Matrix (SOFTmatrix)." URL. `http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html`.
34. "Cell montage gene expression profile search and analysis." URL. `http://cellmontage.cbrc.jp/cgi-bin/index.cgi`.
35. I. Foster, "Globus toolkit version 4: Software for service-oriented systems," in *FIP International Conference on Network and Parallel Computing*, vol. 3779, pp. 2–13, Springer-Verlag LNCS, 2005. `http://www.globus.org/toolkit`.
36. "GridSphere." URL. `http://www.gridsphere.org`.
37. "UK Open Middleware Infrastructure Institute (OMII)." URL. `http://www.omii.ac.uk`.
38. "Life science identifiers." `http://lsid.sourceforge.net/`.
39. "Scottish Bioinformatics Research Network (SBRN)." URL. `http://www.nesc.ac.uk/hub/projects/sbrn`.