

A Distributed System for E-Support of Microarray Data Analysis and Management

Barton, G^{*}., Sternberg^{*}, M.J., Game, L[§]., Chiba, N[§]., Huang, Y. ⁺, Darlington J⁺, Tomlinson, C[§]., Saleem, A⁺., and Butcher, S^{*}.

** Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London*

§ CSC/IC Microarray Centre, Hammersmith Hospital, Imperial College London

+ London e-Science Centre, Department Computing, Imperial College London

Abstract

Microarray technology is increasingly being used as a powerful tool in functional genomics related investigations. However, the proliferation of various tools and techniques has made it very challenging for a lab scientist to keep up with all the latest developments. Presented here is an extensible microarray data analysis portal which provides access for the user to distributed data analysis and management resources from a single user interface.

Introduction

The new post-genomic technologies, such as microarrays and proteomics, are producing new challenges for the biological community. All new technologies require new skills for interpretation but these latest high throughput technologies require expertise in data-management and exploration of multivariate data; new challenges for many biological researchers. All too often researchers are unable to extract the full benefit from the investment in their research due to difficulties in applying best bioinformatics practice to their experiments. This problem is set to increase as more researchers adopt these methodologies, particularly with the move towards integrative and systems Biology. The microarray technology is relatively new, being first described in 1996 [1] and, as such, new methods and software for analysis and data management are continuously being developed. The nature of the experiment requires the application of complex analysis methods as well as the use of non-trivial computer technologies to manage and store the large data sets. This, together with the proliferation of various tools and techniques for different stages of microarray data analysis, has made it very challenging for a lab scientist to keep up to date and implement all the latest developments.

Project Aims

This project aims to develop a distributed e-support system for microarray data analysis and management. The principal objective is the development of a portal system providing

simple, robust access to up-to-date resources for microarray data storage and analysis, combined with an integrated system to optimise user support and training using these amenities. The system will build upon several open-source technologies already available for microarray data analysis, combining them to be a fully integrated system. This should allow the user to perform data management and analysis tasks through a single portal. Imperial College is a useful test case for the system due to its microarray users being located across several dispersed campuses in and around the London area.

The proposed infrastructure will support:

- Data transfer between specialised sites and data repositories.
- Fast-track access for biological researchers to new models and algorithms developed in-house and externally e.g. by statisticians, computer scientists.
- Single point access via bespoke microarray portal to a range of tools and packages for microarray analysis with seamless data flow via a range of data reformatting tools.
- Fully automated tracking of the analyses performed.
- A functional system sufficiently robust and simple to use that it can be readily implemented and used by researchers and

support staff without additional e-science-centric system support.

- Access to online live expertise for researchers for data analysis and management support services, including remote audio-visual interaction of ongoing analysis between researchers and staff on different sites (e.g. shared live screen views, audio-video). This can be extended to include access for collaborating experts in other specialities e.g. statisticians.

- Integration of data access tools to publicly available data repositories.
- Design and implementation of structures for holding the intermediate stages of analysis .

The 3 major work components are organised as follows:

The e-science related tasks:

- Portal infrastructure.
- Integration of distance training and support tools.
- Integration of compute, data and analytical software packages into the portal.

The bioinformatics and support related tasks:

- Conceptual design of administrative and management tools infrastructure.
- Development of the user interface.
- Evaluation of the system in use in an active support setting feeding back into all areas of downstream development

The data management related tasks:

- Design of new data interrogation and export interface for MiMiR data warehouse.

The first phase of the project was concerned with the capture of user requirements, defining the software specification and initial implementation of a prototype data analysis portal. A range of potential user types were questioned including biologists new to microarray data analysis (this group will be the main user group of the portal), biologists experienced in microarray data analysis, statisticians and bioinformaticians who have been involved in microarray data analysis and who will potentially be supporting biologists with microarray data analysis.

Some early points of note from the requirements capture phase included:

- All users questioned used Microsoft Excel at some point during the data analysis process. Reasons for this included users familiarity with Excel, ability to change formats of data structure enabling data to be exported from one analysis tool and imported into another, users knowing how they have manipulated the data and what statistical functions they have applied to the data and for using it as a report writing tool. However there are several obvious disadvantages with excel in terms of microarray data analysis including that it is not designed for such large and complex

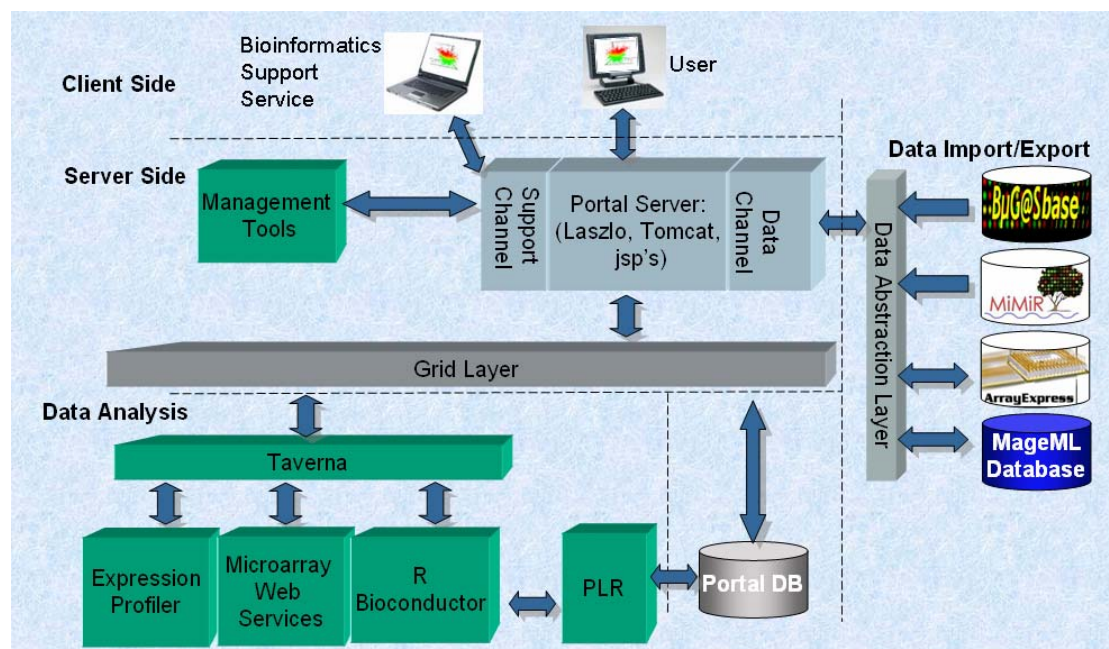


Figure 1: Schematic of Proposed Solution

datasets and analysis of data on this scale is prone to human error being introduced into the analysis steps;

- Users would like to have the ability to automate as much of the process as possible, yet still retaining full control over each step i.e. have the ability to create workflows and fine tune each step of these workflows;
- R, in combination with, Bioconductor is considered to be one of the most powerful, flexible and most widely used microarray data analysis tools, yet at the same time is extremely difficult for the beginner to use. It is driven from the command line and has its own object oriented programming language which is not easy to use, particularly those who are not familiar with programming concepts or statistical methodologies.

Implementation

1. Development of the user interface & portal:

The portal server is based on Apache-Tomcat. The user interface (as shown in fig 2) was developed using the web based Rich Internet Application (RIA) framework **OpenLaszlo** [2]. OpenLaszlo is an open source platform for creating zero-install web applications, which results in the end user does not having to deal with any complex installation issues.

Laszlo code is compiled into a Flash executable and as such the portal can be used in any browser that has Flash capabilities (including some of the most commonly used browsers – Internet Explorer, Firefox and Safari), therefore giving the portal cross platform compatibility. Although web based, the Laszlo user interface provides many of the capabilities and the look and feel of desktop client software due the Flash engine.

Figure 2 shows the microarray data analysis process and the relevant screenshots from the Laszlo portal interface for each stage of this process. Laszlo interacts with the server side portal components, such as the portal database, GridSAM and Taverna through Java Server Pages on a tomcat server.

2. Integration of compute, data and analytical software packages into the portal:

Several data analysis software packages have been integrated into the portal. The main analysis tool which has been integrated is **R**. R is an open source programming language and environment for statistical computing and graphics. Associated with R is Bioconductor [3] - a collection of packages which in the main part have been developed in R, dedicated to the analysis and comprehension of genomic data, in particular to microarray data analysis.

R is being used in this project because of the wealth of microarray functionality available. Several technologies for calling R were assessed including Rserve[4], Taverna[5], shell scripts and the Postgres PLR library[6]. For data analysis, the Postgres PLR library is being used. This allows R functions to be called programmatically within SQL statements. One of the main reasons for using PLR is that it allows data to be passed directly from the database to the R environment and back again. Therefore the results from each step of the data analysis can be easily serialised into the portal database, which greatly facilitates data handling, data persistence and analysis tracking.

A further advantage of using R for microarray data analysis is that it has a large number of quality assessment (QA) plots available. However, these plots can take some time to generate - from a few seconds up to several minutes - which can tie-up the analysis resources. To overcome this problem grid compute clusters were used to generate the plots in parallel using the GridSAM[7] and Grid Engine tools. GridSAM is an open source job submission and monitoring service which is used to pass and launch the QA plot jobs on Grid Engine managed cluster. The R environment is installed on each of the cluster nodes. The data, experiment description text file and a QA R script are passed to a compute node and R is invoked. The resulting QA plot is returned to the portal database and the image displayed in the interface. The small overhead in time required to copy the data is negated by the saving in time it takes to produce all the plots, and the freeing up of resources for other analysis tasks. This system is scalable to allow other computationally intensive analysis algorithms to be integrated.

The Taverna Workbench is an open source software package that allows users to construct complex analysis workflows from components

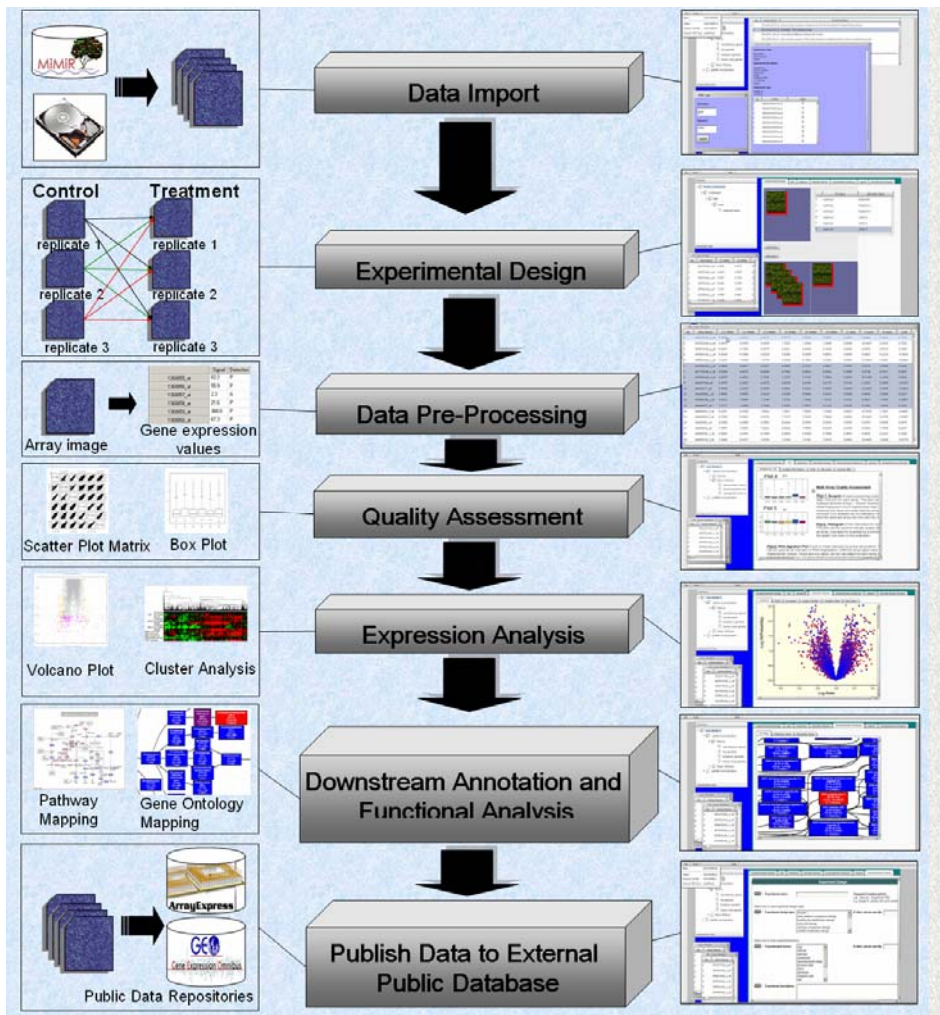


Figure 1: Overview of Microarray data analysis process and portal screenshots.

located on both remote and local machines. Taverna is being used in the prototype portal to develop standard microarray data analysis workflows. The workflow is then saved to the portal database in the Taverna SCUFL format. The portal can then run the workflow when invoked from the interface using the Taverna runtime engine. In the first instance Taverna workflows have been used to retrieve information on microarray results such as retrieving gene annotations, Gene Ontologies and Kegg Pathways.

3. Integration of distance training and support tools:

Several tools have been assessed for integration within the portal to provide real time analysis support for users. These tools include NetMeeting, VNC, VRVS, Webex and Co-pilot. Each has its own advantages (free, cross platform compatibility) and disadvantages (cost, single platform, hard for user to set up) over the others. Further

assessment of these tools will be carried out when the prototype portal is in use.

Training videos have been developed on are available from the portal interface to guide users through data analysis. Also being currently assessed is the use of several flash based multi-media tools available in Laszlo for real time user interactivity, such as web camera API's, coupled with the open source streaming media server red5 to provide live screen grabs of the support staff and user's interfaces.

4. Evaluation of the system in use in an active support setting:

A user feedback tool has been placed in the initial prototype which allows users to provide feedback as they use the portal. The inbuilt Laszlo debugger and tomcat server logs will be used to capture session information on the user's data analysis process. This will be used to determine any bottlenecks, the sources of

any bugs and ways in which to improve the user's analysis workflow.

5. The data management related tasks:

MiMiR[8] is a MIAME[9] (Minimum Information About a Microarray Experiment) compliant microarray data warehouse. Microarray experiment data and annotations are stored in MiMiR. Users can access and query MiMiR through the portal interface and import any data they have permission over into the portal. The portal accesses MiMiR via services on a MiMiR Data Access Server.

Users can also upload their data from the local file system through the portal interface

Future work

Although the functionality of the current portal is limited to standard Affymetrix data analysis process, the infrastructure and interface is in place to extend the number of features, according to user requirements in the testing phase. The functionality to analysis import data from different array platforms such as Agilent, Illumina and custom 2-colour arrays is already available in R, as is the functionality for the analysis of other array methodologies such as Exon Arrays and ChIP-chip arrays. Therefore the portal can be extended to use these features with relative ease.

There are plans to implement data access tools to publicly available data repositories. There are 2 main public microarray data repositories – ArrayExpress[10] and the Gene Expression Omnibus(GEO), as well as several more specialised databases such as Bugasbase and Oncomine.

The portal will allow users access to these resources through the use of mageML compliant tools and services. MageML (MicroArray and Gene Expression Markup Language)[11] is an xml schema used to describe microarray experiments and data, facilitating the exchange of such data. There are several tools available from EBI and also within R for the handling of mageML compliant data.

Milestones:

Phase I - It is proposed for the first phase, due April 2007, that the prototype system will build upon open-source software to develop a user friendly interface with automated data analysis pipelines, based on R for single

channel microarray data. The portal will have ongoing implementation of structures for holding the intermediate stages of analysis.

Phase II - A fully integrated, GRID-enabled system for E-support of microarray analysis, in a sufficiently robust, easy-to-use state that it can be effectively used in a support environment without extensive additional e-science IT support. Users will be able to Import of data from several repositories including ArrayExpress, Gene Expression Omnibus and BμG@Sbase. Access will be made available to the wider community through the portal, as well as the necessary components and installation scripts being available for download.

Acknowledgements

This project draws on the expertise of the Imperial College Bioinformatics Support Service in provision of high-quality bioinformatics core services and support (www.imperial.ac.uk/bioinformatics), the London e-Science Centre for GRID technology (www.lesc.imperial.ac.uk) and the CSC/MRC Microarray Centre for microarray data generation and management (microarray.csc.mrc.ac.uk).

We gratefully acknowledge the BBSRC for funding this work (BBS/B16488), the EBI Expression Profiler team, in particular Misha Kapusheshky (EP Software Development Coordinator), the Taverna Group and the MyGrid Group.

References

- [1] D Shalon, SJ Smith, and PO Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, 6(7):639–645, 1996.
- [2] Open Laszlo: open source platform for creating rich internet applications. <http://www.openlaszlo.org>.
- [3] RC Gentleman, VJ Carey, DM Bates, B Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, AJ Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, JY Yang, and J Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80, 2004.
- [4] Rserve - TCP/IP interface to R. <http://stats.math.uni-augsburg.de/Rserve/index.shtml>.

- [5] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Mark Greenwood, Tim Carver, Matthew R. Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, page bth361, 2004.
- [6] PL/R - R Procedural Language for PostgreSQL. <http://www.joeconway.com/plr/>
- [7] GridSAM - Grid Job Submission and Monitoring Web Service. <http://www.lesc.ic.ac.uk/gridsam/index.html>.
- [8] Mahendra Navarange, Laurence Game, Derek Fowler, Vihar Wadekar, Helen Banks, Nicola Cooley, Fatimah Rahman, Justin Hinshelwood, Peter Broderick, and Helen Causton. MiMiR: a comprehensive solution for storage, annotation and exchange of microarray data. *BMC Bioinformatics*, 6(1):268, 2005.
- [9] A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, CA Ball, HC Causton, T Gaasterland, P Glenisson, FC Holstege, IF Kim, V Markowitz, JC 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)-toward standards for microarray data. *Nat Genet*, 29:365–371, 2001.
- [10] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucl. Acids Res.*, 33, 2005.
- [11] PT Spellman, M Miller, J Stewart, C Troup, U Sarkans, S Chervitz, D Bernhart, G Sherlock, C Ball, M Lepage, M Swiatek, WL Marks, J Goncalves, S Markel, D Iordan, M Shojatalab, A Pizarro, J White, R Hubley, E Deutsch, M Senger, BJ Aronow, A Robinson, D Bassett, CJJ Stoeckert, and A Brazma. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biology*, 3:RESEARCH0046, 2002.