MICROARRAY SYSTEM
A System for Managing Data Produced by DNA-microarray Experiments

Alberto Calvi, Pietro Lovato, Simone Marchesini, Barbara Oliboni
Department of Computer Science, University of Verona, Verona, Italy

Massimo Delledonne, Alberto Ferrarini
Department of Biotechnology, University of Verona, Verona, Italy

Keywords: Microarray, Microarray data management.

Abstract: In this paper, we present the Microarray System which is based on a MIAME-compatible database and allows the users to store and retrieve data produced by experiments made with the DNA-microarray technology. This system was designed and implemented for managing data coming from the Functional Genomics Centre (FGC) of the University of Verona.

1 INTRODUCTION

Biological data produced by using the DNA-microarray technology are usually stored in the form of tab-delimited text files or excel spread-sheets produced by the microarray platform. Further information about the experiment, such as the name of the organism and various properties of the microarray chip, and people that participate to the experiment, are stored in different ways. For example, the title of the files is used to represent information related to organisms and samples used in the experiment.

For storing and managing data produced by microarray experiments, and for accessing and analysing information, several systems have been proposed (Fang et al., 2009; Zhu et al., 2008; Gattiker et al., 2009; Marzolf et al., 2006; Demeter et al., 2007; Vallon-Christersson et al., 2009). Moreover, with the goal of making data coming from microarray experiments fully available to the research community, some public online systems have been proposed. The most widely used are Gene Expression Omnibus (GEO) (Edgar et al., 2002), maintained by the NCBI, and ArrayExpress (Parkinson et al., 2005), maintained by the EBI.

Public systems require data standardization for easier storage and interoperability. The standard that has been proposed in the literature and is the most widely used for representing the information related to a microarray experiment is MIAME (Minimum Information About a Microarray Experiment) (Brazma et al., 2001). MIAME describes the minimum information that has to be provided to ensure that data and results can be easily interpreted; moreover, data standardization allows the comparison and the verification of experimental results.

In this paper, we consider data coming from the Functional Genomics Centre of the University of Verona (FGC), which works as an internal facility of the University and collaborates with other national and international academic institutions. The Centre is based on a Combimatrix array synthesizer that produces microarrays carrying either 12000 or 90000 oligonucleotide probes (Maurer et al., 2006), and currently has collected more than 650 gene expression experiments performed with more than 40 different chip designs.

In this context, an ad hoc database system for storing, managing and querying the huge amount of microarray data the FGC produces, is needed. The database should not only store microarray experiment results, but also descriptions of samples used in the experiment itself, and the way they were treated. The Microarray System was designed by considering the FGC specific needs. Moreover, it was developed to meet the MIAME standard to deal with the problem of sharing experimental results among researchers. Thus, the system supports at the same time a specific analytical method, adopted by the FGC, and the interoperability with other institutions.

Calvi A., Lovato P., Marchesini S., Oliboni B., Delledonne M. and Ferrarini A..
MICROARRAY SYSTEM - A System for Managing Data Produced by DNA-microarray Experiments.
DOI: 10.5220/0003137202930296
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2011), pages 293-296
ISBN: 978-989-8425-36-2

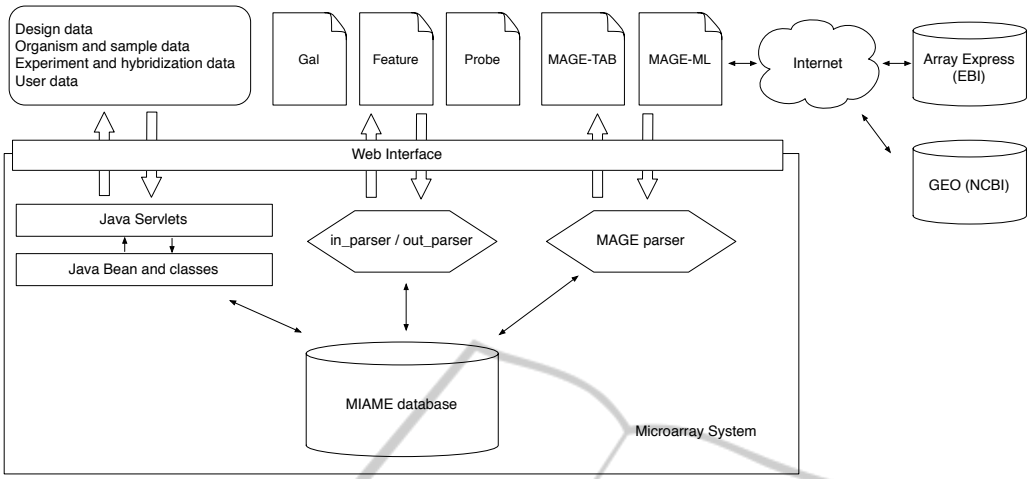

Figure 1: Overview of the Microarray System.

The Microarray System allows the user to retrieve data about every single microarray experiment and the associated biological information. Data collected from multiple arrays can be easily accessed; this means that genes expression of a set of related microarrays, belonging to a particular biological experiment, can be evaluated. Moreover, the system protects the information from unauthorized access, use, modification or destruction through an access control mechanism which, by means of user management and permission assignment, grants confidentiality, integrity and availability of data.

2 THE MICROARRAY SYSTEM

The Microarray System is a web application based on a MIAME compatible database, and achieves two different goals: it allows scientists of the FGC to collect and manage data produced by microarray experiments, and it ensures interoperability among different facilities by granting the possibility to express data in the MIAME format. In particular, the system allows users (i) to store data in the database, and manage them by interacting with the database itself, (ii) to extract data by the original files related to the use of the Combimatrix array synthesizer, and reconstruct them by querying the database, and (iii) to represent data extracted from the database by using MAGE-TAB (Rayner et al., 2006) and MAGE-ML (Spellman et al., 2002) standard formats.

Fig. 1 gives an overview of the system. An authenticated user can access, through a web interface, to data stored in the MIAME database by uploading and downloading information related to chip design, organisms, samples, experiment details and hybridization data. The same information can be exported, in the MAGE-TAB or MAGE-ML formats, and consequently uploaded to public repositories, i.e. ArrayExpress and GEO, or shared with other facilities.

2.1 The System Architecture

The Microarray System is a web application that has been developed following the *Model-View-Controller* architectural pattern (Leff and Rayfield, 2001) in order to isolate the application logic from the user interface (input and presentation), and to permit independent development, testing and maintenance of each layer.

The part related to the *Model* is the domain-specific representation of the data upon which the application operates and consists of a set of Java classes that manipulate data interacting with a persistent storage mechanism. In our case, data are stored in a relational database implemented in MySQL (Widenius and Axmark, 2002). The *View* part, implemented with the JavaServer Pages (JSP) technology (Bergsten, 2003), provides the user with an interface, via web browser, to interact in a transparent way with all the information stored in the database. For example, a user can perform several tasks like the search of experiment data, the execution of BLAST (Altschul et al., 1990), or the execution of information retrieval techniques. The servlet acts as the *Controller* and is in charge of the request processing and the creation of any beans or objects used by the JSP, as well as deciding, depending on the user's actions, which JSP page to forward the request to.

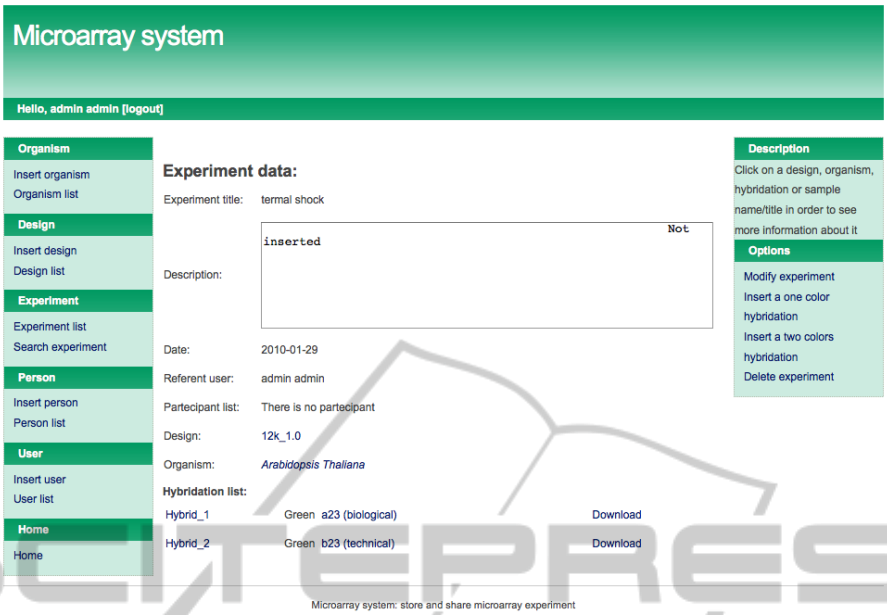

Figure 2: The *thermal shock* experiment page.

2.2 The System Features

The Microarray System allows the management of data produced by microarray experiments. The System allows users to store and manage information related to chip design, organisms, samples, experiment details and hybridization data (see left side of Fig. 1).

Moreover, the System is able to store information extracted from specific uploaded files. Information about the chip are usually maintained in a GenePix Array List (GAL) (Zhai, 2001) file that describes several aspects of a microarray design, such as the type of the chip, the number of spots, their dimension and position, and the respective probe sequence. GAL files are produced by using the Combimatrix GAL File Tool. The fluorescence data are instead organized in the tab-delimited text files named Feature and Probe files produced by the specialized software Combimatrix Microarray Imager that analyzes the scanned microarray image.

For storing into the database information extracted from the uploaded (input) GAL, Feature, and Probe files we designed and implemented the `in_parser`. In order to guarantee the backward-compatibility with external applications used for processing data, the Microarray System allows the user to reconstruct the GAL, Feature, and Probe files starting from data stored into the database by using the `out_parser` (see the central part of Fig. 1).

A user can interact with the system through a web interface, therefore we had to consider some information security aspects in order to realize mechanisms to protect the information and the system from unauthorized access, use, modification or destruction.

User creation is performed by the system administrator in order to bound the number of users accessing the system and meanwhile to record who performs operations on data. A system user can insert new information in the database, being automatically in charge of them, so he/she becomes responsible of assigning access permission for the information to other users (i.e., he/she can set reading and modifying privileges).

The search functionality is based on the experiment. From a search web page a user can set a single field, or a combination of fields. The entries that match searching parameters are filtered and showed according to the privileges the user has on the experiments.

At chip design level, it is possible to launch BLASTx within the system, to find homologies between genes used to design the array and known proteins in a database (UniProt/Swiss-Prot (Consortium, 2008)). The result is a tab-delimited text file containing a mapping between sequence *ids* and the UniProt/Swiss-Prot accession number, as well as other information about the execution. This file is then parsed in order to extract information and store them into the database of the Microarray System allowing for a future implementation of search feature of sequence *ids* to create direct links to the specific UniProt/Swiss-Prot online page. The original tab-

delimited file is also stored and can be downloaded by a user for usage with external software or further analysis.

In Fig. 2 we show the web page of an experiment. In the left side of the page a menu allows the user to access to suitable pages either to insert new information (i.e. *Insert organism*, *Insert design*, etc.), or to require the visualization of stored data (i.e., *Organism List*, *Experiment list*, etc.).

3 CONCLUSIONS AND FUTURE WORK

The Microarray System copes with the problem of managing the large amount of data collected from experiments made with the DNA-microarray technology by the FGC of the University of Verona. The proposed system is able to manage MIAME compatible information, achieving the needs of sharing data with the scientific community working on DNA-microarrays.

As future work, we plan to extend the database for considering other types of microarray, and in order to avoid the use of external applications, we plan to upgrade the Microarray System by implementing modules for data normalization and data analysis.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, W., and Lipman, D. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410.

Bergsten, H. (2003). *JavaServer Pages, 3rd Edition*. O'Reilly & Associates, Inc., Sebastopol, CA, USA.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., and Spellman, P. (2001). Minimum Information about a Microarray Experiment (MIAME) - toward standards for microarray data. *Nature Genetics*, 29:365 – 371.

Consortium, T. U. (2008). The Universal Protein Reasource (UniProt). *Nucleic Acids Res.*, 36:D190–D195.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J., Nitzberg, M., Wymore, F., Zachariah, Z., Brown, P., Sherlock, G., and Ball, C. (2007). The stanford microarray database: implementation of new analysis tools and open source release of software. *Nucl Acids Res*, 35:D766–770.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene Expression Omnibus (GEO): NCBI gene expressione and hybridation array data repository. *Nucleic Acids Research*, 30:207–210.

Fang, H., Harris, S., Su, Z., Chen, M., Qian, F., Shi, L., Perkins, R., and Tong, W. (2009). Arraytrack: An fda and public genomic tool. *Methods Mol Biol*, 563:379–98.

Gattiker, A., Hermida, L., Liechti, R., Xenarios, I., Collin, O., Rougemont, J., and Primig, M. (2009). Mimas 3.0 is a multiomics information management and annotation system. *BMC Bioinformatics*, 10:151.

Leff, A. and Rayfield, J. (2001). Web-Application Development Using the Model/View/Controller Design Pattern. *Enterprise Distributed Object Computing Conference, IEEE International*, 0:0118.

Marzolf, B., Deutsch, E., Moss, P., Campbell, D., Johnson, M., and Galitski, T. (2006). Sbeams-microarray: database software supporting genomic expression analyses for systems biology. *BMC Bioinformatics*, 7:286.

Maurer, K., Cooper, J., Caraballo, M., Crye, J., Suciu, D., Ghindilis, A., Leonetti, J., Wang, W., Rossi, F., Stöver, A., Larson, C., Gao, H., Dill, K., and McShea, A. (2006). Electrochemically generated acid and its containment to 100 micron reaction areas for the production of dna microarrays. *PLoS ONE*, 1(1):e34.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-serra, P., Sharma, A., Sansone, S., and Brazma, A. (2005). ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 33:553–555.

Rayner, T., Rocca-Serra, P., Spellman, P., Causton, H., Farne, A., Halloway, E., Irizarry, R., Liu, J., Maier, D., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C., White, J., Whetzel, P., Wymore, F., Parkinson, H., Sarkans, U., Ball, C., and Brazma, A. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BioMed Central*, 7:18.

Spellman, P., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B., Robinson, A., Bassett, D., Stoeckert, C., and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9):research0046.1–research0046.9.

Vallon-Christersson, J., Nordborg, N., Svensson, M., and Hakkinen, J. (2009). Base - 2nd generation software for microarray data management and analysis. *BMC Bioinformatics*, 10(1):330.

Widenius, M. and Axmark, D. (2002). *Mysql Reference Manual*. O'Reilly & Associates, Inc., Sebastopol, CA, USA.

Zhai, J. (2001). Making GenePix Array List (GAL) Files. robinsonlab.stanford.edu/microarrays/gal/docs/Making_GAL_Files.pdf.

Zhu, Y., Zhu, Y., and Xu, W. (2008). Ezarray: a web-based highly automated affymetrix expression array data management and analysis system. *BMC Bioinformatics*, 9:46.