Genomics & Informatics Vol. 3(1) 39-42, March 2005

Xperanto: A Web-Based Integrated System for DNA Microarray Data Management and Analysis

Ji Yeon Park[1], Yu Rang Park[1], Chan Hee Park[1], Ji Hoon Kim[1] and Ju Han Kim[1,2]*

[1]Seoul National University Biomedical Informatics (SNUBI), [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

DNA microarray is a high-throughput biomedical technology that monitors gene expression for thousands of genes in parallel. The abundance and complexity of the gene expression data have given rise to a requirement for their systematic management and analysis to support many laboratories performing microarray research. On these demands, we developed Xperanto for integrated data management and analysis using user-friendly web-based interface. Xperanto provides an integrated environment for management and analysis by linking the computational tools and rich sources of biological annotation. With the growing needs of data sharing, it is designed to be compliant to MGED (Microarray Gene Expression Data) standards for microarray data annotation and exchange. Xperanto enables a fast and efficient management of vast amounts of data, and serves as a communication channel among multiple researchers within an emerging interdisciplinary field.

Keywords: DNA microarray, database, information system, MAGE-ML, MIAME

Introduction

DNA microarray is a representative of high-throughput biomedical technologies that allow monitoring of gene expression for thousands of genes in parallel. DNA microarray experiment generates enormous amounts of data and they are meaningful only in the context of a detailed description of microarrays, biomaterials, and conditions under which they were generated. As microarrays are increasingly used in a variety of biomedical research, the systematic management of diverse and large amount of gene expression data has received much attention (Quackenbush, 2001). In particular, structured data storage and use of controlled terms with a well-being meaning are important for easy access and efficient use of the data.

With the growing needs of data sharing, Microarray Gene Expression Data (MGED) society has established common standards for microarray data. MIAME (Minimum Information About a Microarray Experiment) is a data annotation standard to specify the minimum information that should be reported about a microarray experiment (Brazma *et al.*, 2001). MAGE (MicroArray Gene Expression) standards an object model (MAGE-OM) and an XML-based language (MAGE-ML) have been developed for the representation of microarray expression data, facilitating the exchange of microarray information between different systems (Spellman *et al.*, 2002). MGED Ontology group defines sets of common terms and annotation rules for microarray experiments (Stoeckert *et al.*, 2002).

Although there are a number of microarray databases in existence, it is not easy to find the database meeting the specialized requirements of individual institutions or groups for local data archiving, analysis and sharing. The complexity of microarray data requires an integrated environment that does much of the data storage, visualization, analysis, and transfer. Based on this, we present Xperanto, meaning eXpressionist's Esperanto in XML, which is a microarray information system for microarray studies. To facilitate the comprehensive interpretation of microarray data, Xperanto system has links to the analytical tools and rich sources of biological annotation. It also has a capability of publishing the microarray data in a standardized format MAGE-ML to permit greater data sharing. As a result, it serves as a communication channel among multiple researchers within an emerging interdisciplinary field.

Methods

Database design

Xperanto database is basically designed to collect all contents defined by the MIAME standard and meet the specific requirements of laboratories. The database is comprised of six components: experiment, hybridization, biomaterial, array, protocol, and gene expression data

*Corresponding author: E-mail juhan@snu.ac.kr, Tel +82-2-740-8320, Fax +82-2-747-4830
Accepted 7 March 2005

(experimentally measured and computationally processed). By experiment we understand a set of one or more hybridizations that are in some way related, often corresponding to a particular publication. Values for data items come from look-up table that manages controlled terms from the MGED Ontology (http://www.mged.org/ontology/). To separate the data utilization according to research group, we implemented a security system allowing each investigator to securely manage their own microarray data and analysis results.

Development process

For the efficient development of the system, we analyzed the user's requirements so as to define the functional and behavioral aspects of the system. All the results were well documented to describe the various views of the microarray information system, and the document helps the communication among a wide range of the scientists involved in the production and analysis of microarray data.

Developmental environment

The system is implemented using MySQL relational database for data storage and has an Apache web server with PHP extension for the generation of server side application. For MAGE-ML export, we took advantage of MAGE-stk (MAGE Software Toolkit) software developed by MGED society, which act as converters between MAGE-OM and MAGE-ML.

User interface

Xperanto provides easy use for researchers by modeling a natural workflow in the microarray experiment. It also aims at encouraging more complete and accurate recordings with the structured data entry system such as drop-down lists of the MGED Ontology terms. Submissions are validated syntactically so that the records are organized correctly. The experimental data and their biological and statistical context are intra-linked on the system, thereby increasing the accessibility to the associated records. To accept various types of quantitative data from image-analysis programs, utility programs are available for modifying their format in system-defined way.

Results and Discussion

Xperanto is developed as a large, flexible, web-based database application that stores data from multiple microarray studies in the biomedical research. The system stores all types of data related to a microarray

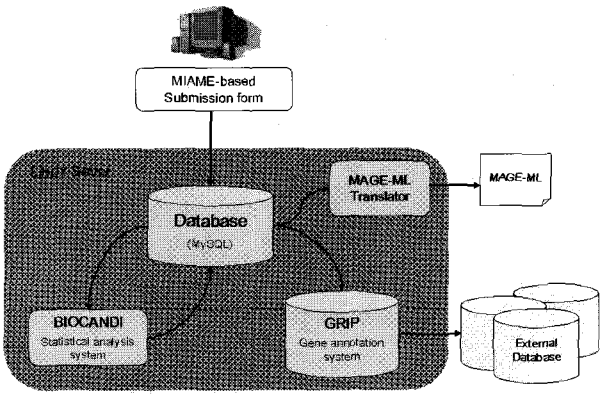

Fig. 1. System Architecture of Xperanto

experiment, including genes, biomaterials, experimental models, and quantitative data. It can accommodate different kinds of microarray-based experiments using single or two channel arrays measuring the abundance of mRNA molecules, as well as array CGH (Comparative Genomic Hybridization) measuring DNA copy number. Xperanto is currently used to manage 63 experiments and 1213 hybridizations from 22 research groups.

Fig. 1 shows overall architecture of Xperanto system.

Efficient data storage and management

Xperanto stores well-annotated data guided in MIAME standards, in addition to experimental gene expression data. Array type used in the experiment is recorded with precise descriptions of each element (spot) under the control of a system administrator. Biomaterial information is hierarchically separately managed; the source of the sample (e.g., organism, cell type or line), its treatment, the extract preparation, and its labeling. Protocols represent laboratory procedures and processing method (e.g., normalization), and are divided into several types such as treatment, labeling, hybridization, and analysis.

To ensure the comparison of transcriptional profiling data with same array design, the experiment should refer to the array assayed. Also, unique identifiers are assigned to all reusable parts of experiment description including biomaterial and protocol in order to store them in a granular, computer interpretable manner.

Once each microarray experiment is completed, an experimental researcher uploads all files of raw data, which are deposited in a protected directory structure. Each hybridization has image files scanned on an array and quantitation files generated from the image analysis (Fig. 2). However, in the case of quantitative data, the several data fields inside the file are decomposed into individual data elements, and their values are transferred to the database for further analysis.

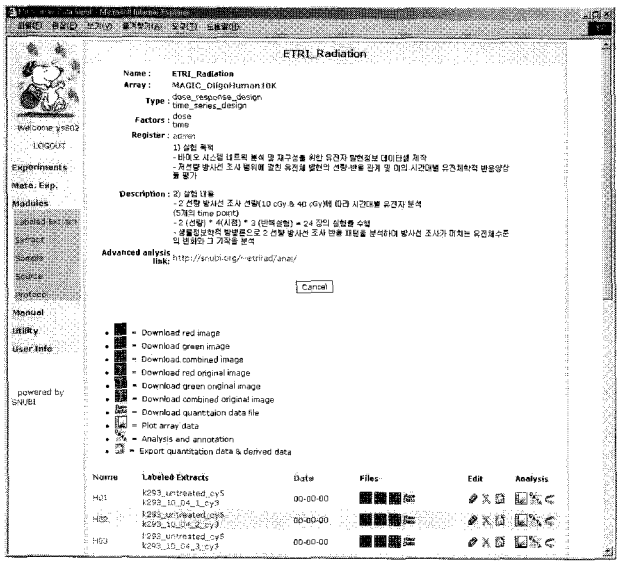

Fig. 2. Experiment description and the gene expression data uploaded

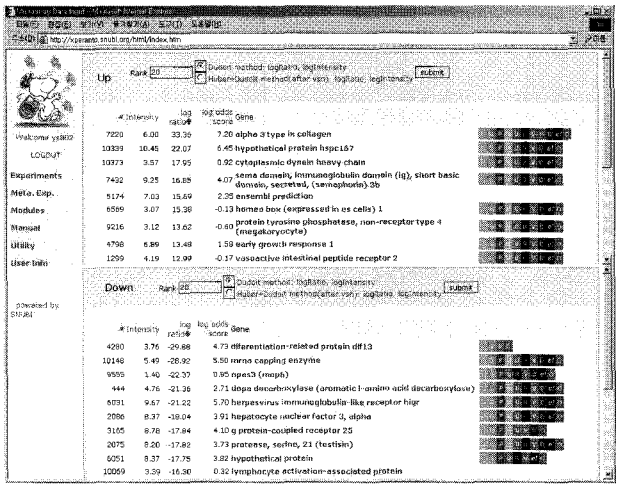

Fig. 3. Significant genes from statistical analysis and their annotation

Linking Xperanto to external resources

The inherent complexities of high-throughput gene expression data require the tight integration with a number of computational tools for their statistical analysis, data annotation and visualization. The integrated system accelerates the analysis of genome-scale datasets by reducing the time-consuming and repetitive activities.

For the basic data processing, Xperanto is linked to a statistical analysis package, SNUBI's BIOCANDI (BIOChip Analysis and Data Integration), which pipelines microarray data analysis procedures implemented in R statistical language. Once initiating the analysis, microarray intensity values are extracted from the database and prepared for the processing in the BIOCANDI. The data set is processed according to algorithms derived from diverse technological sources, and finally a list of significantly expressed genes in a single chip is detected. Plots are generated at each processing step such as filtering, transformation, normalization and discrimination, allowing the users to evaluate the experimental quality at a time.

After the statistical analysis, the genes with the significant expression change are represented with integrated annotations through SNUBI's GRIP (Genome Research Informatics Pipeline, http://grip.snubi.org/) system (Fig. 3). It refers to the functional genomic information including Gene Ontology categories, sequence feature information, protein information, and bibliographic information. Therefore, the integrated data annotation and analysis assists the researchers in discovering biologically useful knowledge from the massive data.

Easy data exchange with other systems

Growing needs of data sharing encourage the researchers to make frequent data releases compliant to microarray data standards. For them, we developed a program that translates Xperanto microarray data into MAGE-OM structure, and then exports it as MAGE-ML. By the user's request, the program reconstructs the related data fragmented in the relational tables, and generates a MAGE-ML data with three types of document such as array, hybridization and experiment (Fig. 4). The published XML file supporting the MIAME standards can be directly deposited to the public databases such as ArrayExpress (Brazma *et al.*, 2003) and GEO (Edgar *et al.*, 2002). For further analysis using various software tools, the system has a data release function to export Xperanto data in user-defined format (Fig. 5).

Future direction

As the microarray technology is developing fast, there are many demands for flexible solutions for the data management and analysis. For this reason, we are trying to add various analysis routines to perform more comprehensive investigations on microarray data, including clustering and data mining. Xperanto also continues to extend by integrating new resources to enhance the data visualization and annotation features.

As biomaterial data varies greatly according to the specific domains of application, it is difficult to capture their detailed and structured annotation in the database. Currently, limited sets of data fields are provided for the characteristics of sample and experiment because of the

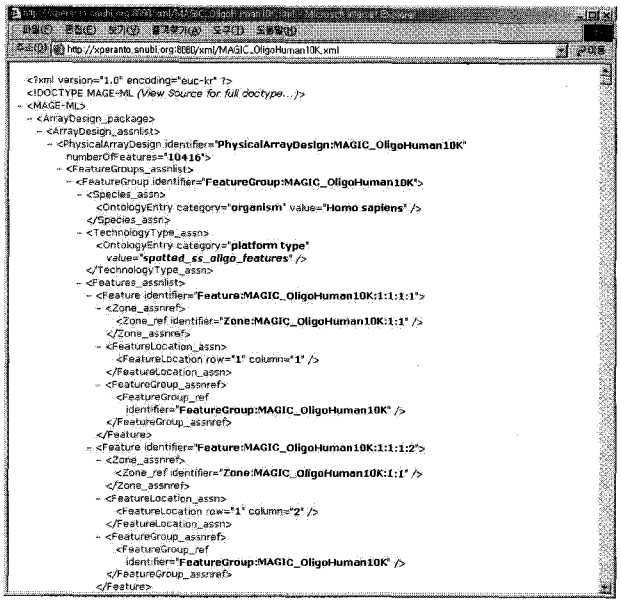

Fig. 4. MAGE-ML document showing microarray information

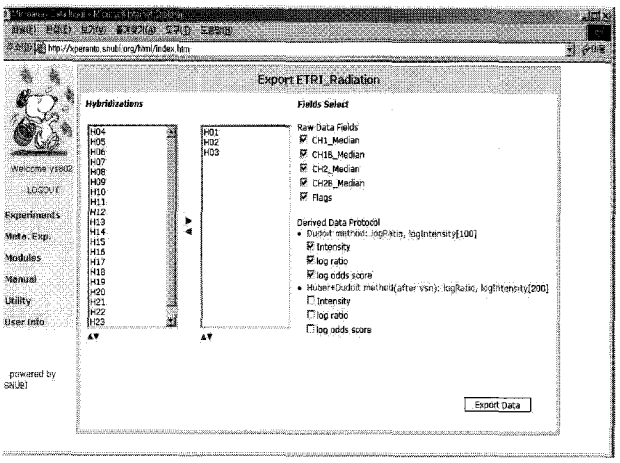

Fig. 5. Data export

great diversity of reporting structure for describing the data. We plan to provide different kinds of structured templates according to research area (e.g., toxicogenomics, cancer genomics). Through continual updates and modifications, Xperanto can serve as an important resource for genomics information and be used to develop data mining for the discovery of meaningful knowledge.

Acknowledgements

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0405-BC02-0604-0004).

Reference

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365-371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68-71.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization microarray data repository. *Nucleic Acids Res.* 30, 207-210.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418-427.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert, C.J. Jr., and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 3, RESEARCH0046.

Stoeckert, C.J. Jr., Causton, H.C., and Ball, C.A. (2002). Microarray databases: standards and ontologies. *Nat. Genet.* suppl.S32, 469-473.