

Comparison and Analysis of MicroArray Databases

Hui-Huang Hsu

Department of CSIE
Tamkang University,
Taipei, Taiwan, R.O.C.

h_hsu@mail.tku.edu.tw

Andy C. Yang

Department of CSIE
Tamkang University,
Taipei, Taiwan, R.O.C.

andyyung@mail.mine.tku.edu.tw

Ming-Da Lu

Department of CSIE
Tamkang University,
Taipei, Taiwan, R.O.C.

dlj0938@yahoo.com.tw

Abstract

Microarray technology has become widely utilized by biologists as an important functional genomics approach. Combined with the Internet and web technologies, a load of microarray databases have been developed. Each of the available microarray databases so far is built for specific purposes, and it also provides biologists and researchers with various microarray data. Nevertheless, it is difficult for the user to choose which bioinformatics systems meet his needs most, for the large amounts of microarray databases. In order to better understand the available microarray databases, a comparison and a brief analysis are performed in this paper. The paper includes some online microarray databases to date that falls into four categories: (i) database platform and interface, (ii) data transmission standard, (iii) data variation, and (iv) system extensibility. The surveyed microarray databases are considered and analyzed within above four aspects. In the end of this paper, we conclude that the integration between several microarray databases is essential.

Keywords: Microarray, database, analysis

1. Introduction

Microarray technology has been widely used by biologists nowadays. Not only for its high-throughput characteristic but it also provides an opportunity to gain gene-expression data from a large number of genes [1]. With microarray technology, biologists can study the expression of thousands of genes simultaneously, and it also gives an overview of the gene activities in a given situation or a chemical impulse.

Probes on the microarray are the identifying units which are typically cDNA samples. After the hybridization of a labeled nucleic acid extract with a large array of different DNA segments, biologists can progress a series of subsequent analyzing work. By varying the origin of the RNA to the extracted samples, microarray can reveal the gene's relative expression under the specific environments.

Physical parameters or chemical effects may result in different microarray experimental effects, and these include different temporal states, disease states, or certain treatment conditions. Followed by normalizing the array

data, the intensity of the signal for each gene can be compared. Biologists will then find the genes with various intensity for signals under a specific condition relative to another condition. Some statistical clustering methods may aid in analyzing the gene expression data. With microarray technology, biologists can efficiently obtain the gene expression data.

There is a great deal of existing microarray databases recording large amounts of microarray. Each of them provides the user with different functions and various user interfaces. Even the data stored in the systems differ a lot. Some of the available microarray databases are developed for a specific purpose. For example, there are microarray databases designed for data of a certain species, or just for a particular experiment. Besides, some of the available microarray databases also provide all kinds of functional interfaces for the user to utilize. For example, the user can preview his query results with a plain graphical visualization web page or something else. Moreover, some of the microarray databases are compliant with newly-developed data transmission or data definition standards. It is not easy for the user to be acquainted with the usage of these comprehensive and large amounts of microarray databases.

To better understand the available microarray databases, a survey is presented in this paper. The survey consists of comparisons of the microarray databases based on four aspects, and also a brief analysis of these systems. The four consideration categories are database platform and interface, data transmission standard, data variation, and system extensibility. We raise a few issues for

discussion in each category, and give some available microarray databases as examples. Finally, a brief conclusion is drawn in the end part of this paper. The discussion shows that the integration between existing microarray databases remains space for improvement.

2. Issues

In this section, discussions about microarray databases in four categories are presented. These issues are database platform and interface, data transmission standard, data variation, and system extensibility

2.1. Database platform and interface

Database platform and interface can be a crucial issue for microarray databases. With user-friendly interfaces designed for the user on the microarray database, biologists can easily perform their query commands, and can also obtain well-regulated information they want. Some of the available microarray databases provide a web browser interface for the user to upload and query data, while others do not. In a few systems, the user can even only set the parameters of his operation, and just wait for the response from the database uncertainly. This leaves the user with great inconvenience. On the contrary, some of the microarray databases fulfill the user's need to browse data responded with a colorful and clear graphical visualization interface. The user can employ a microarray database fast and efficiently if the system is well-designed with novel and ease-of-use interfaces. In addition, functions given by the

system should also be considered.

2.2. Data transmission standard

Data transmission between the microarray databases and the client users remains important. Data sharing among certain similar microarray databases also plays an important role to speed up the communication of information. For the surveyed microarray databases, some of them come with an standard for the data transmission or definition, while others do not. With a unified standard, data transmission can be formulated that helps conservation and circulation of data. For example, biologists coming from different laboratories or operating different microarray experiments cannot share their experimental results without a specific standard. Even for a microarray biologist, it is not easy to arrange his experimental results and transform them onto the microarray databases due to the lack of a unified standard. Moreover, errors or ambiguity may occur often when a biologist is going to transmit his information via the Internet using the microarray databases without a data transmission standard. It will be improved for the conservation and sharing of microarray data if the biologists can follow a unified standard to operate their microarray experimental results.

2.3. Data variation

Some microarray databases were developed for a specific purpose or a particular species data. Among surveyed microarray databases, recorded data themselves range from general microarray data to data of specific species. Biologists can choose a microarray

database that fits their need most to analyze their data of interest.

2.4. System extensibility

As for computer software and web databases, one critical task is to maintain the extensibility of the developed systems. In terms of surveyed microarray databases, only few of them fit this need. These microarray databases have open-source codes, which allow the user to modify a portion of the original system to a new custom one. The user can modify the master copy of the microarray database or can design an additional program to access the system depending on the allowance of the system author. Biologists can thus maximize the benefits of the systems.

3. Comparison and analysis of surveyed databases

In this section, surveying results for microarray databases are discussed. Discussions are raised based on mentioned categories.

3.1. Database platform and interface

3.1.1. Stanford microarray database (SMD)

The SMD [2] stores raw and normalized data experimented from microarray researchers. It also provides a novel web interface for the user to retrieve and visualize his data. This system allows the user to view an enlargement of a selected microarray pixel. A close-up of an individual spot is then shown with spot parameters and associated biological annotation. The user can easily choose the spot in which he

is interested, and obtain corresponding information.

3.1.2. ArrayExpress database

ArrayExpress [3][4] is a public database for high throughput functional genomics data. This microarray database provides sufficient amounts of data within the repository behind the system. The user can operate his query commands via given very simple web interfaces. Queries are divided into two sets: by experiments or by expression profiles. A detailed view of a corresponding search result from the repository displays on the interface. Submissions of large amounts of data are allowed within a specific page. ArrayExpress provides several various functions for the user within different web interfaces.

3.1.3. MUSC DNA microarray database

The Medical University of South Carolina (MUSC) DNA microarray database [5] is a web-accessible archive of DNA microarray data. Public searching and retrieval of data stored in the database can be achieved via the web interface of the system. The system allows the user to search different fields of data. MUSC stands as an academic repository for other microarray databases, and analyzing functions can be added to the system as plug-ins.

3.1.4. MaxdLoad2 and MaxdBrowse

The MaxdLoad2 [6] system is a relational database schema for microarray experimental annotation and storage. MaxdBrowse is a browser implementation of MaxdLoad2. This system presents easy-to-use interfaces for the

user to submit microarray data and to browse, search, and edit the data stored easily. MaxdLoad2 provides several convenient interfaces so that the user can search, edit, or view data very easily. This facilitates the usage of the system, also gives the user more opportunity to utilize MaxdLoad2 completely.

3.2. Data transmission standard

3.2.1. The MIAME standard

The Minimum Information About a Microarray Experiment (MIAME) standard was developed by the Microarray Gene Expression Data society (MGED). This standard aims to specify the microarray experiment data and metadata that should be recorded and described to enable other biologists to understand the microarray experiment clearly [7][8]. MIAME is a data content standard but not a format standard. Almost all surveyed microarray databases are compliant with MIAME standard, except for certain particular purpose databases [9].

3.2.2. MAGE-ML standard

The Microarray Gene Expression (MAGE) standard contains two major units: MAGE Object Model (MAGE-OM) and MAGE Markup Language (MAGE-ML) [10][11]. These two standards were designed to achieve software interoperability, also to encode MIAME required information. Essential parts of the microarray experiments are fetched to fit the MIAME standard, and then MAGE-ML can be used to communicate within MAGE object models. A few surveyed microarray databases are compliant with MAGE-ML standard. For

example, ArrayExpress, MaxdLoad2, MAMA [12], dbZach [13], SMD [14], LAD [15], and GEO [16]. Only microarray databases which are compliant with MAGE-ML can be software interoperable, and the user can utilize these systems more conveniently.

3.3. Data variation

3.3.1. BarleyExpress

BarleyExpress is a web-based microarray experiment data submission tool for BarleyBase [17]. This system mainly focuses on the data of Barley 1 and other plant GeneChips.

3.3.2. dbZach database

dbZach is a modular relational database with associated data insertion, retrieval, and mining tools, that manages toxicology and complementary toxicogenomic data. This system presents enterprise data management interfaces for the user to deal with toxicogenomic data.

3.3.3. READ database

The RIKEN Expression Array Database (READ) is a database of expression profile data from the RIKEN mouse cDNA microarray. This system gives an opportunity for the user to work on the research of RIKEN mouse data.

3.3.4. SGMD database

The Soybean Genomics and Microarray Database (SGMD) provides an integrated view of interactions of soybean with the soybean cyst-nematode (SCN) [18]. The database collects large amounts of relative EST and

microarray data. Biologists can study the influence of SCN on soybean with the information provided by this system.

3.4. System extensibility

3.4.1. MAMA system

The Meta-Analysis of MicroArray (MAMA) data system integrates several analytical methods, including an open- source framework offering other database designers flexibility to link additional statistical algorithms as their plug- in sub-routines.

3.4.2. SMD database

The Stanford Microarray Database (SMD) is a research tool and archive that allows the user to store and share microarray data [19][20]. Significantly, SMD makes its source code fully and freely available for other programmers to create a local installation of SMD. It also shifts data to Linux platforms so that the data stored can be accessed more freely.

3.4.3. LAD database

The Longhorn Array Database (LAD) is another open- source and MIAME-compliant system of SMD. This database is a fully open-source version of the SMD. LAD can operate on PostgreSQL and Linux platforms.

3.5. Comparison

Table 1 illustrates a brief comparison among these databases. The result shows that these microarray databases differ in functions and major purposes. All of them come with

individual interfaces, and most of them are MIAME-compliant. A few of the surveyed databases are designed for specific data of species, while others are general-purposed. To sum up, existing microarray databases differ a lot in functions provided or purpose of usage. The integration between microarray databases is therefore very important.

4. Conclusions

We surveyed twenty microarray databases and systems, and a discussion on these microarray databases based on four specific categories was presented in this paper. This paper was raised to give a general overview for existing microarray databases. We aim to help biologists choose which microarray database fit their need most. The user can thus better understand available microarray databases. Moreover, the discussion and analysis of surveyed databases show that the integration between available microarray databases is essential. This provides an important research direction.

5. References

- [1] M. Gardiner-Garden, and T. G. Littlejohn, "A comparison of microarray databases," *Briefings In Bioinformatics*, Vol. 2, no. 2, 143-158, May 2001.
- [2] C. A. Ball, "The Stanford Microarray Database accommodates additional microarray platforms and data formats," *Nucleic Acids Research*, Vol. 33, Database issue, D580-D582, 2005.
- [3] U. Sarkans et al., "The ArrayExpress gene expression database: a software engineering and implementation perspective," *Bioinformatics*, Vol. 21, no. 8, 1495-1501, 2005.
- [4] H. Parkinson et al., "ArrayExpress—a public database of microarray experiments and gene expression profiles," *Nucleic Acids Research*, Vol. 35, Database issue, 747-750, 2007.
- [5] G. L. Argraves, J. L. Barth, and W. S. Argraves, "The MUSC DNA Microarray Database," *Bioinformatics*, Vol. 19, no. 18, 2473-2474, 2003.
- [6] D. Hancock et al., "maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination," *BMC Bioinformatics*, Vol. 6, 264, 2005.
- [7] G. Mukherjee et al., "Plant-Based Microarray Data at the European Bioinformatics Institute. Introducing AtMIAMExpress, a Submission Tool for Arabidopsis Gene Expression Data to ArrayExpress," *Plant Physiology*, Vol. 139, 632-636, October 2005.
- [8] H. Parkinson et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, Vol. 33, Database issue, 553-555, 2005.
- [9] H. Bono, T. Kasukawa, Y. Hayashizaki, and Y. Okazaki, "READ: RIKEN Expression Array Database," *Nucleic Acids Research*, Vol. 30, no. 1, 211-213, 2002.
- [10] P. T. Spellman et al., "Design and

- implementation of microarray gene expression markup language (MAGE-ML),” *Genome Biology*, Vol. 3, no. 9, 46, 2002.
- [11] W. Martin, and R. M. Horton, “MageBuilder: A Schema Translation Tool for Generating MAGE-ML from Tabular Microarray Data,” *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, CA, USA, August 11-14, 2003.
- [12] Z. Zhang, and D. Fenstermacher, “An Introduction to MAMA (Meta-Analysis of MicroArray data) System,” *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, September 1-4, 2005.
- [13] L. D. Burgoon, Paul C. Boutros, Edward Dere, and Timothy R. Zacharewski, “dbZach: A MIAME-Compliant Toxicogenomic Supportive Relational Database,” *Toxicological Sciences*, Vol. 90, no. 2, 558-568, April 2006.
- [14] G. Sherlock et al., “The Stanford Microarray Database,” *Nucleic Acids Research*, Vol. 29, no. 1, 152-155, 2001.
- [15] P. J. Killion, Gavin Sherlock, and Vishwanath R Iyer, “The Longhorn Array Database (LAD): An Open-Source, MIAME compliant implementation of the Stanford Microarray Database (SMD),” *BMC Bioinformatics*, Vol. 4, 32, 2003.
- [16] T. Barrett et al., “NCBI GEO: mining millions of expression profiles—database and tools,” *Nucleic Acids Research*, Vol. 33, Database issue, D562-D566, 2005.
- [17] X. Tang, L. Shen, and J. A. Dickerson, “BarleyExpress: A Web-Based Submission Tool for Enriched Microarray Database Annotations,” *Bioinformatics*, Vol. 21, no. 3, 399-401, 2005.
- [18] N. W. Alkharouf, and B. F. Matthews, “SGMD: the Soybean Genomics and Microarray Database,” *Nucleic Acids Research*, Vol. 32, Database issue, D398-D400, 2004.
- [19] J. Demeter, “The Stanford Microarray Database: implementation of new analysis tools and open source release of software,” *Nucleic Acids Research*, Vol. 35, Database issue, D766-D770, 2007.
- [20] J. Gollub et al., “The Stanford Microarray Database: data access and quality assessment tools,” *Nucleic Acids Research*, Vol. 31, no. 1, 94-96, 2003.

Table 1: Comparison of surveyed microarray databases

Database	Web Interface	MIAME Support	MAGE-ML Support	Data Species	Open-Source	Note
ArrayExpress	Yes	Yes	Yes	General	No	
MUSC	Yes	Future work	Future work	General	No	
maxdLoad2	No	Yes	Yes	General	No	Combined to provide three functional interfaces
maxdBrowse	Yes	Yes	Yes	General	No	
MIMAS	Yes	Yes	No	General	No	
MAMA	No	Yes	Yes	General	Yes	System modification is allowed
BarleyExpress	Yes	Yes	No	Barley & other plants	No	
dbZach	Yes	Yes	Yes	Toxicogenomic	No	New functions available (http://dbzach.fst.msu.edu/dbZACH_info/dbZACH_FAQ.html#interfacing)
READ	Yes	No	No	RIKEN mouse	No	System suspended
SGMD	Yes	Yes	Future work	Soybean	No	For Queries only, not for data uploading
SMD	Yes	Yes	Yes	General	Yes	System modification is allowed
LAD	Yes	Yes	Yes	General	Yes	Transforms SMD data to a convenient platform
GEO	Yes	Yes	Yes	General	No	Multiple data uploading