YMD: A microarray database for large-scale gene expression analysis

Kei-Hoi Cheung, PhD[1], Kevin White, PhD[2], Janet Hager, PhD[3,4], Mark Gerstein, PhD[4]
Valerie Reinke, PhD[2], Kenneth Nelson, PhD[5], Peter Masiar, MS[1], Ranjana Srivastava, PhD[1]
Yuli Li, MS[1], Ju Li, MS[1], Hongyu Zhao, PhD[2,6], Jinming Li, PhD[7], David B. Allison, PhD[8]
Michael Snyder, PhD [4,5], Perry Miller, MD, PhD[1,5], Kenneth Williams, PhD[3,4]

[1]Center for Medical Informatics, Department of Anesthesiology, [2]Department of Genetics
[3] Keck Biotechnology Resource Laboratory, [4]Department of Molecular Biophysics and Biochemistry
[5]Department of Molecular, Cellular and Developmental Biology
[6]Department of Epidemiology and Public Health
Yale University, New Haven, CT
[7]School of Biological Sciences, Nanyang Technological University, Singapore
[8]Section on Statistical Genetics, Department of Biostatistics, University of Alabama, Birmingham

The use of microarray technology to perform parallel analysis of the expression pattern of a large number of genes in a single experiment has created a new frontier of medical research. The vast amount of gene expression data generated from multiple microarray experiments requires a robust database system that allows efficient data storage, retrieval, secure access, data dissemination, and integrated data analyses. To address the growing needs of microarray researchers at Yale and their collaborators, we have built the Yale Microarray Database (YMD). YMD is Web-accessible with the following features: (i) a Web program that tracks DNA samples between source plates and arrays, (ii) the capability of finding common genes/clones across different array platforms, (iii) an image file server, (iv) laboratory-based user management and access privileges, (v) project management, (vi) template data entry, (vii) linking gene expression data to annotation databases for functional analysis. YMD is currently being used on a pilot basis by several laboratories for different organisms and array platforms.

INTRODUCTION

Microarrays represent a high-throughput biotechnology that allows parallel evaluation of the expression pattern of tens of thousands of genes in a single experiment. The differential expression of these genes under different experimental conditions (e.g., different drug treatments) can yield important information about how different genes work together to mediate complex biological processes that constitute the genetic basis of human diseases or basic physiology. For example, a recent breakthrough in medical research involving the use of microarrays has shown that gene expression data can be used to classify different types of embryonal tumors of the central nervous system and to predict the clinical outcome of these tumors [1]. To understand the biological significance of such large amounts of expression data requires a significant bioinformatics effort. A key to such an effort is a robust database system that allows efficient data storage, retrieval, secure data access, data dissemination, and integrative analysis. To this end, we have built the Yale Microarray Database (YMD) (http://info.med.yale.edu/microarray) to meet the growing informatics needs of microarray research at Yale. YMD is intended for use by microarray researchers at Yale and researchers who use the Microarray Resource of the Keck Biotechnology Resource Laboratory at Yale. While the development of YMD is ongoing, it is currently being used on a pilot basis by several laboratories involving different organisms. These include Dr. White's laboratory (Drosophila), Dr. Reinke's laboratory (C. elegans), Keck Microarray Resource (Human, Mouse, and Arabidopsis), and Dr. Snyder's Microarray Facility (Yeast and Human). While our current focus is on spotted array technology, YMD is designed flexibly so that it can be extended to handle other array technologies such as the Affymetrix GeneChip technology.

Similar large-scale microarray database efforts are underway at other universities and research institutes. These efforts include SMD [2] at Stanford, ChipDB at MIT (http://young39.wi.mit.edu/chipdb_public/), ArrayExpress at the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/arrayexpress/), ArrayDB [3] at NHGRI, ExpressDB [4] at Harvard, etc. These microarray databases have been reviewed and compared elsewhere [5]. While some of these databases (e.g., ArrayExpress) are compliant with MIAME (Minimum Information About Microarray Experiments) [6], others are not. YMD is currently considered to be "MIAME-supportive" since its initial design was based on a subset of the MIAME standard recommendations. Our goal is to make YMD MIAME-compliant to facilitate data exchange and publication. There are several tools or features that distinguish YMD from other microarray databases. These features as well as others are described in the next section.

YMD FEATURES

Plate-to-array convolution

Laboratories or facilities that print microarray slides need the capability of automatically tracking samples between plates and array slides. Since a wide variety of arrayers may be used to perform array spotting, the tracking software must be flexible so that it can accommodate different arrayer configurations. Also, it should be independent of computer platforms as users in different laboratories/facilities may use computers running different operating systems (e.g., MAC, Windows, and Unix). To this end, we have developed a Web-launchable program that allows the

user to enter a set of parameters describing (i) the input plates (well locations and identifiers of the samples); (ii) configurations of the arrayer including the number of pins in the print head, the dipping pattern (how the pins dip into the plate wells), and the printing pattern (how the samples are spotted on the surface of the array slide); and (iii) the format of the array list output that can subsequently be read by a specific analysis program (e.g., Axon's GenePixPro and Packard Biochip's QuantArray).

Comparison across multiple array platforms

In planning and designing a microarray experiment, it is important for researchers to ensure that the genes or ESTs of interest are featured on the slide(s) they plan to acquire or purchase. There are a variety of array platforms that are made available by different groups (e.g., Affymetrix GeneChip, Operon, and Yale Keck Microarray Resource), featuring different organisms and the use of different sources of DNA such as cDNA-based PCR products and oligonucleotides. To further complicate the analysis, different DNA fragments identified by different genbank accession numbers may correspond to the same gene. One way to address this problem is to use a common identifier such as the unigene cluster ID to identify these analogous sequences. To facilitate comparison across multiple platforms, we have implemented a Web interface that allows the user to any combination of arrays available from the Yale Keck Microarray Resource and identify among these arrays common DNA fragments based on unigene cluster IDs. In addition, the interface provides the option to attach to these common genes functional annotations that have been obtained from Stanford SOURCE database (genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch).

Laboratory-based user management

In YMD, user groups are represented by the participating laboratories (within Yale and outside of Yale). Each laboratory will have a designated Principal Investigator (PI). There are five types of user roles in the following order: Database Administration (DBA), Laboratory Administration (Lab Admin), Level-1, Level-2, and Guest. A person who has the DBA role oversees the database. The DBA can create a PI user as well as the corresponding laboratory. Once a PI user is created, he is automatically granted the Lab Admin role. This role permits the PI to (a) create new users and assign their individual project roles for his laboratory and (b) have full access to all the projects within his laboratory. A PI can assign the Lab Admin role to one or more of his laboratory members. This way the PI can delegate the laboratory user and project management responsibility. To allow sharing of data between laboratories, a PI or Lab Admin can create a new laboratory member by including an existing member from another laboratory (users may be assigned with different roles in different laboratories). Users who have the Level 1 role cannot create users or assign their project roles, but they can have full access to all the projects within the laboratory. Unlike Level-1 users, Level-2 users are permitted to make changes only to those projects for which they have been granted this privilege. Guest users have read-only access to those projects to which they have been given access. In addition to these user roles, we plan to allow the PI or Lab Admin user the ability to make certain projects (or certain aspects of the projects) within the laboratory accessible on a read only basis to the scientific community.

Project management interface

We are grouping microarray data into a four-level hierarchical structure: projects, subprojects, experiments, and hybridizations. In our database schema, a generic parent-child structure is used to represent this hierarchy. This approach allows us to easily add another level to the structure or to rearrange the structure without changing the database schema. Fig. 1 shows the graphical interface that reflects such a hierarchical data structuring with each level coded in a different color. The folder icon on the left of each project, subproject or experiment can be clicked to expand or collapse the corresponding "children" dynamically. While the project and

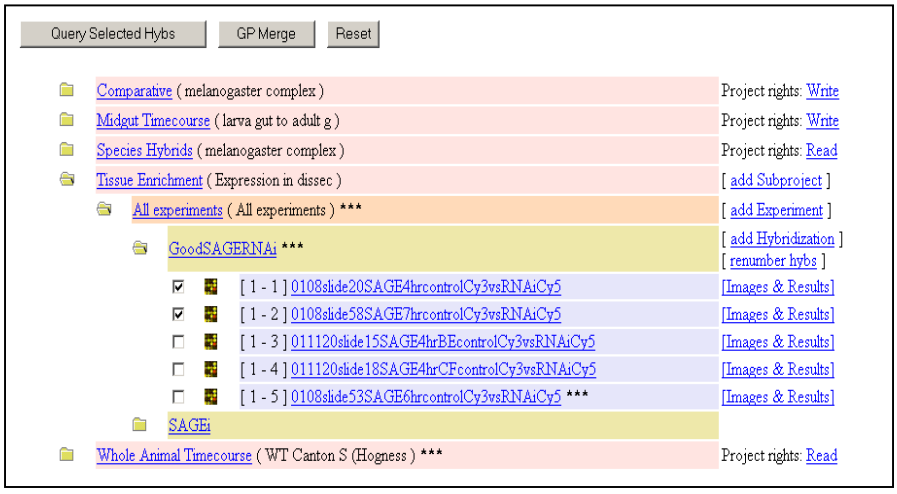

Fig. 1. Project management interface.

subproject folders can be used to separate different categories of microarray experiments, an experiment folder consists of a set of hybridizations (which may correspond to expression data over a series of time points for a time-course experiment). At the hybridization level, we allow grouping of hybridizations that represent biological repeats or technical repeats within the same biological repeat.

Data querying and different types of data analysis can be performed on the quantitative data (Genepix output) associated with the hybridizations. As indicated in Fig. 1, the checkbox on the left of each hybridization (indicated by the image-spots icon) is used to select the hybridizations of interest. For example, a set of replicate hybridizations can be selected and then imported to the GPMerge program developed by Dr. Zhao's laboratory (http://zhao.med.yale.edu/) by pressing the button labeled "GPMerge". The output files containing the analysis results can be downloaded through the Web onto the user's local computer for further analysis.

Template data entry

Very often microarray experiments involve entering the same pieces of information repeatedly. For example, a set of experiments may require the use of the same biological specimen that is treated slightly differently for each experiment. To avoid entering the data redundantly, our system lets the user choose a previous entry (e.g., the entry of a previous experiment within the same subproject) as a template for creating the new entry. In other words, the values of the old entry are copied to the new entry and the user can edit them in the new entry. This template data entry is applied at the experiment level and the hybridization level. By default, the system uses the last entry (either an experiment within a subproject or a hybridization within an experiment) as a template for creating a new one. However, the user can arbitrarily select an existing experiment or hybridization to be a new data entry template.

Other features

Image file server. YMD provides storage for raw image data (TIF files). There has been debate in the microarray community whether or not raw microarray images should be stored. This question was also brought up at previous Microarray Gene Expression Database (MGED) group meetings. It is a tradeoff between storage space and the ability to re-analyze the raw images (especially when new and improved image analysis algorithms are developed). We choose to allow the user to store the raw images mainly because the hardware cost for storage has become less expensive and because of the potential benefits fo being able to reach back to the underlying raw data. We have set up a central image file server that has a current capacity of 500GB and can be expanded. These image files are linked to the microarray experiment data stored in Oracle. The user can readily download these images to their local computers for analysis or re-analysis. Another advantage of storing the images is that the individual image spots can be linked to the gene expression data so that they can be viewed by the user and compared with the corresponding expression values. The morphology of the spots or the visual appearance of the spots may quickly explain outlier and otherwise inexplicable expression datapoints.

Query interface. YMD provides a nested querying capability through a Web query criteria interface (Fig. 2) to allow the user to perform preliminary analysis of

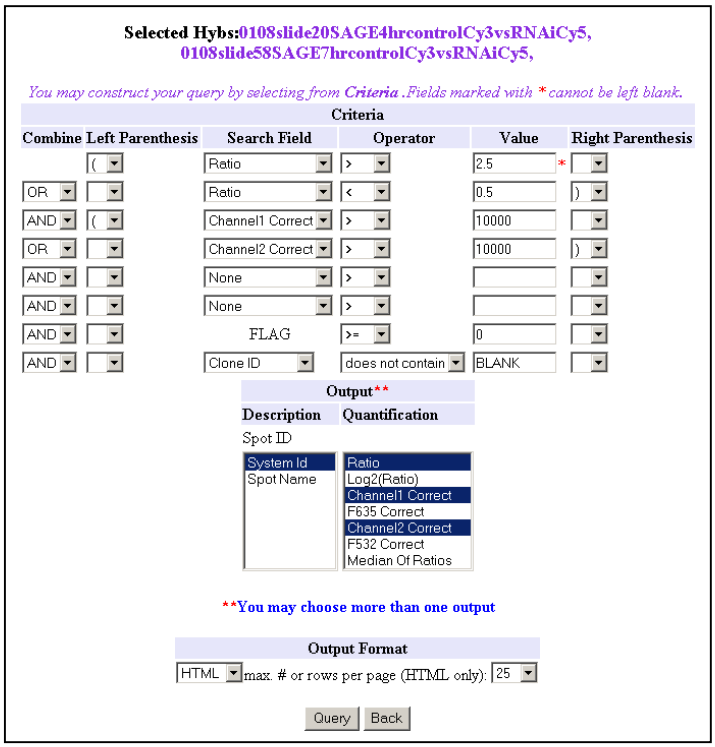

Fig. 2. Query interface.

the expression data for a selected set of hybridizations that belong to the same experiment or different experiments. Fig. 2 shows how a set of complex query criteria can be entered through this form interface to query data associated with two selected hybridizations (their selection was shown in Fig. 1). The query shown in Fig. 2 retrieves genes that are up-regulated by a factor of 2.5 (expression ratio > 2.5) or down-regulated by a factor of 0.5 (expression ratio < 0.5) and at the same time, all of the following conditions must be met: (i) the intensity of one or both of the channels (red and green channels) must be greater than 10,000; (ii) the flag indicating the data quality must be greater than or equal to zero; and (iii) the clone ID does not contain the word BLANK. Notice in this example that parentheses can be used to indicate the precedence of the Boolean operations. The query interface also allows the user to choose the format of the query output. For example, it allows the user to choose which columns (e.g., ratio, channel 1 intensity, channel 2 intensity, etc.) to be included in the query output. Also, the query output can be formatted into one of the following: HTML, EXCEL, TEXT (tab-delimited), and CLUSTER (which can be imported to a standard cluster analysis program). Finally, our system allows the query output to be dynamically linked to external annotation databases such as DRAGON [7] and SOURCE based on genbank accession numbers.

Data unpacking. The image quantitative data produced by the scanning software (e.g., Axon's Genepix) are captured as files initially. For efficient data querying across multiple arrays associated with different experiments, these files are unpacked into an Oracle table and column indexes are created. In YMD, such quantitative data are associated with hybridizations. Once the user specifies the link between the quantitative data files and the hybridizations, the data unpacking process will be done in the background so that the user does not need to wait for the process to finish. This allows the user to unpack a large batch of files without spending a long time. All the user needs to do is to link the files with the corresponding hybridizations through the Web interface. Currently, YMD only allows quantitative data produced by Genepix to be unpacked into the Oracle database. We will expand this to other types of scanning software such as QuantArray by Packard BioScience.

DISCUSSION

We have adopted a distributed client/server architecture for implementing YMD. Currently, YMD is distributed across three different servers, namely, the image file server, the Web server, and the Oracle database server. These servers are linked over the network. There is a general debate in the database community that the federation and the centralization approaches have different merits (this has been discussed in some detail elsewhere [8]). We have chosen a more federated approach toward handling expression information. This has the advantage in that it more closely reflects the social structure of biology laboratories that are doing the experiments and computational analyses – i.e. there are many laboratories, supported by different sources and pursuing different questions, that are naturally linked into a loose federation by their interest in expression experiments. One problem with the federated structure is that it is not very efficient for dealing with large amounts of bulk data, such as that produced by microarrays. We attempt to address this by suggesting that what will be shared by the interoperating databases and analysis has to be more than an interface or common file formats. Rather, we propose that they share effectively a "sub-schema", a number of commonly structured tables that can transfer bulk data much more efficiently than text files or "single datum at a time" network interfaces. A similar federated approach has been applied to microarray data. As described in [9], the RNA Abundance Database (RAD) is subdatabase of the Genomics Unified Schema (GUS) [10] that is a larger framework for gene annotation.

Currently, the YMD Oracle server is used to represent and store basic descriptions of experiments and their associated hybridization raw data (generated by various scanning software such as GenePix). To provide more powerful data analysis and support MIAME data standard more fully, we need to expand YMD to cover the following areas: (i) a more detailed description of experiments, (ii) processed data (storing and retrieving the results of more advanced analyses such as clustering), (iii) gene annotations. We propose to build a subdatabase (it does not need to be in Oracle) to address each area and establish links between the subdatabases and YMD. For example, YMD can potentially be linked with the annotation database that we are currently implementing to store gene annotations for the purpose of cross-array-platform comparison described previously. In addition, we have recently begun a pilot project to explore the use of the TRIAL/DB approach (based on the extension of the entity-attribute-value data modeling) [11] to build a database for describing microarray experiments based on the MIAME standard.

As described previously, we use Unigene IDs as the common identifiers for analogous DNA sequences spotted on different glass slide arrays. One limitation with this approach is that Unigene IDs change over time as the knowledge of how to cluster the sequences improves. Another limitation is that this approach does not allow cross-species comparison because the unigene IDs are different for different species. To address these issues, RESOURCERER [12] developed at TIGR uses the the TIGR Gene Indices (TGI) [13] and TIGR Orthologous Gene Alignment database (TOGA) (http://www.tigr.org/tdb/TOGA/TOGA.shtml) to compare sequences across multiple array platforms with organisms. The Unigene database is more comprehensive in its sequence coverage than TGI that clusters EST data only. One solution is to find a way to integrate all these annotation databases in a meaningful fashion.

To extend the querying capability, we are implementing methods that provide the user with the ability to save queries. The system will provide two query saving methods. One method is to allow the user to save the parameter values that were entered in the query form shown in Fig. 2. This set of parameter values can be saved and retrieved by a user-defined name. Once retrieved, the query form will be filled with those saved values automatically. The user can then edit the values if needed. The other method is to save the SQL statement corresponding to the parameters entered in the query form. These saved SQL statements can later be re-executed individually or in series (these queries can be combined with Boolean AND/OR). In addition to these two methods, we believe that it would be convenient to allow the user to save a hybridization selection for later use.

YMD has been designed to be an institution-wide gene expression database accessed by multiple laboratories/facilities conducting a wide variety of microarray experiments involving different organisms. There is a need to tailor the user interface to specific needs of the individual laboratories. For example, there are organism-specific controlled vocabularies or standard nomenclature (e.g., Flybase for Drosophila and CBIL's controlled vocabularies for human and mouse) available for use in describing microarray experiments. These lists of terms extracted from different sources can be displayed as choice lists to the users when they enter information describing the experiments. However, it would be a burden to the user who only works with a single organism to see all the terms for all organisms. The user interface should be configured dynamically to display to the user only those relevant terms. By the same token, different laboratories may have different needs for data analysis and gene annotation. The user interface should also be tailored to such different needs. We are currently exploring the use of metadata to drive the user interface dynamically. For example, metadata can be designed to store the preferences for individual laboratories (e.g., what particular organism(s), analysis programs and/or annotation).

Acknowledgements

This work was supported in part by NIH grants U24 DK58776 from the National Institute of Diabetes and Digestive and Kidney Diseases, K25 HG02378 from the National Human Genome Research Institute, G08 LM05583 and T15 LM07056 from the National Library of Medicine, R01 CA077808 from the National Cancer Institute and NSF grant 0090286.

References

1. Pomeroy PL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Oson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 2002. 415: 436-442.
2. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. Nucleic Acids Res., 2001. 29(1): 152-5.
3. Ermolaeva, O., M. Rastogi, K.P. KD, S. Schuler, M. Bittner, Y. Chen, R. Simon, P. Meltzer, J. Trent, and M. Boguski, Data management and analysis for gene expression arrays. Nat. Genet., 1998. 20(1): 19-23.
4. Aach J, Rindone W, Church GM. Systematic management and analysis of yeast gene expression data. Genome Res., 2000. 10(4): 431-45.
5. Gardiner-Garden M, Littlejohn TG. A comparison of microarray databases. Brief Bioinform., 2001. 2(2): 143-58.
6. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet., 2001. 29(4): 365-71.
7. Bouton CM, Pevsner J. DRAGON: Database Referencing of Array Genes Online. Bioinform, 2000. 16(11): 1038-1039.
8. Gerstein M. Integrative database analysis in structural genomics. Nat. Struct. Biol., 2000. 7 Suppl: 960-963.
9. Stoeckert, C., A. Pizarro, E. Manduchi, M. Bibson, B. Brunk, J. Crabtree, J. Schug, S. Shen-Orr, and G. Overton, A relational schema for both array-based and SAGE gene expression experiments. Bioinform., 2001. 17(4): 300-308.
10. Davidson, S., J. Crabtree, B. Brunk, J. Schug, V. Tannen, G. Overton, and C. Stoeckert, K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. IBM Systems Journal, 2001. 40(2): 512-531.
11. Brandt CA, Nadkarni P, Marenco L, Karras BT, Lu C, Schacter L, Fisk JM, Miller PL. Reengineering a database for clinical trials management: lessons for system architects. Control Clin. Trials, 2000. 21(5): 440-61.
12. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J. RESOURCERER: a database for annotating and linking microarray resources within and across species. Genome Biology, 2001. 2(11): 1-4.
13. Quackenbush J, Cho J, Lee Y, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana J, White J. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acid Res., 2001. 29: 159-164.