Benchmark

TM4: A Free, Open-Source System for Microarray Data Management and Analysis

BioTechniques 34:374-378 (February 2003)

A.I. Saeed[1], V. Sharov[1], J. White[1], J. Li[1], W. Liang[1], N. Bhagabati[1], J. Braisted[1], M. Klapa[1], T. Currier[1], M. Thiagarajan[1], A. Sturn[1], M. Snuffin[2], A. Rezantsev[2], D. Popov[2], A. Ryltsov[2], E. Kostukovich[2], I. Borisovsky[2], Z. Liu[3], A. Vinsavich[3], V. Trush[3], and J. Quackenbush[1,4]

[1]The Institute for Genomic Research, Rockville, MD, [2]DataNaut, Bethesda, MD, [3]Syntek Systems, Bethesda, MD, and [4]Department of Biochemistry, George Washington University, Washington, D.C., USA

Microarrays have emerged as the premier tool for studying gene expression on a genomic scale. Advances in the precision of array printers and scanners as well as improved laboratory protocols (11) allow for assays of tremendous complexity and scope. Scientists seeking to harness the potential of this technique are often challenged by the large quantities of data produced. Well-designed, user-friendly software is the key to tracking, integrating, qualifying, and ultimately deriving scientific insight from the experimental results. In support of our ongoing work in microarray analysis of gene expression, we developed a suite of software that allow users in the laboratory to capture, manage, and analyze effectively data from DNA microarray experiments. The TM4 suite of tools consist of four major applications, Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database, all of which are freely available to the scientific research community at http://www.tigr.org/software. Although these software tools were developed for spotted two-color arrays, many of the components can be easily adapted to work with single-color formats such as filter arrays and GeneChips™ (Affymetrix, Santa Clara, CA, USA). Three of the TM4 applications, MADAM, MIDAS, and MeV, were developed in Java and have been tested on Microsoft® Windows™, Linux®, Unix®, and MacOS X® platforms; TIGR Spotfinder was written in C/C++ and runs only on Windows systems. The TM4 software system represents a comprehensive, extensible, open-source, and freely available collection of tools that we believe will be of use to a wide range of laboratories conducting microarray experiments. We further hope that by providing source code along with the executable software, we can encourage others to develop new analysis methods and utilities that will further enhance the capabilities of this software system.

Managing array data effectively requires the development and maintenance of a database designed to reflect the experimental process. Central to TM4 is a MySQL database that stores all data related to the microarray process, such as information on experiments, studies, protocols, data normalization, and gene expression. This database supports MIAME (4) and captures the information deemed essential by this standard. While this may serve as a complete microarray database system for some users, others can use it as a starting point for customized database development. A schema for the database is provided, and TM4 can be adapted to work with existing relational databases as the interaction is via JDBC (Sun Microsystem's Java DataBase Connectivity Application Program Interface).

A database is only as good as the data that are entered into it. MADAM (Figure 1A), implemented in Java, facilitates data entry into the database. MADAM guides users through the microarray process from RNA procurement to data analysis, offering intelligent forms to simplify the tracking of experimental parameters and results that are essential for the interpretation of expression results in downstream analyses. Canned reports provide information on RNA samples, studies, slide maps, and other pertinent data, and a general SQL query window allows freeform access to the underlying database. MADAM also serves as a platform for launching other data entry and management tools. Through the use of these integrated modules, users can view and score PCR plates, design experiments and studies, and track laboratory materials. Although not yet fully

supported, MADAM is being adapted to read and write MAGE-ML, the XML data exchange format being developed by an international consortium of leading public databases and microarray research centers. A MAGE-ML version of MADAM should be available by the end of this year and will facilitate submission of microarray data to public repositories such as Array Express and GEO.

Image analysis is a crucial step in the microarray process. TIGR Spotfinder (Figure 1B) was designed for the rapid, reproducible, and computer-aided analysis of microarray images and the quantification of gene expression. TIGR Spotfinder reads paired 16-bit TIFF image files generated by most microarray scanners. Semi-automatic grid construction defines the areas of the slide where spots are expected. Automatic and manual grid adjustments help to ensure that each rectangular grid cell is centered on a spot. A histogram segmentation method defines the boundaries between each spot and the surrounding local background. Spot intensities are calculated as an integral of non-saturated pixels, although other options including spot medians are available. Local background is subtracted from each intensity value. These calculated intensities, along with each spot's position on the array, spot area, background values, and quality-control flags, are written to a TIGR ArrayViewer ("*.tav*") file format, a Microsoft Excel® workbook, or the database. Reusable grid geometry files and automatic grid adjustment allow the user to analyze large quantities of images in a consistent and efficient manner. To complement the automated methods, particularly in noisy areas of the slide, the user may manually identify or discard spots. Quality-control views allow the user to assess systematic biases in the data. TIGR Spotfinder is written in C++.

Before the intensity values measured in TIGR Spotfinder can be compared, normalization is necessary. This critical step can help compensate for variability between slides and fluorescent dyes, as well as other systematic sources of error, by appropriately adjusting the measured array intensities. Data filtering can reduce the dataset by removing poor or questionable data, in addition to data deemed uninteresting or irrelevant to the analysis. TIGR's MIDAS (Figure 1C), a Java application, provides users an intuitive interface to design analysis protocols combining one or more normalization and filtering steps. In this way, data from many individual hybridizations can be treated in a uniform and reproducible manner. MIDAS reads "*.tav*" files generated by TIGR Spotfinder or retrieved from the database via MADAM. Normalization modules include locally weighted linear regression [lowess (7,19)] and total intensity normalization. These can be linked with filters, including low-intensity cutoff, intensity-dependent Z-score cutoffs, and replicate consistency trimming, creating a highly customizable method for preparing expression data for subsequent comparison and analysis. Data analysis methods are constructed using an intuitive graphical scripting language and can be saved for application to other datasets. MIDAS provides scatterplots that illustrate the effects of each algorithm on the data. When the normalization and filtering steps are complete, MIDAS outputs the data in "*.tav*" format.

Normalized and filtered expression files are ready for analysis using TIGR MeV (Figure 1D). MeV is capable of loading "*.tav*" files, including those normalized by MIDAS, and generates informative and interrelated displays of expression and annotation data from single or multiple experiments. At this final stage of the TM4 pipeline, flexibility and the variety of analysis techniques are critical, as every algorithm has strengths that can be exploited when used on certain datasets and experimental designs. The concept of modularization lends itself particularly well to this system, as novel algorithms and existing codebases from the microarray community can be integrated with the Java-based MeV using a well-defined module application program interface (API).

Analysis modules currently implemented in MeV include hierarchical clustering (8), k-means clustering (18), self-organizing maps (15), principal components analysis (17), cluster affinity search technique (3), self-organizing trees (13), template matching, be-

tween-groups tests (including t-tests), QT_Clust (14), support vector machines (5), gene shaving (10), and relevance networks (6). Bootstrapping and jackknifing resample the dataset to generate consensus clusters. Figure of merit graphs (20) suggest appropriate input parameters for algorithms such as k-means. Modules to allow metabolic pathways and genomic/chromosomal maps to be viewed with expression data overlaid are in development and testing. A wizard to handle links to public database Web sites is also being developed. Clusters identified through any analysis method can be labeled and tracked through other analyses, providing the user with the ability to compare the results of several clustering algorithms to determine consensus and focus on genes with specified expression patterns and biological profiles.

The TM4 system was developed in our laboratory at TIGR and has been used for the collection and analysis of a gene expression data from a number of microarray experiments in a wide variety of systems, including studies of expression in human cancer, rodent models of human disease, expression in the

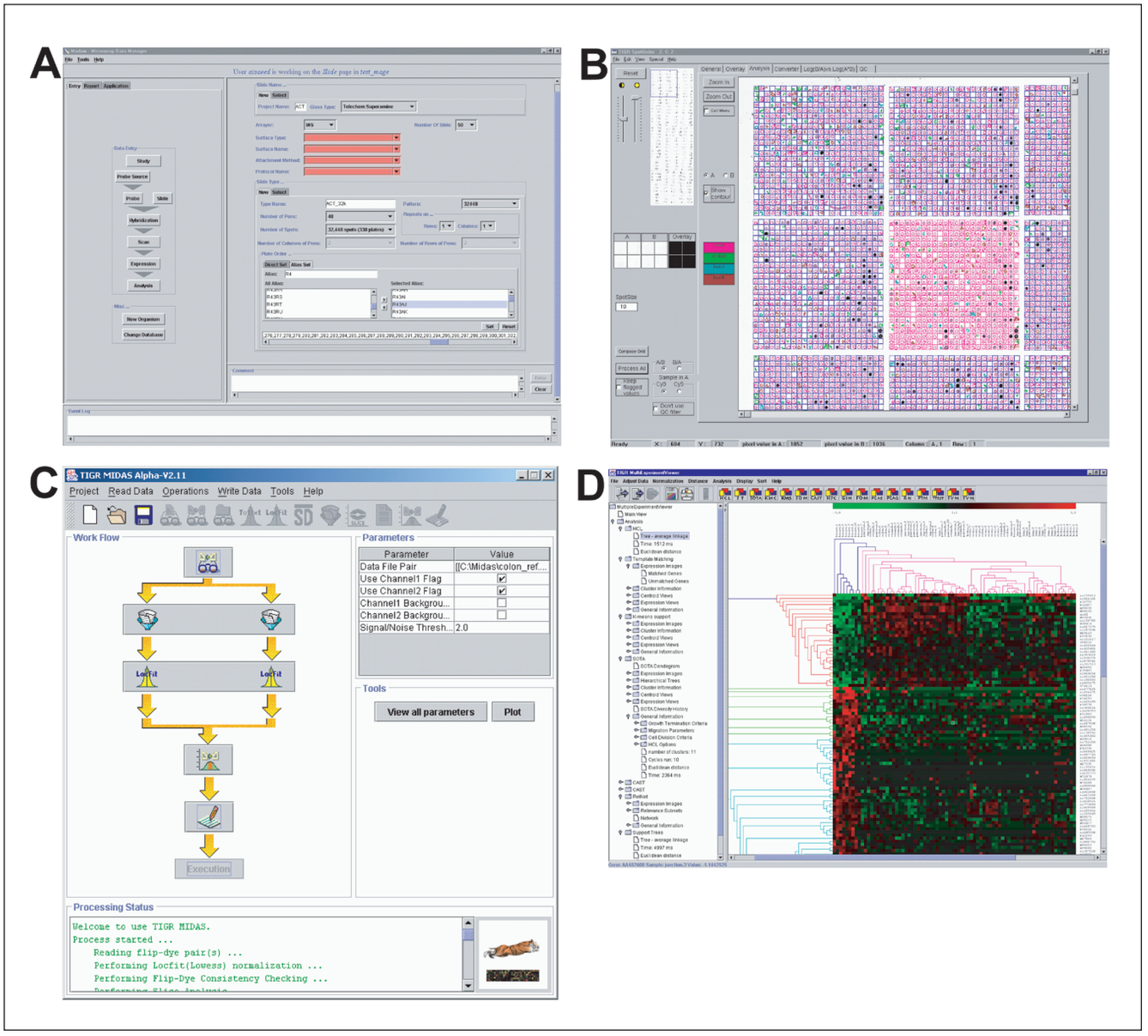

Figure 1. Representative screen shots from TM4. (A) The MADAM data entry interface provides access to data associated with a microarray study. A navigation panel on the left-hand side leads users through the process of data entry during a microarray experiment. Shown here is a page for entering information about the material printed onto each array slide used in a particular study. (B) TIGR Spotfinder provides image processing with direct connections to the microarray database. Here, a grid has been defined and overlaid on the scanned image file. Good (pink contours) and bad spots have been identified and spot contours are clearly displayed. (C) MIDAS allows users to define a data normalization and filtering protocol using a simple graphical scripting interface. The diagram on the left shows the steps in the analysis to be carried out; the panel on the right allows users to enter parameters for each stage in the analysis. (D) MeV allows users to apply a number of sophisticated data mining tools to their array data and provides integrated graphical depictions of the results of the analyses conducted.

model plant *Arabidopsis thaliana*, and expression in a variety of other plant, animal, and microbial species. Our databases contain data from well over 3500 hybridization assays, and the system has evolved through the comments of our many users at TIGR and the more than 2000 users of the system elsewhere. The analysis enabled by TM4 has resulted in a number of publications (1,2,9,11,12,16,19), and the system continues to evolve to meet the challenges of new experimental designs and systems.

As we look to the future, we can expect increasing quantities of gene expression data. The creation of novel algorithms and growing acceptance of a system for information exchange will support this trend. TM4 stands ready to provide researchers with the necessary tools for managing all stages of the microarray process, and we invite the microarray community to participate in its ongoing development by providing new analysis modules that can be made available to researchers. Our goal is to ensure that TM4 remains a full-featured and community-driven solution to the challenge of microarray data management and analysis.

ACKNOWLEDGMENTS

The authors wish to thank Eric Snesrud, Robin Cline, Ivana Yang, Hong-Ying Wang, Yonghong Wang, Simon Kwong, Heenam Kim, Jeremy Hasseman, Priti Hegde, Annie Simpson, and other members of the faculty and staff at TIGR for valuable contributions in the development of this software. We also thank Susan Lo and Michael Heaney for expert assistance in database development and the remaining members of the TIGR IT staff for their support. This work was funded by grants from the US National Cancer Institute, the US National Heart, Lung, Blood Institute, US National Science Foundation, and the US Department of Energy.

REFERENCES

1. **Agrawal, D., T. Chen, R. Irby, J. Quackenbush, A.F. Chambers, M. Szabo, A. Cantor, D. Coppola, and T.J. Yeatman.** 2002. Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. J. Natl. Cancer Inst. *94*:513-521.
2. **Andersson, T., P. Unneberg, P. Nilsson, J. Odeberg, J. Quackenbush, and J. Lundeberg.** 2002. Monitoring of representational difference analysis subtraction procedures by global microarrays. BioTechniques *32*:1348-1358.
3. **Ben-Dor, A., R. Shamir, and Z. Yakhini.** 1999. Clustering gene expression patterns. J. Comput. Biol. *6*:281-297.
4. **Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, et al.** 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*:365-371.
5. **Brown, M.P., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler.** 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA *97*:262-267.
6. **Butte, A.J., P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane.** 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc. Natl. Acad. Sci. USA *97*:12182-12186.
7. **Cleveland, W. and S. Devlin.** 1988. Locally weighted linear regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. *83*:596-609.
8. **Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
9. **El-Sayed, N.M., P. Hegde, J. Quackenbush, S.E. Melville, and J.E. Donelson.** 2000. The African trypanosome genome. Int. J. Parasitol. *30*:329-345.
10. **Hastie, T., R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown.** 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. *1*:RESEARCH0003.
11. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush.** 2000. A concise guide to cDNA microarray analysis. BioTechniques *29*:548-556.
12. **Hegde, P., R.R. Qi, R. Gaspard, K. Abernathy, S. Dharap, J. Earle-Hughes, C. Gay, N.U. Nwokekeh, et al.** 2001. Identification of tumor markers in models of human colorectal cancer using a 19,200-element complementary DNA microarray. Cancer Res. *61*:7792-7797.
13. **Herrero, J., A. Valencia, and J. Dopazo.** 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics *17*:126-136.
14. **Heyer, L.J., S. Kruglyak, and S. Yooseph.** 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*:1106-1115.
15. **Kohonen, T.** 1992. Self-organized formation of topologically correct feature maps. Biol. Cybernetics *43*:59-69.
16. **Malek, R.L., R.B. Irby, Q.M. Guo, K. Lee, S. Wong, M. He, J. Tsai, B. Frank, et al.** 2002. Identification of Src transformation fingerprint in human colon cancer. Oncogene *21*:7256-7265.
17. **Raychaudhuri, S., J.M. Stuart, and R.B. Altman.** 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 455-466.
18. **Soukas, A., P. Cohen, N.D. Socci, and J.M. Friedman.** 2000. Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev. *14*:963-980.
19. **Yang, I.V., E. Chen, J.P. Hasseman, W. Liang, B.C. Frank, S. Wang, V. Sharov, A.I. Saeed, et al.** 2002. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol. *3*:RESEARCH0062.
20. **Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo.** 2001. Validating clustering for gene expression data. Bioinformatics *17*:309-318.

Received 23 October 2002; accepted 2 December 2002.

Address correspondence to:
Dr. John Quackenbush
The Institute for Genomic Research (TIGR)
Rockville, MD, USA
e-mail: johnq@tigr.org