2005

Microarray Databases for Biotechnology

Richard S. Segall
Arkansas State University - Jonesboro, rsegall@astate.edu

Microarray Databases for Biotechnology

Richard S. Segall
Arkansas State University, USA

INTRODUCTION

Microarray informatics is a rapidly expanding discipline in which large amounts of multi-dimensional data are compressed into small storage units. Data mining of microarrays can be performed using techniques such as drill-down analysis rather than classical data analysis on a record-by-record basis. Both data and metadata can be captured in microarray experiments. The latter may be constructed by obtaining data samples from an experiment. Extractions can be made from these samples and formed into homogeneous arrays that are needed for higher level analysis and mining.

Biologists and geneticists find microarray analysis as both a practical and appropriate method of storing images, together with pixel or spot intensities and identifiers, and other information about the experiment.

BACKGROUND

A Microarray has been defined by Schena (2003) as "an ordered array of microscopic elements in a planar substrate that allows the specific binding of genes or gene products." Schena (2003) claims microarray databases as "a widely recognized next revolution in molecular biology that enables scientists to analyze genes, proteins, and other biological molecules on a genomic scale."

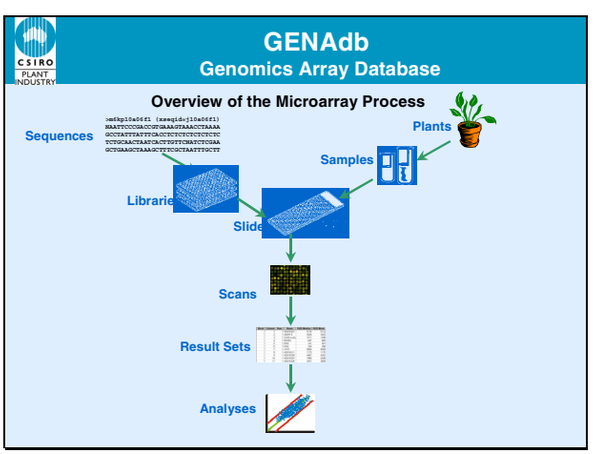
According to an article (2004) on the National Center for Biotechnology Information (NCBI) Web site, "because microarrays can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression," and "this technology is still considered to be in its infancy."

The following *Figure 1* is from a presentation by Kennedy (2003) of CSIRO (Commonwealth Scientific & Industrial Research Organisation) in Australia as available on the Web, and illustrates an overview of the microarray process starting with sequence data of individual clones that can be organized into libraries. Individual samples are taken from the library as spots and arranged by robots onto slides that are then scanned by lasers. The image scanned by lasers is than quantified according to the color generated by each individual spot that are then organized into a results set as a text file that can then be subjected to analyses such as data mining.

Jagannathan (2002) of the Swiss Institute of Bioinformatics (SIB) described databases for microarrays including their construction from microarray experiments such as gathering data from cells subjected to more than one conditions. The latter are hybridized to a microarray that is stored after the experiment by methods such as scanned images. Hence data is to be stored both before and after the experiments, and the software used must be capable of dealing with large volumes of both numeric and image data. Jagannathan (2002) also discussed some of the most promising existing non-commercial microarray databases of ArrayExpress, which is a public microarray gene expression repository, the Gene Express Omnibus (GEO), which is a gene expression database hosted at the National Library of Medicine, and GeneX, which is an open source database and integrated tool set released by the National Center for Genome Resources (NCGR) in Santa Fe, New Mexico.

Grant (2001) wrote an entire thesis on microarray databases describing the scene for its application to genetics and the human genome and its sequence of the three billion-letter sequences of genes. Kim (2002) presented improved analytical methods for micro-array based genome composition analysis by selecting a signal value that is used as a cutoff to discriminate present

Figure 1. Overview of the microarray process (Kennedy, (2003)



and divergent genes. Do et al. (2003) provided comparative evaluation of microarray-based gene expression databases by analyzing the requirements for microarray data management, and Sherlock (2003) discussed storage and retrieval of microarray data for molecular biology.

Kemmeren (2001) described a bioinformatics pipeline for supporting microarray analysis with example of production and analysis of DNA (Deoxyribonucleic Acid) microarrays that require informatics support. Gonclaves & Marks (2002) discussed roles and requirements for a research microarray database.

An XML description language called MAML (Microarray Annotation Markup Language) has been developed to allow communication with other databases worldwide (Cover Pages 2002). Liu (2004) discusses microarray databases and MIAME (Minimal Information about a Microarray Experiment) that defines what information at least should be stored. For example, the MIAME for array design would be the definite structure and definition of each array used and their elements. The Microarray Gene Expression Database Group (MGED) composed and developed the recommendations for microarray data annotations for both MAIME and MAML in 2000 and 2001 respectively in Cambridge, United Kingdom.

Jonassen (2002) presents a microarray informatics resource Web page that includes surveys and introductory papers on informatics aspects, and database and software links. Another resourceful Web site is that from the Lawrence Livermore National Labs (2003) entitled Microarray Links that provides an extensive list of active Web links for the categories of databases, microarray labs, and software and tools including data mining tools.

University-wide database systems have been established such as at Yale as the Yale Microarray Database (YMD) to support large-scale integrated analysis of large amounts of gene expression data produced by a wide variety of microarray experiments for different organisms as described by Cheung (2004), and similarly at Stanford with Stanford Microarray Database (SMD) as described by both Sherlock (2001) and Selis (2003).

Microarray Image analysis is currently included in university curricula, such as in Rouchka (2003) Introduction to Bioinformatics graduate course at University of Louisville.

In relation to the State of Arkansas, the medical school is situated in Little Rock and is known as the University of Arkansas for Medical Sciences (UAMS). A Bioinformatics Center is housed within UAMS that is involved with the management of microarray data. The software utilized at UAMS for microarray analysis includes BASE (BioArray Software Environment) and AMAD, which is a Web driven database system written entirely in PERL and JavaScript (UAMS, Bioinformatics Center, 2004).

MAIN THRUST

The purpose of this article is to help clarify the meaning of microarray informatics. The latter is addressed by summarizing some illustrations of applications of data mining to microarray databases specifically for biotechnology.

First, it needs to be stated which data mining tools are useful in data mining of microarrays. SAS Enterprise Miner, which was used in Segall et al. (2003, 2004a, 2004b) as discussed below contains the major data mining tools of decisions trees, regression, neural networks, and clustering, and also other data mining tools such as association rules, variable selection, and link analysis. All of these are useful data mining tools for microarray databases regardless if using SAS Enterprise Miner or not. In fact, an entire text has been written by Draghici (2003) on data analysis tools for DNA microarrays that includes these data mining tools as well as numerous others tools such as analysis of functional categories and statistical procedure of corrections for multiple comparisons.

Scientific and Statistical Data Mining and Visual Data Mining for Genomes

Data mining of microarray databases has been discussed by Deyholos (2002) for bioinformatics by methods that include correlation of patterns and identifying the significance analysis of microarrays (SAM) for genes within DNA. Visual data mining was utilized to distinguish the intensity of data filtering and the effect of normalization of the data using regression plots.

Tong (2002) discusses supporting microarray studies for toxicogenomic databases through data integration with public data and applying visual data mining such as ScatterPlot viewer.

Chen et al. (2003) presented a statistical approach using a Gene Expression Analysis Refining System (GEARS).

Piatetsky-Shapiro and Tamayo (2003) discussed the main types of challenges for microarrray data mining as including gene selection, classification, and clustering. According to Piatetsky-Shapiro and Tamayo (2003), one of the important challenges for data mining of microarrays is that "the difficulty of collecting microarray samples causes the number of samples to remain small" and "while

the number of fields corresponding to the number of genes is typically in the thousands" this "creates a high likelihood of finding false positives."

Piatetsky-Shapiro and Tamayo (2003) identify areas in which micorarrays and data mining tools can be improved that include "better accuracy, more robust models and estimators" as well as better appropriate biological interpretation of the computational or statistical results for those microarrays constructed from biomedical or DNA data.

Piatetsky-Shapiro and Tamayo (2003) summarize up the areas in which microarray and microarry data mining tools can be improved by stating:

Typically a computational researcher will apply his or her favorite algorithm to some microarray dataset and quickly obtain a voluminous set of results. These results are likely to be useful but only if they can be put in context and followed up with more detailed studies, for example by a biologist or a clinical researcher. Often this follow up and interpretation is not done carefully enough because of the additional significant research involvement, the lack of domain expertise or proper collaborators, or due to the limitations of the computational analysis itself.

Draghici (2003) discussed in-depth other challenges in using microarrays specifically for gene expression studies, such as being very noisy or prone to error after the scanning and image processing steps, consensus as to how to perform normalization, and the fact that microarrays are not necessarily able to substitute completely other biological factors or tools in the realm of the molecular biologist.

Mamitsuka et al. (2003) mined biological active patterns in metabolic pathways using microarray expression profiles. Mamitsuka (2003) utilized microarray data sets of gene expressions on yeast proteins.

Curran et al. (2003) performed statistical methods for joint data mining of gene expressions and DNA sequence databases. The statistical methods used include linear mixed effect model, cluster analysis, and logistic regression.

Zaki et al. (2003) reported on an overview of the papers on data mining in bioinformatics as presented at the International Conference on Knowledge Discovery and Data Mining held in Washington, DC in August 2003. Some of the novel data mining techniques discussed in papers at this conference included gene expression analysis, protein/RNA (ribonucleic acid) structure prediction, and gene finding.

Scientific and Statistical Data Mining and Visual Data Mining for Plants

Segall et al. (2003, 2004a, 2004b) performed data mining for assessing the impact of environmental stresses on plant geonomics and specifically for plant data from the Osmotic Stress Microarray Information Database (OSMID). The latter databases are considered to be representative of those that could be used for biotech application such as the manufacture of plant-made-pharmaceuticals (PMP) and genetically modified (GM) foods.

The Osmotic Stress Microarray Information Database (OSMID) database that was used in the data mining in Segall et al. (2003, 2004a, 2004b) contains the results of approximately 100 microarray experiments performed at the University of Arizona as part of a National Science Foundation (NSF) funded project named the "The Functional Genomics of Plant Stress" whose data constitutes a data warehouse.

The OSMID microarray database is available for public access on the Web hosted by Universite Montpellier II (2003) in France, and the OSMID contains information about the more than 20,000 ESTs (Experimental Stress Tolerances) that were used to produce these arrays. These 20,000 ESTs could be considered as components of data warehouse of plant microarray databases that was subjected to data mining in Segall et al. (2003, 2004a, 2004b). The data mining was performed using SAS Enterprise Miner and its cluster analysis module that yielded both scientific and statistical data mining as well as visual data mining. The conclusions of Segall et al. (2003, 2004a, 2004b) included the facts about the twenty-five different variations or levels of the environmental factor of salinity on plant of corn, as also evidenced by the visualization of the clusters formed as a result of the data mining.

Other Useful Sources of Tools and Projects for Microarray Informatics

- A bibliography on microarray data analysis created as available on the Web by Li (2004) that includes book and reprints for the last ten years.
- The Rosalind Franklin Centre for Genomics Research (RFCGR) of the Medical Research Council (MRC) (2004) in the UK provides a Web site with links for data mining tools and descriptions of their specific applications to gene expressions and microarray databases for genomics and genetics.
- Reviews of data mining software as applied to genetic microarray databases are included in an

annotated list of references for microarray soft ware review compiled by Leung et al. (2002).

- Web links for the statistical analysis of microarray data are provided by van Helden (2004).
- Reid (2004) provides Web links of software tools for microarray data analysis including image analysis.
- *Bio-IT World Journal* Web site has a Microarray Resource Center that includes a link of extensive resources for microarray informatics at the European Bioinformatics Institute (EBI).

FUTURE TRENDS

The wealth of resources available on the Web for microarray informatics only supports the premise that microarray informatics is a rapidly expanding field. This growth is in both software and methods of analysis that includes techniques of data mining.

Future research opportunities in microarray informatics include the biotech applications for manufacture of plant-made-pharmaceuticals (PMP) and genetically modified (GM) foods.

CONCLUSION

Because data within genome databases is composed of micro-level components such as DNA, microarray databases are a critical tool for analysis in biotechnology. Data mining of microarray databases opens up this field of microarray informatics as multi-facet tools for knowledge discovery.

ACKNOWLEDGMENT

The author wishes to acknowledge the funding provided by a block grant from the Arkansas Biosciences Institute (ABI) as administered by Arkansas State University (ASU) to encourage development of a focus area in Biosciences Institute Social and Economic and Regulatory Studies (BISERS) for which he served as Co-Investigator (Co-I) in 2003, and with which funding the analyses of the Osmotic Stress Microarray Information Database (OSMID) discussed within this article were performed.

The author also wishes to acknowledge a three-year software grant from SAS Incorporated to the College of Business at Arkansas State University for SAS Enterprise Miner that was used in the data mining of the OSMID microarrays discussed within this article.

Finally, the author also wishes to acknowledge the useful reviews of the three anonymous referees of the earlier version of this article without whose constructive comments the final form of this article would not have been possible.

REFERENCES

Bio-IT World Inc. (2004). *Microarray resources and articles*. Retrieved from *http://www.bio-itworld.com/resources/microarray/*

Chen, C.H. et al. (2003). Gene expression analysis refining system (GEARS) via statistical approach: A preliminary report. *Genome Informatics*, *14*, 316-317.

Cheung, K.H. et al. (2004). *Yale Microarray Database System*. Retrieved from *http://crcjs.med.utah.edu/bioinfo/abstracts/Cheung,%20Kei.doc*

Curran, M.D., Liu, H., Long, F., & Ge, N. (2003,December). Machine learning in low-level microarray analysis. *SIGKDD Explorations*, *5*(2),122-129.

Deyholos, M. (2002). An introduction to exploring genomes and mining microarrays. In *O'Reilly Bioinformatics Technology Conference*, January 28-31, 2002, Tucson, AZ. Retrieved from *http://conferences.oreillynet.com/cs/bio2002/view/e_sess/1962 - 11k - May 7, 2004*

Do, H., Toralf, K., & Rahm, E. (2003). *Comparative evaluation of microarray-based gene expression databases*. Retrieved from *http://www.btw2003.de/proceedings/paper/96.pdf*Draghici, S. (2003). *Data analysis tools for DNA microarrays*. Boca Raton, FL: Chapman & Hall/CRC.

Goncalves, J., & Marks, W.L. (2002). Roles and requirements for a research microarray Database. *IEEE Engineering Medical Biol Magazine*, *21*(6), 154-157.

Grant, E. (2001, September). *A microarray database*. Thesis for Masters of Science in Information Technology. The University of Glasgow.

Jagannathan, V. (2002). *Databases for microarrays*. Presentation at Swiss Institute of Bioinformatics (SIB), University of Lausanne, Switzerland. Retrieved from *http://www.ch.embnet.org/CoursEMBnet/CHIP02/ppt/Vidhya.ppt*

Jonassen, I. (2002). *Microarray informatics resource page*. Retrieved from *http://www.ii.uib.no/~inge/micro*

Kemmeren, P.C., & Holstege, F.C. (2001). *A bioinformatics pipeline for supporting microarray analysis*. Retrieved

from *http://www.genomics.med.uu.nl/presentations/Bioinformatics-2001-Patrick2.ppt*

Kennedy, G. (2003). *GENAdb: Genomics Array Database*. CSIRO (Commonwealth Scientific & Industrial Research Organisation) Plant Industry, Australia. Retrieved from *http://www.pi.csiro.au/gena/repository/GENAdb.ppt*

Kim, C.K., Joyce E.A., Chan, K., & Falkow, S. (2002). Improved analytical methods for microarray-based genome composition analysis. *Genome Biology,* 3(11). Retrieved from *http://genomebiology.com/2002/3/11/research/0065*

Lawrence Livermore National Labs. (2003). *Microarray Links*. Retrieved from *http://microarray.llnl.gov/links.html*.

Leung, Y.F. (2002). Microarray software review. In D. Berrar, W. Dubitzky & M. Granzow (Eds.), *A practical approach to microarray data analysis* (pp. 326-344). Boston: Kluwer Academic Publishers.

Li, W. (2004). *Bibliography on microarray data analysis*.Retrieved from *http://www.nslij-genetics.org.microarray/2004.html*

Liu, Y. (2004). *Microarray Databases and MIAME (Minimum Information About a Microarray Experiment)*. Retrieved from *http://titan.biotec.uiuc.edu/cs491jh/slides/cs491jh-Yong.ppt*

Mamitsuka, H., Okuno, Y., & Yamaguchi, A. (2003, December). Mining biological active patterns in metabolic pathways using microarray expression profiles. *SIGKDD Explorations*, *5*(2), 113-121.

Medical Research Council. (2004). *Genome Web: Gene expression and microarrays*. Retrieved from *http://www.rfcgr.mrc.ac.uk/GenomeWeb/nuc-genexp.html*

Microarray Markup Language (MAML). (2002, February 8). *Cover Pages Technology Reports*. Retrieved from *http://xml.coverpages.org/maml.html*

National Center for Biotechnology Information (NCBI). (2004, March 30). Microarrays: Chipping away at the mysteries of science and medicine. National Library of Medicine (NLM), National Institutes of Health (NIH). Retrieved from *http://www.ncbi.nlm.nih.gov/About/primer/microarrys.html*

Piatetsky-Shapiro, G., & Tamayo, P. (2003, December). Microarray data mining: Facing the challenges. *SIGKDD Explorations*, *5*(2), 1-5.

Reid, J.F. (2004). Software tools for microarray data analysis. Retrieved from *http://www.ifom-firc.it/MICROARRAY/data_analysis.htm*

Rouchka, E. (2003). *CECS 694 Introduction to Bioinformatics*. Lecture 12. Microarray Image Analysis, University of Louisville. Retrieved from *http://kbrin.a-bldg.louisville.edu/~rouchka/CECS694_ 2003/Week12.html*

Schena, M. (2003). *Microarray analysis*. New York: John Wiley & Sons.

Segall, R.S., Guha, G.S., & Nonis, S. (2003). *Data mining for analyzing the impact of environmental stress on plants: A case study using OSMID.* Manuscript in preparation for journal submission.

Segall, R.S., Guha, G.S., & Nonis, S. (2004b, May). Data mining for assessing the impact of environmental stresses on plant geonomics. In *Proceedings of the Thirty-Fifth Meeting of the Southwest Decision Sciences Institute* (pp. 23-31). Orlando, FL.

Segall, R.S., & Nonis, S. (2004a, February). Data mining for analyzing the impact of environmental stress on plants: A case study using OSMID. Accepted for publication in *Acxiom Working Paper Series of Acxiom Laboratory of Applied Research (ALAR)* and presented at *Acxiom Conference on Applied Research and Information Technology*, University of Arkansas at Little Rock (UALR).

Selis, S. (2003, February 15). *Stanford researcher advocates far-reaching microarray data exchange*. News release of Stanford University School of Medicine. Retrieved from *http://www.stanfordhospital.com/newsEvents/mewsReleases/2003/02/aaasSherlock.html*

Sherlock, G. et al. (2001). The Stanford microarray database. *Nucleic Acids Research*, *29*(1), 152-155.

Sherlock, G., & Ball, C.A. (2003). Microarray databases: Storage and retrieval of microarray data. In M.J. Brownstein & A. Khodursky (Eds.), *Functional genomics: Methods and protocols* (pp. 235-248). Methods in Molecular Biology Series (Vol. 224). Totowa, NJ, Humana Press.

Tong, W. (2002, December). *ArrayTrack-Supporting microarray studies through data integration*. U. S. Food and Drug Administration (FDA)/National Center for Toxicological Research (NCTR) Toxioinformatics Workshop: Toxicogenomics Database, Study Design and Data Analysis.

Universite Montpellier II. (2003). *The Virtual Library of Plant-Array: Databases*. Retrieved from *http://www.univ-montp2.fr/~plant_arrays/databases.html*

University of Arkansas for Medical Sciences (UAMS) Bioinformatics Center. (2004). Retrieved from *http://bioinformatics.uams.edu/microarray/database.html*

Van Helden, J. (2004). *Statistical analysis of microarray data: Links*.Retrieved from *http://www.scmbb.ulb.ac.be/~jvanheld/web_course_microarrays/links.html*

Zaki, M.J., Wang, H.T., & Toivonen, H.T. (2003, December). Data mining in bioinformatics. *SIGKDD Explorations*, *5*(2), 198-199.

KEY TERMS

Data Warehouses: A huge collection of consistent data that is both subject-oriented and time variant, and used in support of decision-making.

Genomic Databases: Organized collection of data pertaining to the genetic material of an organism.

Metadata: Data about data, for example, data that describes the properties or characteristics of other data.

MIAME (Minimal Information about a Microarray Experiment): Defines what information at least should be stored.

Microarray Databases: Store large amounts of complex data as generated by microarray experiments (e.g., DNA).

Microarray Informatics: The study of the use of microarray databases to obtain information about experimental data.

Microarray Markup Language (MAML): An XML (Extensible Markup Language)-based format for communicating information about data from microarray experiments.

Scientific and Statistical Data Mining: The use of data and image analyses to investigate knowledge discovery of patterns in the data.

Visual Data Mining: The use of computer generated graphics in both 2-D and 3-D for the use in knowledge discovery of patterns in data.