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."

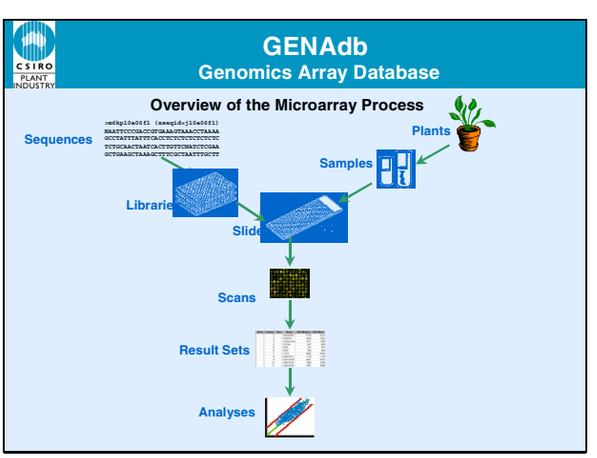
According to an article (2004) on the National Center for Biotechnology Information (NCBI) Web site, "because microarrays can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression," and "this technology is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