DRUG DISCOVERY TODAY BIOSILICO

Storage and transmission of microarray images

Yu Luo and Stefano Lonardi

With the recent explosion of interest in microarray technology, massive amounts of microarray images are currently being produced. The storage and transmission of these types of data are becoming increasingly challenging. This article reviews the latest technologies that allow for the compression and storage of microarray images in dedicated database systems.

The first attempts at global analysis of gene expression were undertaken in the mid-1970s with studies of the hybridization of an mRNA pool with radioactively labeled cDNA. Interest in gene expression increased steadily during the 1980s, and in the 1990s a new era of high throughput gene expression studies unfolded with the development of microarrays [1,2]. Although microarrays are relatively new, their penetration has been phenomenal. Biologists and physicians have enthusiastically embraced this technology, and are currently producing an unprecedented quantity of microarray data. As of March 2005, for example, the Stanford microarray database (SMD) [3] contained sets of images for ~54,000 microarrays, relating to the biology of 35 organisms. The database has been growing exponentially since its inception.

Microarrays allow studies of gene expression on a massively parallel scale; a single microarray experiment can provide information on the expression of thousands of genes. More specifically, a microarray experiment is designed to compare the transcriptional activity of a set of genes under two conditions (hereafter called 'reference' and 'experimental'). The outcome is a quantitative measure of the relative change of expression of each gene in an experimental condition compared with the reference condition.

Currently, a few variants of microarray technology exist. To simplify the discussion in this review, we will describe the so-called cDNA microarrays. Proprietary technologies developed by Affymetrix and NimbleGen Systems employ somewhat different strategies – although the general principles remain the same. cDNA microarrays are fabricated on a glass or a nylon substrate by specialized high-speed robots. The fabrication process creates thousands of microscopic spots containing DNA probes that are immobilized in the substrate. The DNA probes are chosen to hybridize to unique sequences in the genes being studied.

The mRNA from the two different conditions is obtained separately and reverse transcriptase is used to transcribe the mRNA into cDNA. The cDNA is labeled with a green or red dye, depending on which conditions it corresponds to. The microarray chip is then exposed to a mix of the two populations of cDNAs. A given strand of cDNA will hybridize with the DNA probe that was selected from the gene that produced that transcript. The chip is then washed to remove any unbound cDNAs (see [1,2] for more details on the protocols). The microarray chip is finally scanned using a confocal laser or a charge-coupled device (CCD) to generate two digital images,

Yu Luo*
Stefano Lonardi**
Department of Computer Science & Engineering, University of California, Riverside, CA 92521, USA
*e-mail: yuluo@cs.ucr.edu
**e-mail: stelo@cs.ucr.edu

1359-6446/05/$ – see front matter PII: S1359-6446(05)03647-0

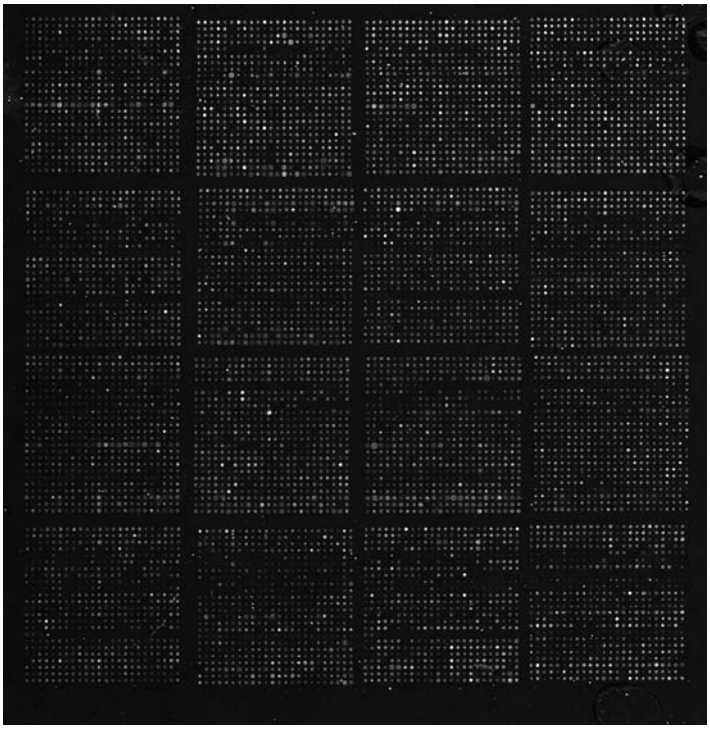

FIGURE 1

A typical microarray image, composed by 4x4 sub grids. Each sub grid is composed of a 24×24 matrix of spots. The original resolution of this image is 1872×1916 pixels, 16 bpp.

each corresponding to one of the colors [4,5]. The output of a single microarray experiment is a pair of 16 bits per pixel (bpp) digital images whose total size is typically in the tens of megabytes [6].

Microarray images are usually structured as a series of high intensity spots located on a regular grid. For example, Figure 1 shows a microarray with 16 subgrids for a total of ~9000 spots. A variety of methods and software tools are available to extract the gene expression information from microarray images (reviewed in [4]). However, there is still a debate in the scientific community about the accuracy, reliability and robustness of the analytical and statistical methods that need to be used. In practice, microarray images are challenging to process. For example, when the image contains high levels of noise, the detection of low-intensity spots is easily compromised. Because there is not yet a clear choice of a specific analytic method, it is believed crucial to keep the raw image data in some permanent storage medium [7]. Simply discarding the raw data and repeating the experiment is not an option because of the high costs associated with each experiment. Given the massive amount of data produced and the need of long-term storage and efficient transmission, *ad hoc* compression methods and dedicated database systems are currently an active area of research in computational biology.

Before turning the attention to the current research in microarray image compression, we report on the image compression formats employed by some of the microarray databases currently available (Table 1 is a summary of some commonly used image formats). The focus of this review is mainly on databases in the public domain because of the difficulties involved in finding technical specifications of proprietary platforms.

Storage of microarray images

The importance of organizing and storing the data of microarray experiments in relational databases cannot be overemphasized. Many microarray users are still struggling in the transition from spreadsheets to databases designed to handle the explosive growth of their microarray datasets. An international effort led by the Microarray Gene Expression Data (MGED) Society (www.mged.org) is under way to 'establish standards for microarray data annotation and exchange, facilitate the creation of microarray databases and related software implementing these standards, and promote the sharing of high quality, well annotated data within the life sciences community'. The MGED group has prepared standards on the 'minimum information about a microarray experiment' (MIAME) that requires all the information needed to interpret, share [8,9] and possibly replicate the results of a microarray experiments to be recorded in the database and made public [10–12]. A relational database system allows users to process and store the large quantities of data produced by microarray experiments and thereby accommodate the enforcements of standards like MIAME. Moreover, because microarray experiments typically depend on the work of several people in the laboratory, database systems can easily enforce common principles in data format and data entry [13]. To facilitate the exchange of information between databases, the MGED group has also designed a standard called microarray mark-up language (MAML).

A proliferation of microarray databases has occurred in response to the demands of the life science communities [14–16]. We have reviewed several databases both for local installation and public data repositories, and here we present our findings. Public data repositories allow multiple users to access the data remotely via a browser, whereas local installations allow access only through the machine in which the database is installed. Some databases are not designed to manage images but only gene expression data [e.g. ChipDB (http://chipdb.wi.mit.edu) or AMAD (www.microarrays.org)].

Local installation databases, such as GeneDirector (www.biodiscovery.com), mAdb (http://madb.niaid.nih.gov), maxdSQL (http://bioinf.man.ac.uk/microarray/maxd) and the SMD (http://genome-www5.stanford.edu) [3,17], store file images in TIFF format and allow users direct access to them. SMD allows images in GIF format to be obtained for viewing purposes or web posting and the mAdb database can export images in JPEG format for presentation. Among the databases for public data repositories, ArrayExpress (www.ebi.ac.uk/arrayexpress) [17–20], the RNA Abundance