Mining Microarray Data

Nanxiang Ge
Aventis, USA

Li Liu
Aventis, USA

INTRODUCTION

During the last 10 years and in particularly within the last few years, there has been a data explosion associated with the completion of the human genome project (HGP) (IHGMC and Venter *et al*., 2001) in 2001 and the many sophisticated genomics technologies. The human genome (and genome from other species) now provides an enormous amount of data waiting to be transformed into useful information and scientific knowledge. The availability of genome sequence data also sparks the development of many new technology platforms. Among the available different technology platforms, microarray is one of the technologies that is becoming more and more mature and has been widely used as a tool for scientific discovery. The major application of microarray is for simultaneously measuring the expression level of thousands of genes in the cell. It has been widely used in drug discovery and starts to impact the drug development process.

The mining microarray database is one of the many challenges facing the field of bioinformatics, computational biology and biostatistics. We review the issues in microarray data mining.

BACKGROUND

The quantity of data generated from a microarray experiment is usually large. Besides the actual gene expression measurement, there are many other data available such as the corresponding sequence information, gene's chromosome location information, gene ontology classification and gene pathway information. Management of such data is a very challenging issue for bioinformatics. The following is a list of requirements for data management.

- **Data Organization:** different microarray platforms generate data with different formats, different gene identifiers, etc. Sequence data, gene ontology data, and gene pathway information all with diverse formats. Sensible organization of such diverse data types will ease the data mining process.
- **Data Standards:** the microarray data analysis community has developed a standard for microarray data: the Minimum Information About a Microarray Experiment (MIAME: http://www.mged.org). The MIAME standard is needed to enable the consistent interpretation of experiment results and potentially to reproduce the experiment.

MAIN THRUST

Mining microarray data is a very challenging task. The raw data from microarray experiments usually comes in image format. Images are then quantified using image analysis software. Such quantified data are then subject to three steps of analysis:

- Pre-processing microarray data
- Mining microarray data.
- Joint mining of microarray data and sequence database.

We review the data analysis methods in these three aspects. We focus our discussion on Affymetrix (http://www.affymetrix.com) technology, but many of the methods are applicable to data from other platforms.

Pre-Processing Microarray Data

While effectively managing microarray and related data is an important first step, microarray data have to be pre-processed so that downstream analysis can be performed. Array-to-array and sample-to-sample variations are the main reason for the requirement of pre-processing. Typically, pre-processing involves the following four steps:

- **Image Analysis:** image analysis in microarray experiment involves gridding of image, signal extraction and back ground adjustment. Affymetrix's

microarray analysis suite (MAS) provides the quantification software.

- **Data Normalization:** normalization is a step necessary for remove systematic array to array variations. Different normalization methods have been proposed. Cyclic loess method (Dudoit et al., 2002), quantile normalization method and contrast-based method normalize the probe level data; Scaling method and non-linear method normalize expression intensity level data. Bolstad, Irizarry, Astrand, and Speed (2003) provides a comparison of all these methods and suggests that simple quantile normalization performs relatively stable.
- **Estimation of Expression Intensities:** different methods to summarize expression intensity from probe level data have appeared in literature. Among them, Li and Wong (2001) proposed a model based expression intensity (MEBI), Irizarry's (2003) robust multi-array (RMA) methods and the method provided in Affymetrix MAS5 software.
- **Data Transformation:** data transformation is a very important step and it enables the data to fit to many of the assumptions behind statistical methods. Log transformation and glog (Durbin, Hardin, Hawkins, & Rocke, 2002) transformation are the two commonly used methods.

Mining Microarray Data

In principal, the current data mining activities in microarray data can be grouped into two types of studies: unsupervised and supervised. Unsupervised analysis has been used widely for mining microarray experiment. Cluster analysis has been the dominant method for unsupervised mining.

Examples of unsupervised data mining:

- Eisen, Spellman, Brown, and Botstein (1998) studied the gene expression of budding yeast *Saccharomyces cerevisiae* spotted on cDNA microarrays during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks. Hierarchical clustering was applied to this gene expression data, and the result was represented by a tree whose branch lengths reflected the degree of similarity between genes, which was assessed by a pair-wise similarity function. The computed tree was then used to order genes in the original data table, and genes with similar expression pattern were grouped together. The ordered gene expression table can be displayed graphically in a colored image, where cells with log ratios of 0's were colored black, cells with positive log ratios were colored red, cells with negative log ratios were colored green, and the intensities of reds or greens were proportional to the absolute values of the log ratios. The clustering analysis efficiently grouped genes with similar functions together, and the colored image provided an overall pattern in the data. Clustering analysis can also help us understand the novel genes if they were co-expressed with genes with known functions.
- Standard clustering analysis, such as hierarchical clustering, k-means clustering, self-organizing maps, are very useful in mining the microarray data. However, these data tables are often corrupted with extreme values (outliers), missing values, and non-normal distributions that preclude standard analysis. Liu, Hawkins, Ghosh, and Young (2003) proposed a robust analysis method, called rSVD (Robust Singular Value Decomposition), to address these problems. The method applies a combination of mathematical and statistical methods to progressively take the data set apart so that different aspects can be examined for both general patterns and for very specific effects. The benefits of this robust analysis will be both the understanding of large-scale shifts in gene effects and the isolation of particular sample-by-gene effects that might be either unusual interactions or the result of experimental flaws. The method requires a single pass, and does not resort to complex "cleaning" or imputation of the data table before analysis. The method rSVD was applied to a micro array data, revealed different aspects of the data, and gave some interesting findings.

Examples for supervised data mining:

- Golub *et al*. (1999) studied the gene expression of two types of acute leukemias, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), and demonstrated the feasibility of cancer prediction based on gene expression data. The data comes from Affymetrix arrays with 6817 genes, and consists of 47 cases of ALL and 25 cases of AML. 38 samples (27 ALL, 11 AML) were used as training data. A set of 50 genes with the highest correlations with an "idealized expression pattern" vector, where the expression level is uniformly high for AML and uniformly low for ALL, were selected. The prediction of a new sample was based on "weighted votes" of these 50 genes. The method made strong prediction for 29 of the 34 test samples, and the accuracy was 100%. Golub's method is actually a minor variant of the maximum likelihood linear discriminate analysis for two