Microarray Data Mining

Li-Min Fu
Southern California University of Health Sciences, USA

INTRODUCTION

Based on the concept of simultaneously studying the expression of a large number of genes, a DNA microarray is a chip on which numerous probes are placed for hybridization with a tissue sample. Biological complexity encoded by a deluge of microarray data is being translated into all sorts of computational, statistical or mathematical problems bearing on biological issues ranging from genetic control to signal transduction to metabolism. Microarray data mining is aimed to identify biologically significant genes and find patterns that reveal molecular network dynamics for reconstruction of genetic regulatory networks and pertinent metabolic pathways.

BACKGROUND

The idea of microarray-based assays seemed to emerge as early as of the 1980s (Ekins & Chu, 1999). In that period, a computer-based scanning and image-processing system was developed to quantify the expression level in tissue samples of each cloned complementary DNA sequence spotted in a 2D array on strips of nitrocellulose, which could be the first prototype of the DNA microarray. The microarray-based gene expression technology was actively pursued in the mid-1990s (Schena et al., 1998) and has seen rapid growth since then (Bier et al., 2008).

Microarray technology has catalyzed the development of the field known as functional genomics by offering high-throughput analysis of the functions of genes on a genomic scale (Schena et al., 1998). There are many important applications of this technology, including elucidation of the genetic basis for health and disease, discovery of biomarkers of therapeutic response, identification and validation of new molecular targets and modes of action, and so on. The accomplishment of decoding human genome sequence together with recent advances in the biochip technology has ushered in genomics-based medical therapeutics, diagnostics, and prognostics.

MAIN THRUST

The laboratory information management system (LIMS) keeps track of and manages data produced from each step in a microarray experiment, such as hybridization, scanning, and image processing. As microarray experiments generate a vast amount of data, the efficient storage and use of the data require a database management system. While some databases are designed to be data archives only, other databases such as ArrayDB (Ermolaeva et al., 1998) and Argus (Comander, Weber, Gimbrone, & Garcia-Cardena, 2001) allow information storage, query and retrieval as well as data processing, analysis and visualization. These databases also provide a means to link microarray data to other bioinformatics databases (e.g., NCBI Entrez systems, Unigene, KEGG, OMIM). The integration with external information is instrumental to the interpretation of patterns recognized in the gene-expression data. To facilitate the development of microarray databases and analysis tools, there is a need to establish a standard for recording and reporting microarray gene expression data. The MIAME (Minimum Information about Microarray Experiments) standard includes a description of experimental design, array design, samples, hybridization, measurements and normalization controls (Brazma et al., 2001).

Data Mining Objectives

Data mining addresses the question of how to discover a gold mine from historical or experimental data, particularly in a large database. The goal of data mining and knowledge discovery algorithms is to extract implicit, previously unknown and nontrivial patterns, regularities, or knowledge from large data sets that can be used to improve strategic planning and decision-mak-

ing. The discovered knowledge capturing the relations among the variables of interest can be formulated as a function for making prediction and classification or as a model for understanding the problem in a given domain. In the context of microarray data, the objectives are identifying significant genes and finding gene expression patterns associated with known or unknown categories. Microarray data mining is an important topic in bioinformatics, a field dealing with information processing on biological data, particularly, genomic data.

Practical Factors Prior to Data Mining

Some practical factors should be taken into account prior to microarray data mining. At first, microarray data produced by different platforms vary in their formats and may need to be processed differently. For example, one type of microarray with cDNA as probes produces ratio data from two channel outputs whereas another type of microarray using oligonucleotide probes generates non-ratio data from a single channel. Also, different platforms may pick up gene expression activity with different levels of sensitivity and specificity. Moreover, different data processing techniques may be required for different data formats.

Normalizing data to allow direct array-to-array comparison is a critical issue in array data analysis since there are several variables in microarray experiments that can affect measured mRNA levels (Schadt, Li, Ellis, & Wong, 2001; Yang et al., 2002). Variations may occur during sample handling, slide preparation, hybridization, or image analysis. Normalization is essential for correct microarray data interpretation. In simple ways, data can be normalized by dividing or subtracting expression values by a representative value (e.g., mean or median in an array) or by taking a linear transformation to zero mean and unit variance. As an example, data normalization in the case of cDNA arrays may proceed as follows. The local background intensity is first subtracted from the value of each spot on the array; and the two channels are normalized against the median values on that array; and then the Cy5/Cy3 fluorescence ratios and $\log_{10}$-transformed ratios are calculated from the normalized values. In addition, genes that do not change significantly can be removed through a filter in a process called data filtration.

While data analysis is a central issue in data mining, experimental design is critical as well. In particular, the use of replication in controlled experiments can significantly improve the outcome (Lee, Kuo, Whitmore, & Sklar, 2000).

Differential Gene Expression

To identify genes differentially expressed across two conditions is one of the most important issues in microarray data mining. In cancer research, for example, we wish to understand what genes are abnormally expressed in a certain type of cancer, so we conduct a microarray experiment and collect the gene expression profiles of normal and cancer tissues, respectively, as the control and test samples. The information regarding differential expression is derived from comparing the test against control sample.

To determine which genes are differentially expressed, a common approach is based on fold-change. In this approach, we simply decide a fold-change threshold (e.g., 2X) and select genes associated with changes greater than that threshold. If a cDNA microarray is used, the ratio of the test over control expression in a single array can be converted easily to fold change in both cases of up-regulation (induction) and down-regulation (suppression). For oligonucleotide chips, fold-change is computed from two arrays, one for test and the other for control sample. In this case, if multiple samples in each condition are available, the statistical t-test or Wilcoxon tests can be applied but the catch is that the Bonferroni adjustment to the level of significance on hypothesis testing would be necessary to account for the presence of multiple genes. The t-test determines the difference in mean expression values between two conditions and identifies genes with significant difference. The non-parametric Wilcoxon test is a good alternative in the case of non-Gaussian data distribution. SAM (Significance Analysis of Microarrays) (Tusher, Tibshirani, & Chu, 2001) is a state-of-the-art technique based on balanced perturbation of repeated measurements and minimization of the false discovery rate (FDR). FDR is the expected proportion of false positives among all declared positives. FDR, as an alternative to the *p*-value, has been widely accepted for gene selection from microarray data (Yang & Yang, 2006). In addition, multivariate statistical analysis techniques, such as singular value decomposition (Alter, Brown, & Botstein, 2000) and multi-dimensional scaling, can be applied to reduce the high dimensionality of microarray data.