CS 2427 - Algorithms in Molecular Biology

Lecture #13 : 1 March 2006
Presenter: Dr. Tim Hughes

Scribe Notes by: Pingzhao Hu

1. Introduction to Microarrays

Gene expression profiling using nucleic acid hybridization-based methods has become popular in medical and biological research and development. These methods use high-density Microarrays to allow the researcher to screen for the expression of thousands of genes simultaneously in a single experiment. Basically, these arrays can be mainly divided into two categories: One is Affymetrix GeneChip Microarray, also called one-color microarray and another is cDNA Microarray, also called two-color Microarray.

Affymetrix GeneChip Microarray uses Oligonucleotides of length 25 base pairs. Typically, a mRNA molecule of interest (gene) is represented by a probe set composed of 11-20 probe pairs of these oligonucleotides. Each probe pair is composed of a perfect match (PM) probe (a section of the mRNA molecule of interest), and a mismatch (MM) probe. The difference between PM and MM is that MM has changed the middle (13th) base of the PM for measuring non-specific binding. There probe pairs are usually called a probe set as follows.

Perfect match	**AGGCTATCGCACTCCAGTGG**
Mismatch	**AGGCTATCGTACTCCAGTGG**

Since the Oligos is short, a gene on the array can be included many probe sets. RNA samples are prepared, labeled and hybridized with the arrays.

cDNA Microarray contains probe from a known gene on each spot. There probes on the array are longer pieces of DNA that are complementary to the genes in study. Usually, cDNA probes for making the array can be produced from a commercially available cDNA library so that a closer representation of the entire genome of an organism can be made on the array. It is also be possible to use PCR to amplify specific genes from genomic DNA to generate the cDNA probes. The produced cDNA probes then can be mechanically spotted onto a glass slide. Since cDNA probs are much longer than oligos in Affymetrix GeneChip, a probe is almost identical to a gene in a successful hybridization with a clone in the cDNA Microarrays. Usually, two samples, experimental sample and control sample (also called baseline sample), are prepared for being hybridized to the arrays. The control sample can be labeled with a green-fluorescing dye called Cy3 and the experimental sample labeled with a red-fluorescing dye called Cy5. If there is more of an mRNA transcript in the control sample than in the experimental sample, more Cy3 will bind to the probe on the array and the spot will fluoresce green.

Otherwise, the spot fluoresce red. In many cases, the two samples have the same amount of transcript. Therefore, the spot will fluoresce yellow.

Besides these two types of arrays, other arrays include Ink-jet array and mastless array, but they are not widely used.

Though there are many different types of Microarrays as discussed above, the analysis for the data obtained from these Microarrays are mainly divided into two levels: One is low level analysis, also called preprocessing, which is focusing on summarizing and normalized data; another is high-level analysis, which is focused on mining the summarized and normalized data. The detail of these two analyses is discussed as follows.

2. Preprocessing (Low level analysis) microarray expression data

Preprocessing microarray gene expression data includes many issues. For different types of microarray, the preprocessing methods are also very different. Here we focus on two issues: one is summarization and another is normalization. Summarization is to get an expression value for probe (spot) with two colors on cDNA Microarray and probe set with 11-20 probe pairs on Affymetrix Microarray. Normalization is to make all data directly comparable.

For the cDNA Microarray data, the summarization and normalization are two separately steps. Usually, we use log2(Cy3/Cy5) to get an expression value for each spot. There are many normalization methods for this type of Microarray, such as variance stabilizing normalization (vsn). For the Affymetrix Microarray data, the summarization and normalization are usually a unified step. Some methods for this purpose include MAS5 developed by Affymetrix, which use PM-MM to adjust for non-specific-binding and background noise; model-based method developed by Li and Wang (2001), which use PM-MM values, a non-linear normalization and takes multi-array summaries into account for detecting and removal outlier.

3. High level analysis of multiple experiments

There are different goals to analyze multiple microarray experiments. In the course, the following goals were discussed:

1. Identify differentially expressed genes by comparing experiments from different biological conditions
2. Clustering experiments and transcripts
3. Predicting gene functions from microarray experiments.

3.1 Identifying differentially expressed genes

To identify differentially expressed genes is to compare the expression levels of genes in samples obtained from different biological conditions, such as people with cancers and people without cancers. One of useful tools to aid for this purpose is to draw the scatter plot in which x-axis is the expression values of one condition (say, cancers), and y-axis is the expression values of another condition (say, without cancers). The genes that are highly expressed in one condition and lowly expressed in another condition will be most interesting. Fig.1. (modified based on Prof. Hughes's slide) illustrates a similar example, where each condition (Cup5 and Vma8, respectively. Note: WT is baseline) has only one array. As we can see, gene A is highly expressed in Cup5, but lowly expressed in Vma8; while gene B is highly expressed Vma8 but lowly expressed in Cup5. For most other genes, they either highly or lowly expressed in both conditions. Therefore, genes A and B are the most interesting genes.

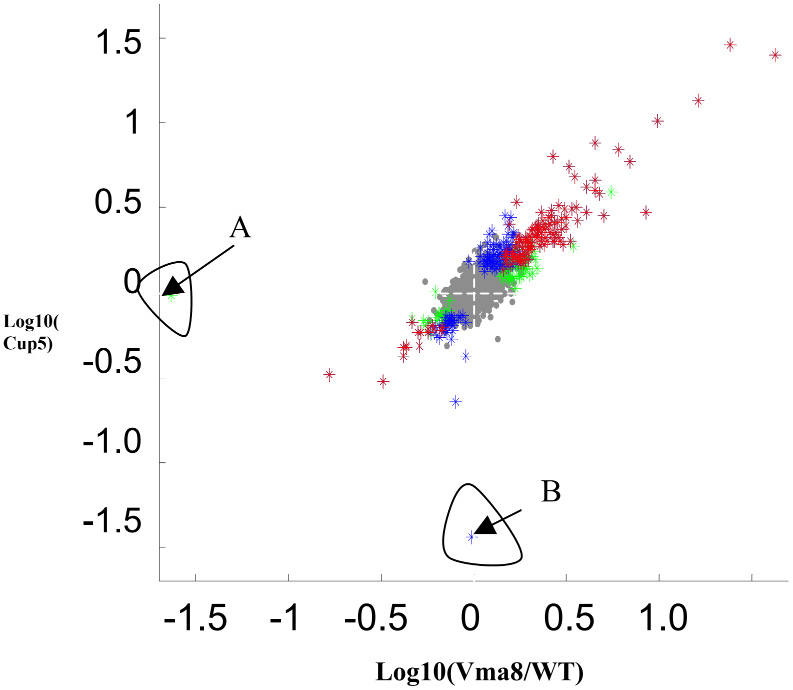

Figure 1 Scatter plot of expression values in two conditions

The drawback of this method is that it cannot tell how significant the set of selected genes are. A more precise method to make the conclusion may be obtained by calculating p-value for each gene using a supervised statistical test, such as t-test, or a model-based method, such as mixture model. For example, if we have more microarrays in the above experiments, say we have n_{Cup5}, n_{Vma8} arrays in the group Cup5 and Vma8, respectively, a t-test to evaluate the significance of each gene, say gene A, in this experiment will be

$$t_{geneA} = \frac{\overline{X}^{geneA}_{Cup5} - \overline{X}^{geneA}_{Vma8}}{\sqrt{\frac{Var^{geneA}_{Cup5}}{n_{Cup5}} + \frac{Var^{geneA}_{Vma8}}{n_{Vma8}}}},$$

where $\overline{X}^{geneA}_{Cup5}$, $\overline{X}^{geneA}_{Vma8}$, Var^{geneA}_{Cup5}, Var^{geneA}_{Vma8} are the means and variances of expression values of gene A in the two conditions. p-value can be calculated based on the t-statistic value, t_{geneA}.

Since a typical microarray dataset includes thousands of genes, an immediate concern is multiple testing, which means that when many hypothesis (here we have thousands of hypothesis) are tested, the probability that the number of false positives (genes that are found to be statistically different between two conditions, but are not in reality) will be increase sharply with the number of hypothesis. In order to control the false positives at an acceptable level, say 0.05 (5 of 100 positives are false), lots of methods have been developed to adjust the calculated p-values (Dudoit et al. 2003).

3.2 Clustering experiment conditions and gene expression

Clustering is an unsupervised learning method. The objective is to find clusters of samples, such as different disease groups, or clusters of genes, such as genes with the same biological functions in a cluster. If the objective is to cluster experiment conditions or experiment conditions and genes simultaneously (Figure 2), a pre-processing step to filter out genes that have no variations among conditions is needed. This requirement is also due to the fact that some clustering methods require that there be more samples (arrays here) than variables (here genes), otherwise, the clustering methods will fail. It should be noted that the filtering methods should be in an unsupervised way. Otherwise, the clustering analysis cannot find any useful information except the prior information used in supervised analysis. The common ways to filter genes are (1) coefficient of variation (CV); (2) the threshold of expression values. If the objective is to cluster genes, the filtering process is an optional choice. In either case, some other key issues in clustering analysis are:

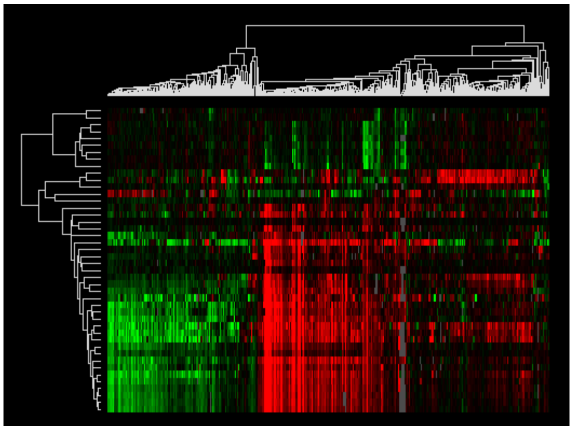

Figure 2 Hierarchical clustering of genes and samples

The first issue is to choose a suitable similarity measure. It is widely known that similarity measure significantly affects the final results. There are two common ways to calculate similarity: One is based on distance; another is based on correlation. The decision about which measure should be used depends on the biological questions we are interested in. For example, Pearson correlation may be more useful to identify over- and under-expressed genes.

The second issue is that although there are many clustering methods available for microarray analysis, such as hierarchical clustering, k-means clustering, self-organization maps (SOM) and principal component analysis, the commonly used one by biologists are hierarchical clustering. However, we usually have no statistical test to guide the decision of where to cut the dendrogram of hierarchical clustering results. In other words, deciding the number of clusters in a mciroarray data set is a challenge. The external criteria are based on the background knowledge of biological experiments.

The third issue is how to interpret the clustering results. It is out of doubt that any clustering algorithms can identify clusters from a given dataset, but we usually have no testing method to convince us whether the clusters are natural/biological clusters or just the artifacts of the algorithm.

3.3 Predicting gene functions

General speaking, both unsupervised and supervised methods can be used to predict gene functions based on microarray expression data. The rationale of using clustering methods for gene function prediction is that co-expressed genes in a cluster are expected to play the same or similar functions in a biological process/biological pathway. The basic idea is that (1) clustering genes; (2) annotating the functions of genes in clusters based on public biological databases, such as gene ontology (GO).

The assignment of gene to a cluster can be somewhat arbitrary and is generally quite sensitive to the choice of clustering algorithm, the parameter values used and the robustness of the experimental data. Moreover, most current incarnations of unsupervised algorithms generally cannot assign a gene to more than one ‘dominant’ functional class (clusters), in other words, the genes in different clusters are non-overlapped, which likely does not reflect the pleitrophic, and dynamic functional interactions. In this respect, supervised learning methods, such as support vector machines (SVMs) (Zhang et al. 2004), are likely to be far more effective at devising biologically inclusive multi-function prediction models since some prior information from annotation databases can be used in the model construction.

Using supervised method for gene function prediction, one must first pre-define the collection of ‘functions’ one aims to consider. The definition can be based on external annotation databases, such as GO. GO is a database of controlled vocabulary gene annotations describing the biological processes, molecular functions and cellular localizations of genes. Function is defined within the GO framework using a series of inter-linked, multi-tiered, hierarchically ordered ontological terms assigned to curated gene products, as defined by human experts. The functions defined on the top level of the hierarchy are much broader than those defined at the bottom level. In the function prediction, we are more interested in predicting functions at the middle or low level of the hierarchy, since the functions at those levels are more biological-specific. There are also other ways to pre-define gene functions. For example, we can define it based on either Enzyme Commission (EC) convention, a hierarchical classification nomenclature established to define the six principal classes of enzymes, or Munich Information Center for Protein Sequences (MIPS) reference database, which uses an extended EC-like methodology to capture and represent the available information concerning the subunit composition of well-studied protein complexes and biochemical pathways for several representative model organisms.

A recent study (Zhang et al. 2004) explored to use mouse gene expression (mRNA levels) for function prediction based on pre-defined functions in GO. They collected 992 GO biological process categories for 7779 genes, in which each GO category has at least 3 genes and less than 500 genes. They demonstrated that their SVM algorithm could assign lots of uncharacterized genes to these gene function categories with at least 50% precisions. However, it is not clear how many other functional gene classes can be recognized from mRNA expression data by this (or any other) method. Other functional classes may require different mRNA expression experiments, or may not be recognizable at all from mRNA expression data alone since some genes are regulated at the translation and protein levels. Integrating other data, such as the presence of transcription factor binding sites in the promoter region or sequence features of the protein with the expression data can improve gene function prediction. Other challenges in using this dataset for function prediction includes: (1) set the multi-label issue (a gene can be annotated into more than one function categories) in the standard support vector machine algorithm, which was developed for a binary classification problem; (2) unbalanced classification. As we can see, the minimum number of genes (positives) in a function category is 3 while they had the total number of 7779 characterized genes in the dataset. It is obviously a big challenge to train such an unbalanced dataset for function prediction.

Reference

Dudoit S, Shaffer JP, Boldrick JC. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 8:71-103.

Zhang W, Morris QD, Chang R, et al. (2004). The functional landscape of mouse gene expression. *Journal of Biology*, 3:21.