StatNews #58

A Short Introduction to Microarrays
May 2003
Revised 2012

Breakthroughs in technology have transformed molecular biology into a field with abundant data. Microarray technology has become a powerful genetic research method, allowing for massive screening of information in an intensive manner. Similar to the invention of the microscope, microarrays promise remarkable discoveries in biology. The purpose of this newsletter is to give a brief introduction to microarrays and to statistical issues associated with microarray data.

Microarrays allow studying expression (i.e., activity) of thousands of genes simultaneously. With many sequences of genes already known (e.g., yeast, many bacteria, and human), an important task now is assigning biological functions to genes and understanding the functional relationships among genes. There are many applications for microarray information, such as identifying disease genes, classification of tumors, or monitoring the effect of drugs. Types of microarrays include cDNA (or spotted) and Affymetrix.

A cDNA array is a glass slide on which DNA with known sequences representing different genes are spotted. In principle, such a slide can contain all known genes expressed in a given organism. Consider an experiment comparing cancerous vs. normal cells. In this experiment, the desired endpoint is to determine which genes are activated or repressed in the cancerous sample as opposed to the normal sample. The sample cells are labeled with fluorescent dyes; for example, red dye is used for the tumor sample and green for the normal. These two differentially labeled samples are combined and hybridized (meaning they will bind with complementary DNA on the slide) to the cDNA array. Suppose a gene that is present on the array is expressed in the tumor sample. That spot on the array will then bind to the corresponding red-labeled cDNA in the tumor sample. If that gene is also present in the normal sample, then that same spot will also bind to the corresponding green-labeled cDNA in the normal sample. The ratio of expression (red/green) indicates whether or not that spot is more highly expressed in the tumor sample (more red than green), repressed in the tumor sample (more green than red), or unchanged (equal red or green signal). The ratios of the red and green intensities across all of the spots on the slide represent the data for the microarray data analysis.

As massive amounts of data are collected, statistical issues emerge in image processing, design of experiments, and grouping and classification of genes. Any microarray study involves a number of steps, including: study design, data pre-processing, and data analysis.

The study design will help answer accurately the biological question, through randomization and replication, by deciding how many arrays will be used, what cell samples will be put on what arrays and with what dye assignment, sample preparation, etc. Several experimental designs have been

proposed for microarray experiments, including loop designs, dye-swap and single or multifactor designs.

The data pre-processing step is required to ensure that efficient biological comparisons can be made. This step removes variation due to technical issues such as printing, scanning, hybridization, or different labeling efficiencies of the dyes. One main purpose of microarray data analysis can be seen as a statistical multiple hypothesis-testing problem: the simultaneous testing of the null hypothesis of no differential expression for each gene. Finding powerful testing procedures, while controlling for family-wise error rate, is of special importance.

Two other purposes are classification (i.e., grouping of genes with similar patterns of expression) and prediction of biological outcomes such as tumor class or response to treatment. Clustering is used to identify clusters of genes when no pre-existing categories exist. When the purpose is to assign genes to pre-determined categories, discriminant analysis is more appropriate. Clustering methods commonly used include hierarchical clustering, k-means clustering and self-organized maps. Prediction methods include nearest neighbor classifiers, linear discriminant analysis, classification trees, and neural networks.

A good computing environment is essential in microarray data analysis. Statistical procedures for microarray data analysis are becoming available in mainstream statistical software. SAS Microarray Solution and S-PLUS Array Analyzer include methods for microarray data analysis. SAM (Significance Analysis of Microarrays) developed at Stanford University, available at no charge for academic users, is a versatile method of finding significant genes, but it does not have data normalization capabilities. A more detailed list of software packages can be found at http://ihome.cuhk.edu.hk/~b400559/arraysoft.html[1].

Please contact Simona Despa in the Office of Statistical Consulting if you would like to discuss analysis of microarray data.

Author: Simona Despa

Back to StatNews Table of Contents

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).

[1] This webpage was expired. 2012, May