DOCTORAL THESIS

Statistical analysis of gene expression data in cDNA microarray experiments

Zhao, Hongya

Date of Award:
2006

Statistical Analysis of Gene Expression Data in cDNA Microarray Experiments

ZHAO Hongya

A thesis submitted in partial fulfillment of the requirments

for the degree of

Doctor of Philosophy

Principal Supervisor: Prof. FANG Kai Tai

Hong Kong Baptist University

January 2006

Abstract

DNA microarray offers a powerful and cost effective approach to monitoring changes in gene expression levels for thousands of genes simultaneously instead of "one gene". In fact the way microarray technology is revolutionizing the biological science. However, the data alone does not constitute knowledge. It must be first analyzed, association studied and results confirmed in order to convert it into knowledge. Therefore there is a large and rapidly increasing literature on microarray data analysis.

In this thesis, the statistical and computational methods are focused. Chapter 1 is a brief introduction of molecular biology and microarray technology. In Chapter 2, the visualization tools are employed to gain insight into microarray data. It is realized that some traditional analytical approaches seem improper to apply because of the characteristics inherent to microarray data. And some typical preprocessing of data is also discussed. Chapter 3 and 4 describe the Bayesian models to analyze microarray data. According to the characteristics of data, Bayesian hierarchical models appear to be suited for the analysis because it can accommodate the complicated error structures, borrow strength across genes, and deduce the dimensionality of inference. In Chapter 3, we improve the LNN and GG hierarchical models to LNNG, considering the dependence between mean and variance of gene expression. Besides the univariate model in Chapter 3, the multivariate Bayesian model is also developed to microarray analysis in Chapter 4 because we discovered that there are strong relations between the measurements within one gene spot.

Some multivariate methods, such as outlier detection and clustering, are also applied in microarray analysis in Chapter 5. Considering the replicated microarray data as multivariate, differentially expressed genes can be identified as outliers with the robust multivariate algorithm. Similarly, the identification can be made in cluster setting because of the complication of gene expression patterns. In the last chapter, a novel three-color cDNA microarray experiment is made in our biological laboratory to assess drug effect on target disease. With the experimental data, a graphical tool, hexaMplot, is first proposed to demonstrate the relations among the expression of normal, disease, and drug samples. Based on the hexaMplot, the hypothesis testing of correlation coefficient provides a reasonable method for the evaluation of drug effect. The Bayesian models in the thesis can be extended to analyze three-color microarray data.

The microarray analysis covers a very broad rang of research. Besides the statistical topics in the thesis, there is a great deal of literatures in the new methodologies and computational advances. All in all the study of microarray is only the beginning in the new period of genome, but it is thriving and growing at a remarkable pace. As the technologies in large-scale and high throughput continue to evolve, the fresh challenges will emerge for the statistical analysis.

Key words: DNA Microarray, gene expression, visualization of data, differentially expressed genes, empirical Bayesian, conjugate prior, hierarchical mixture model, false discovery rate, multivariate Byesian model, generalized likelihood ratio test, outlier, projection pursuit, kurtosis, three-color cDNA microarray, hexaMplot.

Contents