

## DOCTORAL THESIS

### Statistical analysis of gene expression data in cDNA microarray experiments

Zhao, Hongya

*Date of Award:*  
2006

[Link to publication](#)

#### General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

# Statistical Analysis of Gene Expression Data in cDNA Microarray Experiments

**ZHAO Hongya**

A thesis submitted in partial fulfillment of the requirements  
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 range 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 Bayesian model, generalized likelihood ratio test, outlier, projection pursuit, kurtosis, three-color cDNA microarray, hexaMplot.

# Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Table of Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>xii</b>
<b>Chapter 1 Introduction to DNA Microarrays</b>	<b>1</b>
1.1 Genetic Background . . . . .	2
1.1.1 DNA, RNA, and Protein . . . . .	2
1.1.2 Gene Expression . . . . .	4
1.2 DNA Microarray Technology . . . . .	6
1.2.1 cDNA Microarray . . . . .	6
1.2.2 Olithogonome Microarray . . . . .	9
1.3 Applications and Challenges . . . . .	10
<b>Chapter 2 Exploratory Analysis of Microarray Data</b>	<b>16</b>
2.1 Visualization of Data . . . . .	17
2.1.1 AM Plot . . . . .	18
2.1.2 Box Plot . . . . .	18
2.1.3 Contour Plot . . . . .	21
2.1.4 QQ Probability Plots . . . . .	23

2.2	Pre-processing of Data . . . . .	24
2.2.1	Transformation . . . . .	25
2.2.2	Normalization . . . . .	30
<b>Chapter 3 Empirical Bayesian in Microarray Analysis</b>		<b>36</b>
3.1	A Brief Introduction to Bayesian Analysis . . . . .	37
3.1.1	Bayes' Theorem . . . . .	37
3.1.2	Prior Distributions . . . . .	40
3.1.3	Empirical Bayes Approach . . . . .	46
3.2	Hierarchical Mixture Model . . . . .	49
3.2.1	General Framework: Two Conditions . . . . .	49
3.2.2	Multiple Conditions . . . . .	52
3.2.3	False Discovery Rate . . . . .	53
3.3	Parametric EB Models . . . . .	55
3.3.1	Gamma-Gamma and Lognormal-Normal Models . . . . .	56
3.3.2	Lognormal-Normal-Gamma model . . . . .	57
3.3.3	Model Fitting . . . . .	61
3.4	Simulation Study . . . . .	61
3.4.1	Simulation of Data . . . . .	61
3.4.2	Data Analysis . . . . .	62
3.5	Case Study . . . . .	64
3.6	Conclusion . . . . .	67
<b>Chapter 4 Multivariate Bayesian in Microarray Analysis</b>		<b>69</b>
4.1	Multivariate Hierarchical Model . . . . .	70
4.2	Inference . . . . .	75
4.2.1	Generalized Likelihood Ratio . . . . .	76
4.2.2	P-value Adjustment . . . . .	77
4.3	Simulation Study . . . . .	79
4.3.1	Simulation of Data . . . . .	79
4.3.2	Data Analysis . . . . .	80
4.3.3	Effect of Covariance . . . . .	84

4.3.4	Effect of Sample Sizes . . . . .	85
4.4	Case Study . . . . .	87
4.5	Conclusion . . . . .	89
<b>Chapter 5 Identification with Multivariate Data Analysis</b>		<b>90</b>
5.1	Identification with Outlier Detection . . . . .	90
5.1.1	Multivariate Outlier Analysis . . . . .	91
5.1.2	Case Study . . . . .	97
5.1.3	Comparison . . . . .	99
5.1.4	Discussion . . . . .	102
5.2	Identification in Multiple-cluster Setting . . . . .	104
5.2.1	Cluster Techniques in Microarray Analysis . . . . .	105
5.2.2	Robust Gene Clustering . . . . .	111
5.2.3	Case study . . . . .	119
5.2.4	Discussion . . . . .	123
<b>Chapter 6 Three-color cDNA Microarray to Assess Drug Effect</b>		<b>124</b>
6.1	Application to Disease and Drug . . . . .	125
6.2	Three-color cDNA Microarray Technology . . . . .	128
6.3	Statistical Analysis . . . . .	131
6.3.1	Visualization Tool: HexaMplot . . . . .	131
6.3.2	Hypothesis Test of Correlation Coefficient . . . . .	138
6.4	Further Study . . . . .	141
<b>Bibliography</b>		<b>146</b>
<b>Curriculum Vitae</b>		<b>157</b>