Microarray-based gene set analysis in cancer studies

Qin (Sarah) Song

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Statistics,
The University of Auckland, 2008.

Abstract

This work addresses the development and application of gene set analysis methods to problems in microarray-based data sets. The work consists of three parts. In the first part a gene set analysis method (PCOT2) is developed. It utilizes inter-gene correlation to detect significant alteration in gene sets across experimental conditions. The second part is focused on the exploration of correlation-based gene sets in conjunction with the application of the PCOT2 testing method in the investigation of biological mechanisms underlying breast cancer recurrence. In the third part, statistical models for analyzing combined microarray-based expression and genomic copy number data are developed. In addition, an analysis which incorporates tumour subgroups is shown to provide more accurate prognosis assessment for breast cancer patients.

The main results are: (i) PCOT2 is an effective approach for incorporating inter-gene correlation for identifying differentially expressed gene sets; (ii) visualization tools provide insight into the structure of gene sets; (iii) gene functions including "cell cycle", "cell proliferation", "immune response" and "phosphate transport" were detected as significantly altered functions that are strongly associated with tumour recurrence in breast cancer; (iv) amplification regions are associated with patients' recurrence free survival times within subtypes of breast cancer.

Acknowledgments

My studies in New Zealand have been a wonderful and enriching experience. I would like to take this opportunity to thank people who provided help and advice throughout the course of the work.

First and foremost, my deepest appreciation to my supervisor, Dr Michael Alan Black, for his encouragement and guidance in the past three years. Mik, it has been a great pleasure to work with you and benefit from your knowledge, energy, enthusiasm and kindness. Thank you. Also thanks to Dr Sharon Browning and Dr Yong Wang for their kind support.

I would like to thank those collaborators with whom I have interacted, that is, Professor Anthony Reeve, Dr Ahmad Anjomshoaa, Dr Yu-Hsin Lin, Hui-Ming Lin, Debbie Leader, Sarah Glyn-Jones, Dr Dorit Naot. Also thanks go to Les McNoe and Richard Gyde for proof-reading this thesis, and to Dr Christian Röver and Dr Richard Umstätter for the discussions about latex formatting.

I have appreciated every staff member of the Statistics Department, University of Auckland, and the Cancer Genetics Laboratory, Department of Biochemistry, University of Otago, for their kindness and support.

This work was funded by a New Zealand International Postgraduate Research Scholarship. Also thanks to the R/Bioconductor community for providing much of the software used in this work.

Finally special thanks to the many people not named here, who also had their share in making this journey possible and enjoyable.

Preface

High-throughput technologies for the measurement of gene expression provide powerful tools for investigating the transcriptome on a genomic scale across diverse biological conditions. To facilitate the interpretation of statistical results in terms of biological knowledge, incorporating gene functional information is becoming increasingly important when analyzing genome-wide expression data. This work focuses on the development and application of statistical tools for gene set analysis of expression microarrays. The exploration of microarray-based array CGH data will also be discussed. The goal is to identify gene sets that have an impact on clinical outcomes, and may ultimately be used for the identification of disease progression.

The content of this thesis is as follows:

Chapter 1 provides an overview of basic molecular biology and microarray technology. Many commonly used gene set (pathway) databases and software are summarized. A basic introduction to the molecular biology of cancer is also included. *Chapter 2* briefly reviews some current methods for both single gene analysis and gene set analysis. *Chapter 3* describes a new statistical method, PCOT2, used for the analysis of gene sets, that was developed as part of this thesis. A comparison of performance with existing gene set analysis methods are produced in both real and simulated data sets. *Chapter 4* develops a correlation-based method for defining gene sets in conjunction with the PCOT2 testing method, with the goal of investigating the biological mechanisms underlying the recurrence of breast cancer. *Chapter 5* explores recurrence free survival (RFS) using amplified chromosomal regions in conjunction with an assessment of the level of proliferation of each tumour. Additional investigation is performed in the analysis of subgroups of breast cancer patients based on intrinsic gene set signatures. An analysis

method which allows the integration of molecular, genetic and clinical data is presented in this chapter. *Chapter 6* provides a summary of the broad results and conclusions of this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

aCGH	array Comparative Genomic Hybridization
BC	Breast Cancer
cDNA	complementary Deoxyribose Nucleic Acid
CIS	Carcinoma *In Situ*
DNA	Deoxyribose Nucleic Acid
EB	Empirical Bayes
EM	Expectation-Maximization
ER	Estrogen Receptor
ES	Enrichment Score
GEO	Gene Expression Omnibus
GLM	Generalized Linear Model
GPS	Gene Proliferation Signature
GSEA	Gene Set Enrichment Analysis
GT	Global Test
HER2	Human Epidermal growth factor Receptor 2
HMM	Hidden Markov Model
LIMMA	Linear Models for Microarray Data
LN	Lymph Node
LOOCV	Leave-One-Out Cross-Validation
MAD	Median Absolute Deviation
MCP	Multiple Comparison Procedures
MDS	Multi-Dimensional Scaling
MGED	Microarray Gene Expression Data

MIAME	Minimum Information About a Microarray Experiment
mRNA	messenger Ribonucleic Acid
OA	Ontological Analysis
PCOT2	Principal Coordinates and Hotelling's T^2
PCR	Polymerase Chain Reaction
PE	Prediction Error
PgR/PR	Progesterone Receptor
PH	Proportional Hazard
RFS	Recurrence Free Survival
RMA	Robust Multi-Array
RNA	Ribonucleic Acid
SAFE	Significance Analysis of Function and Expression
SAGE	Serial Analysis of Gene Expression
SAM	Significance Analysis of Microarrays
SNR	Signal-to-Noise Ratio
SSP	Single Sample Predictor
SVD	Singular Value Decomposition