



<http://researchspace.auckland.ac.nz>

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form.

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	xv
List of Tables	xix
1 Introduction	1
1.1 DNA and the central dogma of molecular biology	1
1.2 Microarray technology	4
1.2.1 Spotted oligo microarrays	6
1.2.2 Affymetrix microarrays	7
1.2.3 Microarray-based comparative genomic hybridization .	8
1.3 Microarray resources	10
1.4 Pathway databases and software	12
1.5 Cancer genetics	15
1.5.1 Basic molecular biology of cancer	15
1.5.2 Microarray-based studies of cancer	16
1.6 Concluding Remarks	17
2 Statistical methods for microarrays	19
2.1 Overview	19
2.2 Preprocessing procedures	20
2.3 Single gene analysis methods	22
2.4 Gene set analysis	28
2.4.1 Overview	28
2.4.2 Ontological Analysis (OA)	29
2.4.3 Gene Set Enrichment Analysis	29
2.4.4 Significance analysis of function and expression (SAFE)	31

2.4.5	Global test (GT)	31
2.5	Concluding Remarks	33
3	Principal Coordinates and Hotelling's T^2 (PCOT2)	35
3.1	Overview	35
3.2	PCOT2 gene set analysis method	36
3.2.1	Overview	36
3.2.2	Methodology	37
3.2.3	The <code>pcot2</code> Bioconductor package	42
3.2.4	Evaluation	49
3.3	Comparison of PCOT2 with other gene set analysis methods	55
3.3.1	Measurements on gene sets	55
3.3.2	Evaluation methods	56
3.3.3	Results	59
3.4	Concluding Remarks	69
4	Correlation-based gene sets for quantifying the likelihood of breast cancer recurrence	71
4.1	Overview	71
4.1.1	Correlation-based gene set analysis	71
4.1.2	Breast Cancer	73
4.2	Materials and Methods	74
4.2.1	Patient demographics	74
4.2.2	Microarray profiling	75
4.2.3	Defining gene clusters with genes having similar patterns	76
4.2.4	Methods for choosing highly correlated genes	77
4.2.5	Determining cutoff value from re-sampling data sets	79
4.2.6	Choosing highly correlated genes and gene clusters	79
4.2.7	Testing highly correlated gene clusters	80
4.2.8	Annotating significant gene clusters	81
4.3	Results	82
4.3.1	Association between clinical information and prognosis	82
4.3.2	Pre-processing of gene expression data sets	83
4.3.3	Clustering based on the Rule One criterion	83
4.3.4	Clustering analysis based on Rule Two criterion	97

4.4	Concluding Remarks	100
5	Combining gene expression and genome copy number data to predict relapse-free survival time	103
5.1	Overview	103
5.2	Materials and Methods	105
5.2.1	Patients and Data sets	105
5.2.2	Analyzing array CGH data	106
5.2.3	Hidden Markov Models (HMM) for detecting genomic amplification events	107
5.2.4	Identify and visualize focal amplifications	108
5.2.5	Using aCGH and expression data to define gene set predictors	110
5.2.6	Survival analysis for predicting patients' recurrence survival time	111
5.2.7	Prediction model of gene amplification regions	114
5.2.8	Prediction models of subtypes of breast cancer tumours	117
5.3	Results	121
5.3.1	Analysis of aCGH data	121
5.3.2	Predict amplification status using expression data . . .	130
5.3.3	Survival analysis on four test data sets	137
5.3.4	Prediction model of subtypes of breast cancer patients	139
5.3.5	Survival analysis within breast cancer subtypes	145
5.4	Concluding Remarks	154
6	Conclusions and discussion	157
A	Computational considerations	163
B	Some other basic concepts	165
C	Chapter Three Appendix Files	167
C.1	Comparison of the performance of the GSEA-Category method before and after taking the absolute value of t statistic. [the file was saved in CD].	167

C.2	Visualization of gene expression profiles of four types of simulated gene pathways (gene set are changed in the same direction) [CD].	167
C.3	Application of gene set analysis methods to diabetes data (all gene sets) [CD].	167
C.4	Application of gene set analysis methods to leukemia data (all gene sets) [CD].	167
D	Chapter Four Appendix Files	169
D.1	Clinical characteristics of patients and their tumours [CD]. . .	169
D.2	The density plots for Loi <i>et al.</i> data [CD].	169
D.3	The density plots for three data sets [CD].	169
D.4	Summary plots of the maximum correlation distribution [CD].	169
D.5	Venn Diagram of highly correlated genes detected in the three data sets [CD].	169
D.6	Comparison plot of three linkage methods [CD].	170
D.7	A complete list of gene symbols in the reported clusters related to the analysis using 197 common genes [CD].	170
D.8	Difference plots for the detected clusters [CD].	170
D.9	The lists of all the gene symbols in the detected clusters [CD].	170
E	Chapter Five Appendix Files	173
E.1	Clinical information of the overlapped samples [CD].	173
E.2	Grid plots of the amplification status for all the clones on chromosome 8, 11, 17 and 20 with their related samples [CD].	173
E.3	A list of proliferation genes [CD].	173
E.4	A list of 120 Sorlie <i>et al.</i> genes [CD].	173
E.5	A list of 306 Hu <i>et al.</i> genes [CD].	174
E.6	Basic frequency plot of aCGH data [CD].	174
E.7	Grid plot of the amplification status for all the clones across all the samples [CD].	174
E.8	Names of 50 clones reported as amplified [CD].	174
E.9	Density plot of gene expression profiles between before and after removing three samples in Chin <i>et al.</i> data [CD].	174

E.10	The summary table and survival plot of four groups clustered using eight predictors [CD].	174
E.11	Density plot of Wang <i>et al.</i> expression data between the \log_2 and natural log transformations [CD].	175
E.12	Survival plots of four predictors in four data sets [CD].	175
E.13	The clinical information of patients within each cluster (sub-group) that was defined by using 120 intrinsic gene probes on Pawitan <i>et al.</i> data set [CD].	175
E.14	Results of using SSP for the prediction of subtypes based on two different types of correlation [CD].	175
E.15	Survival plots of four predictors using the cohort data set in five subgroups [CD].	175
E.16	Survival plots using subgroups of patients with amplification on V1.Ch8 [CD].	175
E.17	Survival plots using subgroups of patients with amplification on V1.Ch11 [CD].	176
E.18	Survival plots using subgroups of patients with amplification on V1.Ch20 [CD].	176
E.19	Survival plots using subgroups of patients with proliferation [CD].	176
E.20	Survival plots using subgroups of patients (the enlarged version) [CD].	176
E.21	Survival analysis of comparing treated and untreated patients in the combined data [CD].	176
E.22	Survival plots of using the sub-grouped untreated and treated patients separately [CD].	176
F	Publications, Presentations and Posters [CD]	177
	Bibliography	179

List of Figures

1.1	The diagram of the Central Dogma of molecular biology. . . .	3
1.2	Cancer development from normal cells to cancer.	15
3.1	Outline of the PCOT2 methodology for analyzing the diabetes data.	37
3.2	Visualization of a biological pathway using the <code>corplot</code> function in the <code>pcot2</code> package.	45
3.3	Visualization of both expression and correlation data for the OXPHOS_HG-U133A_probes pathway using the <code>pcot2</code> software package.	46
3.4	Using <code>gene.locator</code> to find genes from the correlation plot. .	47
3.5	Visualization of gene expression profiles of four types of simulated gene pathways (gene set are changed in both directions). .	49
3.6	Comparison of detection rates for different combinations of parameters used in the <code>pcot2</code> function in 100 simulated data sets with altered expression patterns, as a function of increasing inter-gene correlation structure.	52
3.7	Comparison of detection rates for different combinations of parameters used in the <code>pcot2</code> function in 100 simulated data sets with altered expression patterns, as a function of increasing inter-class separation	53
3.8	Comparison of detection rates of gene set analysis methods in 100 simulated gene sets with altered expression patterns, as a function of increasing correlation.	61

3.9	Comparison of detection rates of gene set analysis methods in 100 simulated gene sets with altered expression patterns, as a function of increasing inter-class separation.	62
3.10	Venn diagram of the results from the GSEA-Category, Globaltest and PCOT2 gene set analysis methods.	66
4.1	The expression plot generated by applying complete linkage clustering to tumour-relapse samples in the Pawitan <i>et al.</i> data.	86
4.2	The expression plot generated by applying average linkage clustering to tumour-relapse samples in the Pawitan <i>et al.</i> data.	87
4.3	The correlation plot generated by applying complete linkage clustering to tumour-relapse samples in the Pawitan <i>et al.</i> data.	91
4.4	The correlation plot generated by applying average linkage clustering to tumour-relapse samples in the Pawitan <i>et al.</i> data.	92
4.5	Difference plots for clusters 152, 645, 228 and 115 shown in Table 4.6	99
5.1	Average \log_2 -ratio between relapsed and non-relapsed samples on chromosome 8.	121
5.2	Grid plot of amplification on chromosome 8 across all samples.	122
5.3	Seven regions were detected as amplified on four chromosomes.	124
5.4	2D-clustering grid plot using proliferation and amplification variables.	126
5.5	Survival plots using the eight predictors in the Chin <i>et al.</i> data set.	127
5.6	Survival plots using variables selected based on both the AIC and BIC criteria.	128
5.7	Survival plot of four predictors in the cohort data.	137
5.8	Clustering patients using the 120 intrinsic probes in the Pawitan <i>et al.</i> data.	140
5.9	Survival plots of the use of tumour subgroups that are predicted by both the 120 intrinsic probes and SSP methods. . .	142
5.10	Survival plots using subgroups of patients in the five data sets.	143
5.11	Survival plots using subgroups of tumours with amplification on V1.ch8 in the cohort data.	148

5.12 Survival plots using tumour subgroups with amplification on V1.ch11 in the cohort data.	149
5.13 Survival plots using subgroups of tumours with amplification on V1.ch20 in the cohort data.	150
5.14 Survival plots using subgroups of tumours with proliferation in the cohort data.	151

List of Tables

1.1	A list of commonly used microarray databases and tools. . . .	11
1.2	A list of popularly used gene network (pathway) databases and software.	14
3.1	Summary of measurements used for each gene set method. . .	56
3.2	Detection rates for the five gene set analysis methods on simulated data (genes within each set are assumed having the same direction of regulation).	59
3.3	Detection rates for the five gene set analysis methods on simulated data (genes within each set are assumed having the different direction of regulation).	60
3.4	Detection rates for the five gene set analysis methods on simulated data based on the diabetes and leukemia data sets . . .	64
3.5	Application of gene set analysis methods to the diabetes data.	67
3.6	Application of gene set analysis methods to the Leukemia data.	68
4.1	Two procedures used for selecting highly correlated genes. . .	78
4.2	Summary of the maximum pair-wise correlation distribution from the re-sampling data sets.	83
4.3	Annotation of significantly changed clusters based on expression data.	89
4.4	Annotation of significantly changed clusters based on correlation data.	93
4.5	Summary of clustering analysis using Rule One criterion in three breast cancer data sets.	95
4.6	Summary of clustering analysis using Rule Two criterion in three breast cancer data sets.	98

5.1	Mapping information for the seven regions detected as amplified on chromosomes 8, 11, 17 and 20.	124
5.2	Deviance for each model using forward selection.	128
5.3	LOOCV output based on different selection methods and criteria.	131
5.4	Performance of the three rules based on outer LOOCV results for each variable.	132
5.5	Final models using expression probes to predict each amplification region.	134
5.6	Performance of models using K-means and model-based clustering for predicting each amplification region.	135
5.7	The summary table of the final prediction model used for each region.	136
5.8	Results of log rank tests in the cohort data set.	138
5.9	Number and percentage (in parentheses) of patients within each subtype predicted by the SSP on five breast cancer data sets.	141
5.10	Numbers of samples within each breast cancer subtype.	145
5.11	Log-rank p -values within each breast cancer subtype.	146
5.12	Log rank test p -values using subtypes within treated and untreated patients in the combined data.	153

List of Abbreviations

aCGH	array Comparative Genomic Hybridization
BC	Breast Cancer
cDNA	complementary Deoxyribose Nucleic Acid
CIS	Carcinoma <i>In Situ</i>
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