

# PROTOCOLS FOR DIMENSION REDUCTION OF TRANSCRIPTOMIC DATA

By

Tim Peters  
BSc (Bioinformatics)(Hons)

A THESIS SUBMITTED TO MACQUARIE UNIVERSITY  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY  
DEPARTMENT OF STATISTICS  
JULY 2012





# Contents

<b>Abstract</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xv</b>
<b>1 Biological Background and Introduction</b>	<b>1</b>
1.1 Lymphoma . . . . .	1
1.2 Gene Expression Profiling . . . . .	2
<b>2 The Evolution of Dimension Reduction Techniques</b>	<b>9</b>
2.1 Fold Change and $t$ -tests . . . . .	10
2.2 Principal Components of a High-Dimensional Dataset . . . . .	12
2.3 Learning Explicitly in the Feature Space . . . . .	13
2.3.1 Fisher's Linear Discriminant and Linear Discriminant Analysis .	14
2.3.2 Regularisation and the Support Vector Machine . . . . .	15
2.3.3 Non-linear Decision Boundaries and the Kernel Trick . . . . .	17
2.4 Architecture for Dimension Reduction Algorithms . . . . .	19
2.4.1 Filters, Wrappers and Embedded Algorithms . . . . .	21
2.4.2 Stochastic Methods . . . . .	23
2.5 Deterministic Feature Selection via Coefficient Shrinkage . . . . .	28

2.5.1	Least Angle Regression . . . . .	31
2.5.2	Elastic Net and Generalised Path Seeking . . . . .	31
2.6	Decision Trees . . . . .	32
2.7	Random Forests, Bootstrapping and Stability Selection . . . . .	34
2.8	Machine Learning and Classification . . . . .	36
2.8.1	Nearest-neighbour Classification . . . . .	38
2.8.2	Cross Validation . . . . .	40
2.8.3	Machine Learning ‘Hygiene’ . . . . .	41
2.9	Closing Remark . . . . .	42
<b>3</b>	<b>Two-Step Cross-Entropy Feature Selection for Microarrays — Power Through Complementarity</b>	<b>43</b>
3.1	Introduction . . . . .	44
	TWO-STEP CROSS-ENTROPY FEATURE SELECTION FOR MICROARRAYS —	
	POWER THROUGH COMPLEMENTARITY . . . . .	47
	Background . . . . .	47
	Data and Methodology . . . . .	51
	Results and Discussion . . . . .	55
	References . . . . .	57
<b>4</b>	<b>Cancer Microarray Feature Selection Using Support Vector Machines: Comparing Regularisation Techniques</b>	<b>61</b>
4.1	Introduction . . . . .	62
	CANCER MICROARRAY FEATURE SELECTION USING SUPPORT VECTOR	
	MACHINES: COMPARING REGULARISATION TECHNIQUES . . . . .	65
	Background . . . . .	66
	Data and Methodology . . . . .	69
	Results and Discussion . . . . .	72
	References . . . . .	78

---

<b>5</b>	<b>Relieving Feature Selection AECS and Pains; a Consensus Approach to Identifying Biomarkers</b>	<b>83</b>
	Background . . . . .	85
	Data and Methodology . . . . .	92
	Results and Discussion . . . . .	97
	References . . . . .	104
<b>6</b>	<b>Practical Considerations, Conclusions and Future Directions</b>	<b>113</b>
6.1	Class Imbalance . . . . .	113
6.2	Computational Costs . . . . .	115
6.3	What is the Optimal Number of Features a Dataset Should Be Reduced To? . . . . .	116
6.4	Differing Characteristics of Transcriptomic Datasets . . . . .	119
6.5	Future Directions and Conclusion . . . . .	122
<b>A</b>	<b>Appendix</b>	<b>125</b>
A.1	Cross-Entropy Code Example . . . . .	125
A.2	AECS Code Example . . . . .	128
	<b>List of Symbols</b>	<b>143</b>
	<b>References</b>	<b>145</b>



# Abstract

The elucidation of the complex aetiology of human disease has been accelerated by recent advances in systems biology. Generation of high-dimensional datasets through gene expression profiling is an inevitable component of this research. Bioinformaticians are presented with this unabridged data by scientists seeking biological insights. Their role is that of a software engineer as well as a scientist; they are needed to facilitate the analysis by building software that performs dimension reduction. The desired outcome of dimension reduction is to find a handful of genes whose expression values reliably diagnose unlabelled samples. This thesis discusses the issues faced in bioinformatic classification and feature selection, and culminates in the development of a protocol to generate groups of genes that illuminate the nature of human disease.

The literature review describes how microarray technology facilitates the analysis of gene expression profiling, and charts the journey from hybridization to a normalised dataset. It then follows the development of dimension reduction techniques over the last 20 to 30 years. Moving from early techniques, it covers the three main strains of data mining algorithms: Discriminant Analysis, Decision Trees and the shrinkage family.

This thesis contains three articles (two of which have been published), each describing a statistical concept in need of consideration each time a dimension reduction is performed. By way of example, supervised learning of the transcriptomes of lymphoma patients is carried out in each study. The first shows that the common practice of scoring features individually and ranking them by these scores is too superficial a

method of assessing their degrees of biological relevance. We show the need to assess the gestalt discriminatory power of feature sets, the implications of this power in algorithm design and optimisation, and the intuitive relationship of this concept to biological phenomena.

The second article describes the need for regularisation of the linear model. We discuss how the searches for a workable compromise between model bias and variance within each of the three main data-mining strains are performed in quite different ways, yet possess a common theoretical background and yield similar predictive results on validation procedures. We show that a simple forward selection technique that adds features to a model based on the maximisation of the penalised margin width of its regularised Support Vector Machine formation performs competitively against, and in some cases outperforms Random Forest and Least Angle Regression with respect to classifying unlabelled data points.

No feature selection technique has proven to be superior to all others. Since there currently seems to be no ‘silver bullet’ method for extracting the most telling biomarkers from a transcriptome, we develop and test a novel ensemble feature selection method in the third and most ambitious article. We rigorously build an inventory, through sound selection of both regularisation and penalty parameters, of all three major machine learning families, and construct a validation architecture that involves bootstrapping for stability purposes. Testing this selection suite across a range of high-dimensional datasets, including some publicly available ones, lends further weight to a broad range of previous statistical and biological findings. We find significant overlap between the features found using our method and the ones identified as putative biomarkers in the original studies accompanying the publicly available datasets, some whose implication in disease has more recently been further explicated. We also apply our method to an in-house lymphoma gene expression dataset, independently identifying features that are already identified in other studies as biomarkers for the disease subtype in question, as well as discovering some novel putative ones.

Supplementary material and appendices detail studies ancillary to the core of the thesis and may provide starting points for further research.



# Certificate of Originality

Except where otherwise indicated below or in the text herein, the work described in this thesis is entirely my own, and has not been submitted, in any form, for a higher degree at Macquarie University or any other institution.

In this thesis I have used gene expression profile data generously provided by the Blood Stem Cell and Cancer Research (BSCCR) unit, St. Vincent's Centre for Applied Medical Research, as part of their lymphoma project. The following list summarises my particular contribution to the joint papers in this thesis.

## **Chapter 3:**

**Tim Peters**, David Bulger, To-ha Loi, Jean Y. H. Yang, David Ma. *Two-Step Cross-Entropy Feature Selection for Microarrays — Power Through Complementarity*. IEEE/ACM Trans Comput Biol Bioinform. 2011 Jul-Aug; 8(4):1148-51. doi:10.1109/TCBB.2011.30

Conception: 5%, analysis: 100%, writing: 95%

## **Specific contribution of joint authors**

David Bulger: Conceived the hypothesis and methodology for this article. Provided continual feedback and suggestions on both analysis and writing.

To-ha Loi: Was central to generating the expression data from the microarrays in the laboratory and wrote sections in Materials and Methods pertaining to expression data generation and normalisation.

Jean Y. H. Yang: Normalised the expression data.

David Ma: Oversaw all activities of the BSCCR group.

#### **Chapter 4:**

**Tim Peters**, David Bulger, To-ha Loi, Jean Y. H. Yang, David Ma. *Cancer Microarray Feature Selection Using Support Vector Machines: Comparing Regularisation Techniques*. In JSM Proceedings, 2009. Section on Statistical Learning and Data Mining. Alexandria, VA: American Statistical Association. 2951-2965.

Conception: 95%, analysis: 100%, writing: 95%

#### **Specific contribution of joint authors**

David Bulger: Provided continual feedback and suggestions on both analysis and writing.

To-ha Loi: As per Chapter 3.

Jean Y. H. Yang: As per Chapter 3.

David Ma: As per Chapter 3.

#### **Chapter 5:**

**Tim Peters**, David Bulger, To-ha Loi, Jean Y. H. Yang, David Ma. *Relieving Feature Selection AECS and Pains; a Consensus Approach to Identifying Biomarkers*.

Conception: 95%, analysis: 100%, writing: 95%

#### **Specific contribution of joint authors**

David Bulger: As per Chapter 4.

To-ha Loi: As per Chapter 3.

Jean Y. H. Yang: As per Chapter 3.

David Ma: As per Chapter 3.

# Acknowledgements

First and foremost, I would like to thank my primary supervisor, Dr. David Bulger, for the tireless help, instruction and magnanimity that he has given me over the last 5 years. Soon after I received my bachelor's degree, in the process of prevaricating over whether to begin a doctoral thesis or not, an older colleague who had already received his doctorate offered his advice: the relationship that you have with your supervisor is even more important than whether the research you are undertaking interests you. This advice stuck with me, and, in addition to my interest in machine learning and statistics, was the governing factor in my decision to see this project through. I warmed to his dry wit, patience and unpretentious manner, and I have gained many skills and insights under his supervision.

Many thanks to Professor David Ma for his supervision, as well as to Dr. To-ha Loi and the rest of the group at the Blood Stem Cell and Cancer Research (BSCCR) unit, St. Vincent's Centre for Applied Medical Research, for originally offering me a research assistant position back in 2006, which was the gateway to this project. Participating in the research activity of their ambitious group has been an immensely rewarding experience.

I also extend sincere thanks to Dr. Jean Yang, Lecturer in the School of Mathematics and Statistics at the University of Sydney, for her guidance, support and pragmatism; the Department of Statistics at Macquarie University for allowing me to undertake a doctoral degree under their auspices, particularly Professors Graham Wood and Barry Quinn, Associate Professor Jun Ma, Sandra Ticehurst, and fellow

doctoral students Brad, Sheenal, Peter and Annette; Macquarie University Division of Economic and Financial Studies for providing a scholarship for me over the course of my tenure.

An immeasurable amount of appreciation goes to my parents for their love and provision throughout this process; I couldn't have done it without them. Much love and thanks to Angela and Tristan for their support as well.

Any errors or infelicities that remain are entirely my own.

# List of Figures

1.1	Subsection of a two-colour microarray heatmap. Each dot represents the expression value of a single gene, based on its intensity. Red dots represent up-regulated genes, and green dots down-regulated ones. Image obtained under Creative Commons licence from Department of Biology at James Madison University. . . . .	4
2.1	A support vector machine in $\mathbb{R}^2$ separating two (red and blue) classes. .	18
2.2	A more regularized support vector machine using the same dataset in Figure 2.1. . . . .	19
2.3	Comparison of two SVMs drawn on two identical datasets. . . . .	20
2.4	Paths of the top six feature weights in the probability vector $V$ in the first 250 iterations of a CE method run where $q = 6$ , $\alpha = 0.02$ and the top decile of feature sets are extracted to create $V^*$ . . . . .	27
2.5	A selection of estimation pictures for $ B_1 ^\gamma +  B_2 ^\gamma \leq q$ for $\gamma = \{2, 1, 0.5\}$ , $\gamma \rightarrow 0$ . . . . .	30
2.6	A simple decision tree distinguishing between Hodgkin's and Non-Hodgkin's Lymphoma. Terminal nodes are in red and decision nodes are in green.	33
2.7	A non-exhaustive dichotomous taxonomy of well-known two-dimensional shapes. . . . .	37
2.8	(a) A Receiver Operating Characteristic (ROC). (b) ROC with a dot representing the decision boundary. . . . .	39

3.1	Diagnostic tree for lymphoma or suspected lymphoma samples. . . . .	49
3.2	Scatter plots comparing CE performance rank with individual $F$ -value rank. . . . .	55
4.1	Receiver operating characteristics depicting the accuracy of each method on the optimized models from (a) the St. Vincent's training set and (b) St. Vincent's validation set. . . . .	73
4.2	ROCs for (a) Golub training set and (b) Golub validation set. . . . .	74
4.3	ROCs for (a) Armstrong training set and (b) Armstrong validation set. . . . .	75
4.4	ROC for Alizadeh training set. Note that this study used a customized array, and the samples have a much greater homogeneity, since they are diagnosed as subclassifications from a single form of lymphoma. . . . .	76
5.1	Hierarchical chart showing the biological classification of the response variables used in this study. . . . .	94
6.1	Plots of optimal value of $\lambda$ against the most regularised value of $\lambda$ conferring minimum training error along the same regularisation path, using SVM stepwise regression with simulated annealing, on three different datasets. . . . .	118
6.2	Explained risk (ER) for all datasets used in Chapter 5. Datasets with homogeneous samples tend to have a higher ER. . . . .	120
6.3	Proportions of bootstraps for which $\gamma$ is optimal for all data splits used in Chapter 5. 150 bootstraps were used. . . . .	121

# List of Tables

4.1	Attributes of the four datasets used in this study. . . . .	70
4.2	Classification accuracies for all models. The figure represents the highest accuracy achieved along each path for the regression algorithms, and the accuracy at 500 trees for Random Forest. Prediction is calculated from posterior probabilities for both stepwise regressions and Random Forest, and from signed responses for least-angle regression. . . . .	77
4.3	Area under receiver operating characteristics shown in Figure 4.1 . . .	77
4.4	$\ell^1$ -norm from decision boundary (coloured dots shown on ROC curves in Figures 4.1 - 4.4) to $(0, 1)$ . This represents the sum of the false positive and false negative rates. . . . .	78
5.1	Summary of the datasets used in this study. The structure of the response variable splits from the in-house dataset is described in Figure 5.1. . . . .	94
5.2	Fully dimensionally-reduced feature lists from each of the 10 data splits analysed, using AECS. Known links to genes from expressed sequence tags (ESTs) are given. Features are grouped according to the class in which they show a higher degree of expression. . . . .	100

