Methods for Incorporating Biological Information into the Statistical Analysis of Gene Expression Microarray Data

Debbie Leader

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics.

The University of Auckland New Zealand.

December 2009

Abstract

Microarray technology has made it possible for researchers to simultaneously measure the expression levels of tens of thousands of genes. It is believed that most human diseases and biological phenomena occur through the interaction of groups of genes that are functionally related. To investigate the feasibility of incorporating functional information and/or constraints (based on biological and technical needs) into the classification process two approaches were examined in this thesis.

The first of these approaches investigated the effect of incorporating a pre-filter into the gene selection step of the classifier construction process. Both simulated and real microarray datasets were used to assess the utility of this approach. The pre-filter was based on an early method for determining if a gene had undergone a biologically relevant level of differential expression between two classes. The genes retained by the pre-filter were ranked using one of five standard statistical ranking methods and the most highly ranked were used to construct a predictive classifier. To generate the simulated data a selection of different parametric and non-parametric techniques were employed. The results from these analyses showed that when the constraints that the pre-filter contains were placed on the classification analysis, the predictive performance of the classifiers were similar to when the pre-filter was not used.

The second approach explored the feasibility of incorporating sets of functionally related genes into the classification process. Three publicly available datasets obtained from studies into breast cancer were used to assess the utility of this approach. A summary of each gene-set was derived by reducing the dimensionality of each gene-set via the use of Principal Co-ordinates Analysis. The reduced gene-sets were then ranked based on their ability to distinguish between the two classes (via Hotelling's T^2) and those most highly ranked were used to construct a classifier via logistic regression. The results from the analyses undertaken for this approach showed that it was possible to incorporate function information into the classification process whilst maintaining an equivalent (if not higher) level of predictive performance, as well as improving the biological interpretability of the classifier.

Acknowledgements

I could not have embarked on the journey that this thesis has taken me on without the Bright Futures Enterprise Scholarship I was awarded from the TEC in collaboration with Pacific Edge Biotechnology Ltd. Of equal importance to the completion of this thesis was the ongoing guidance, support and inspiration provided by my supervisor Dr. Micheal A. Black, even after he moved to Dunedin.

I would also like to thank my parents, Donna and Andrew Leader for their continued support, encouragement and for taking care of me when I felt that the world was coming to an end.

I wish to thank the Department of Biochemistry at the University of Otago for hosting me during my stay in Dunedin, in particular I wish to thank Moira Robinson, Dr. Fabienne Lecomte, Payal Diwadkar, Dr. Nattanan (Joy) Panjaworayan and Dr. Amonida Zadissa who made my time in Dunedin both informative and enjoyable.

I would also like to thank my fellow PhD sufferer and great friend Dr. Sarah Song for her lively conversions and continued support.

Finally I wish to thank Associate Professor Brian H. McArdle for taking on the official duty of being my main supervisor when Mik joined the Department of Biochemistry at the University of Otago half way through my PhD and the Department of Statistics at the University of Auckland for supporting me in my decision to move to Dunedin so as to remain under the supervision of Dr. Micheal A. Black.

Contents

Abstract i

Acknowledgements iii

List of Figures vii

List of Tables xi

1 Introduction 1

1.1 Molecular Biology . . . 3

1.2 Genomic Technologies . . . 4

1.2.1 High Throughput Genomics: Expression Microarrays . . . 4

1.2.2 Expression Data from Spotted and Affymetrix Microarrays . . . 10

1.3 Summary . . . 19

2 Tools for the Analysis of Gene Expression Data 21

2.1 Methods that Have Been Used Recently in Bioinformatic Applications . 21

2.1.1 Classification Methods . . . 22

2.1.2 Gene Selection Methods . . . 24

2.1.3 Bias Reduction Methods for Microarray Data . . . 28

2.1.4 Simulating Microarray Data . . . 30

2.2 Statistical Methods Suitable for the Analysis of Gene Expression Data 33

2.2.1 Multivariate Analysis Methods . . . 33

2.2.2 Univariate Analysis Methods . . . 42

2.2.3 Machine Learning Methods . . . 48

2.2.4 Resampling-Based Methods . . . 50

2.3 Bioinformatic Databases . . . 53

2.4 Software for the Analysis of Bioinformatic Data . . . 58

3 Gene Selection for Class Prediction 63

3.1 Motivation . . . 63

3.2 Issues in Gene Selection for Classifying Microarray Data 64
3.3 Gene Selection Techniques 65
3.4 Pre-Filtering Options 70
3.5 Predictor Construction 72
3.6 Melanoma Data 74
3.7 Simulated Datasets 75
3.7.1 Initial Simulation Methods 76
3.7.2 Data Driven Simulation Methods 77
3.7.3 Simulation Results 80
3.7.4 Initial Simulation Results 80
3.7.5 Data Driven Simulation Results 85
3.8 Application of the Proposed Methodology to a Melanoma Dataset . . . 95
3.9 Discussion of the Performance of the Methodology when Applied to both Simulated Data and the Melanoma Data 109

4 Incorporating Functional Information into the Classification Process 111
4.1 Motivation 111
4.2 Gene Set Analysis - Overview 112
4.2.1 Per-gene methods: GSEA and SAFE 113
4.2.2 Correlation-based Methods: PCOT2 and Global Test 115
4.2.3 Defining Gene-Sets 116
4.3 Extending Gene Set Analysis to Classification 120
4.4 Breast Cancer Microarray Data 120
4.5 Methods 124
4.5.1 Gene-Based Classification Method 124
4.5.2 Gene-Set-Based Classification Method 125
4.5.3 Comparison of Gene-Set Definition Methods 129
4.6 Results 130
4.7 Comparison of Results 156

5 Conclusions 161

Bibliography 165

List of Figures

1.1 The Central Dogma of Molecular Biology. The transcription of DNA into mRNA followed by the translation of the mRNA into protein. . . . 3

1.2 Spotted Array Experimental Process (Black, 2002). Genomic material is printed onto the slide, and fluorescently labelled cDNA from the two treatment groups is hybridized to the genomic material. The slide is then laser scanned to produce two images with are combined to give ratios of the intensity of fluorescence at each spot. 6

1.3 Loess Normalisation Process on the Mouse Data Contained in the R package **SMA** developed by Dudoit et al. (2002b). (a) Shows the pre-normalised M & A values with the loess smoothed line produced by the `plot.smooth.line` function in the **SMA** package (Dudoit et al., 2002b). (b) Shows the loess normalised M values *versus* the same A values. . . 15

1.4 Print-Tip Loess Normalisation Process on the Mouse Data Contained in the R package **SMA** developed by Dudoit et al. (2002b). (a) Shows the pre-normalised M & A values with a loess line for each print-tip. (b) Shows the print-tip loess normalised M values *versus* the same A values. 15

2.1 Graphical Example of the Difference Between (a) Euclidean and (b) Mahalanobis Spaces. 34

2.2 Graphical Example of the Difference Between (a) LDA and (b) QDA. . 35

2.3 Graphical Example of PCA for Two Variables. (a) shows the data in Table 2.1, (b) has the means for x1 and x2 superimposed as axes, (c) shows these axes rotated so that the first axis has the most variation and (d) shows the second principal component versus the first principal component. 37

2.4 Graphical Examples of Linkage Methods (a) Single Linkage, (b) Complete Linkage, (c) Average Linkage and (d) Centroid Linkage (McArdle, 2001). 41

2.5 An Example of a Dendrogram. The length of each branch represents the distance at which the two clusters were joined and each node represents a cluster (Everitt et al., 2001). 42

2.6 Kaplan-Meier survival curves for the Wang et al. (2005a) dataset which was split into two groups based on the estrogen receptor status. 45

2.7 Example of the maximal margin SVM classifier. The margin is the distance between the linear hyperplane (the solid line) and the support vectors, indicated by the dashed lines. 49

2.8 A Single Hidden Layer, Feed-Forward Neural Network (Hastie et al., 2001). The middle row is know as the hidden layer which forms a link between the input layer (bottom row) and the output layer (top row). This link is known as the *activation function*. 50

3.1 Visualisation of the pre-filter method with range set at -1 to 1 and α equal to 50%. Each panel represents the simulated $\log_2$ fold change with respect to a reference sample for one of three genes across multiple samples from two classes. Using the parameters stated above the genes represented in plots (a) and (b) would be retained by the pre-filter whereas the gene represented in plot (c) would not be retained. 72

3.2 Estimating σ^2 by fitting a Gamma distribution to the inverse of the observed variances of the genes in the "good" response class in the melanoma data (class 1). 79

3.3 Performance statistics obtained when applying the SVM classifier used in combination with the pre-filter and Mann-Whitney gene selection method to simulated data generated using Simulation Method 3. The horizontal axis shows the number of genes used in each classifier and the black lines represent the approximate 95% asymmetric confidence intervals for each classifier. Plot (a) shows the proportion of "poor" response samples that were correctly classified (i.e., the sensitivity), (b) shows the proportion of "good" response samples that were correctly classified (i.e., the specificity) and (c) shows the proportion of all the samples that were correctly classified (i.e., the classification rate) for each of the 99 classifiers. 84

3.4 Comparison of the classification rates with approximate 95% asymmetric confidence intervals obtained when the parameters in the pre-filter were changed. Plot (a) shows the effect of varing α for a fixed range of -1 to 1 and plot (b) shows the effect when the range is varied for a fixed α of 50%. 97

4.1 The GO hierarchical Structure for apoptosis. (Ashbuner et al. (2000), Release Date: 2008-05-15) 118
4.2 A graphical representation of the KEGG pathway for apoptosis (Kanehisa and Goto (2000); Kanehisa et al. (2006, 2008), Entry Number 04210, Release Date 4/14/04) 119
4.3 Flow Chart of Proposed Gene-Set-Based Classification Methodology . . 125
4.4 Venn diagrams displaying the overlap between (a) the three probe lists generated using the probe-based baseline method and (b) three gene lists generated using the gene-based baseline method 133
4.5 Venn diagrams displaying the overlap between (a) the three KEGG pathway lists obtained when the first two principal components of the top ranked KEGG pathways (extracted using the probe IDs) were used to construct the classifiers, and (b) the three KEGG pathway lists obtained when the first two principal components of the top ranked KEGG pathways (extracted using the gene symbols) were used to construct the classifiers. 137
4.6 A Venn diagram displaying the overlap between the three GO Term lists obtained when the first principal component of the top ranked GO Terms (which contained between 10 and 100 probes per term) were used to construct the classifiers. 141
4.7 Venn diagrams displaying the overlap between the GO Terms associated with each of the clusters used in the classifiers for the first 20 random splits of each dataset. The classifiers were constructed using the first principal component of the top ranked clusters, with the incorporation of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50 and the genes were standardised to $\mu = 0$, sd $= 1$. (a) compares all of the GO Terms associated with each of the clusters, (b) compares only the top 10 highly associated GO Terms for each of the clusters. 149
4.8 There were 13 GO Terms that were highly associated with at least 50% of the clusters used in the classifiers from the first 20 random splits for at least one of the three datasets. This barchart compares the proportion of clusters that each of these 13 GO Terms were highly associated with for each of the three datasets. The classifiers were constructed using the first principal component of the top ranked clusters, with the incorporation of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50 and the genes were standardised to $\mu = 0$, sd $= 1$. 154

4.9 Venn diagrams displaying the overlap between the GO Terms associated with each of the clusters used in the 11 classifiers for each dataset. The classifiers were constructed using the first principal component of the top ranked clusters, with the incorporation of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50 and the genes were standardised to $\mu = 0$, sd $= 1$. (a) compares all of the GO Terms associated with each of the clusters, (b) compares only the top 10 highly associated GO Terms for each of the clusters. . . 154

4.10 There were 15 GO Terms that were highly associated with at least 50% of the clusters used in the 11 classifiers for at least one of the three datasets. This barchart compares the proportion of clusters that each of these 15 GO Terms were highly associated with for each of the three datasets. The classifiers were constructed using the first principal component of the top ranked clusters, with the incorporation of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50 and the genes were standardised to $\mu = 0$, sd $= 1$. 155

List of Tables

2.1 Example Data for PCA 36

2.2 Default Subset of packages installed when using the `biocLite()` function, with a brief description quoted from the Release 2.0 Package Index and the Developmental Package Index pages found on the Bioconductor web-page 61

3.1 The first column shows the initial parameter values used in simulation method five. The columns labeled A - D show the four different sets of parameter values used to determine if any improvement in the average number of differentially expressed genes retained in the classified and/or the average classification rate could be attained using simulation method six. 78

3.2 Comparison of the SVM leave-one-out (LOO) classifier prediction results achieved for simulation method 1a 82

3.3 Comparison of the SVM double LOO (DLOO) classifier prediction results achieved for simulation method 1a 82

3.4 Comparison of the SVM LOO and DLOO classifier prediction results achieved for simulation method 1b 82

3.5 Comparison of the average SVM LOO $B.632$ bootstrap bias corrected prediction results obtained when each of the five gene selection methods were applied to 500 simulated data sets generated using simulation method 4. 95% confidence intervals (CI) are provided in brackets. . . . 88

3.6 Comparison of the average SVM LOO $B.632$ bootstrap bias corrected prediction results obtained when the pre-filter was combined with each of the five gene selection methods and were applied to 500 simulated data sets generated using simulation method 4. 95% confidence intervals (ECI) are provided in brackets. 89

3.7 Comparison of the average SVM LOO $B.632$ bootstrap bias corrected prediction results obtained when each of the five gene selection methods were applied to 100 simulated data sets generated using simulation method 6. 95% confidence intervals (CI) are provided in brackets. . . . 92

3.8 Comparison of the average SVM LOO $B.632$ bootstrap bias corrected prediction results obtained when the pre-filter was combined with each of the five gene selection methods and were applied to 100 simulated data sets generated using simulation method 6. 95% confidence intervals (CI) are provided in brackets. . . . 93

3.9 Comparing whether the SVM LOO classification was correct (Yes, below) when the moderated t-statistic was used to account for the technical replication in the melanoma data at the gene selection stage, compared to those obtained when the technical replication was overlooked at the gene selection stage. The comparison was also considered with the combination of the moderated t-statistic and the pre-filter. Approximate 95% asymmetric confidence intervals are included in brackets for the proportion correctly classified by each method. . . . 98

3.10 Comparison of the SVM LOO prediction results without bias corrections obtained when each of the five statistical gene selection methods were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets. . . . 102

3.11 Comparison of the SVM LOO prediction results without bias corrections obtained when the pre-filter was used in combination with each of the five gene selection methods and were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets. 103

3.12 Comparison of the SVM DLOO prediction results obtained when three of the gene selection methods were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets bellow each result. . . . 104

3.13 Comparison of the SVM DLOO prediction results obtained when the Pre-Filter was used in combination with each of the three chosen gene selection methods and were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets bellow each result. . . . 104

3.14 Comparison of the SVM LOO prediction results with $B.632$ bootstrap bias corrections obtained when each of the five gene selection methods were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets. . . . 105

3.15 Comparison of the SVM LOO prediction results with $B.632$ bootstrap bias corrections obtained when the pre-filter was used in combination with each of the five gene selection methods and were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals for the classification rates are included in brackets. 106

3.16 Comparison of the SVM apparent prediction results with $B.632$ bootstrap bias corrections obtained when each of the five gene selection methods were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets. 107

3.17 Comparison of the SVM apparent prediction results with bootstrap bias corrections obtained when the pre-filter was used in combination with each of the five gene selection methods and were applied to the Melanoma data. Approximate 95% asymmetric confidence intervals are included in brackets. 108

4.1 Clinical and pathological characteristics of the three datasets. Values for GSE2034 were adapted from Table 1 in Wang et al. (2005a) and supplementary information associated with the GSE2034 dataset obtained through the Genebank Gene Expression Ominbus database. Values for the GSE4922 and GSE6532 datasets were obtained from the clinical information available through the Genebank Gene Expression Ominbus database as supplementary information associated with each datasets. . 123

4.2 Comparison of the average performance statistics obtained when the probe-based method was used to construct a linear SVM classifier on the Wang et al. (2005a) dataset, with and without probe selection being stratified by estrogen receptor (ER) status. The performance statistics were based on 10 random splits of the dataset into training and test sets and the t-statistic was used to rank the probes. This table displays the maximum of the average classification rates obtained along with the average sensitivity and specificity for the classifiers associated with the maximum average classification rate. 95% empirical confidence intervals are shown in parentheses. 132

4.3 Comparison of the median performance statistics obtained when the probe-based method was used to construct a classifier, based on logistic regression, on the Wang et al. (2005a) dataset, with and without the use of a threshold at the point of classification. The performance statistics were based on 10,000 random splits of the dataset into training and test sets with only the probes contained in the 76-probe signature published by Wang et al. (2005a) used in the classifiers. The threshold value at which a test sample was predicted as belonging to class "1" (recurrent) was chosen as the minimum predicted class "1" probability attained by the recurrent samples in the training set that were correctly classified. Approximate 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 132

4.4 Summary of the median performance statistics obtained when the probe-based method was used to construct a classifier, based on logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 133

4.5 Summary of the median performance statistics obtained when the gene-based method was used to construct a classifier, based on logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. To obtain an estimate for the expression level of each gene, the expression levels for all of the probes associated with the same gene were averaged. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 133

4.6 Comparing the effect on the performance statistics, from one random split of the GSE2034 dataset into training and test sets, when the reduced KEGG pathways (extracted using the probe IDs) were ranked within estrogen receptor (ER) status to when the reduced KEGG pathways were ranked ignoring ER status. The classification models were constructed using the first principal component of the highly ranked reduced pathways as explanatory variables. 95% asymmetric confidence intervals are shown in parentheses. 136

4.7 Summary of the median performance statistics obtained when the first principal component of the top ranked KEGG pathways (extracted using the probe IDs) were used to construct a classifier, using logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 137

4.8 Summary of the median performance statistics obtained when the first two principal components of the top ranked KEGG pathways (extracted using the probe IDs) were used to construct a classifier, using logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 137

4.9 Summary of the median performance statistics obtained when the first principal component of the top ranked KEGG pathways (extracted using the gene symbols) were used to construct a classifier, using logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 138

4.10 Summary of the median performance statistics obtained when the first two principal components of the top ranked KEGG pathways (extracted using the gene symbols) were used to construct a classifier, using logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 138

4.11 Comparing the effect on the performance statistics of the KEGG-based classifier when each of the five standard deviation pre-filters were applied to the training data prior to defining the KEGG pathways (which were extracted using the probe IDs). The performance statistics were based on 1,000 random splits of the GSE2034 dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 139

4.12 Comparing the effect on the performance statistics of the KEGG-based classifier (with the pathways extracted using the gene symbols) when the number of top ranked pathways used as a source to construct the classifiers was varied. The performance statistics were attained using 11 fold cross validation on the GSE2034 dataset with the fist two principal components used to construct the classifiers. Approximate 95% asymmetric confidence intervals for the performance statistics are shown in parentheses. 139

4.13 Summary of the median performance statistics obtained when the first principal component of the top ranked GO Terms (which contained between 10 and 100 probes per term) were used to construct a classifier, using logistic regression, for each of the three datasets. The performance statistics were based on 1,000 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 141

4.14 Comparing the effect on the performance statistics of the GO Term based classifier when each of the six standard deviation pre-filters were applied to the training data prior to defining the GO Terms (which contained between 10 and 100 probes per term). The performance statistics were based on 1,000 random splits of the GSE2034 dataset into training and test sets. 142

4.15 Comparing the median performance statistics of the GO Term based classifier when the range on the number of probes per GO Term was increased from 10-100 to 10-1,000. The classification models were built using the first two principal components of the top ranked GO Terms as explanatory variables and the median performance statistics were based on 1,000 random splits of the GSE4922 dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 143

4.16 Comparison of the median performance statistics obtained when only GO Terms associated with cell cycle were used in the classifier to those obtained based on all of the GO Terms with between 10 and 1,000 probes per term. The comparison was made using the GSE4922 dataset which was randomly split 1,000 times into training and test sets. The classifiers were constructed using the first two principal components of the top ranked GO Terms as explanatory variables. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 143

4.17 Comparing the performance statistics of three linkage methods and three cutting points of the resulting dendrograms, for the correlation-based method for classifier construction applied to the GSE4922 dataset. One random split of the dataset into training and test sets was used and the classifier was constructed using only the first principal component of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 150

4.18 Comparing the performance statistics of three linkage methods and three cutting points of the resulting dendrograms, for the correlation-based method for classifier construction applied to the GSE4922 dataset. One random split of the dataset into training and test sets was used and the classifier was constructed using the first two principal components of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 150

4.19 Comparing the performance statistics when the classifier was constructed using clusters derived from the recurrent (R) training samples to the performance statistics obtained when the classifier was constructed using clusters derived from the non-recurrent (NR) training samples. The correlation-based method using Average Linkage in the construction of the clusters was applied to the GSE4922 dataset in each case. The performance statistics were also compared across three different cutting points of the resulting dendrograms. One random split of the dataset into training and test sets was used and the classifier was constructed using only the first principal component of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 151

4.20 Comparing the performance statistics when the classifier was constructed using clusters derived from the recurrent (R) training samples to the performance statistics obtained when the classifier was constructed using clusters derived from the non-recurrent (NR) training samples. The correlation-based method using Average Linkage in the construction of the clusters was applied to the GSE4922 dataset in each case. The performance statistics were also compared across three different cutting points of the resulting dendrograms. One random split of the dataset into training and test sets was used and the classifier was constructed using the first two principal components of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 151

4.21 Comparison of the performance statistics obtained when the correlation-based method was applied to one random split of the GSE4922 dataset into training and test sets, with and without the use of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50. Complete linkage was used in the construction of the clusters and the genes were standardised to $\mu = 0$, sd $= 1$. The classifiers were constructed using the first principal component of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 152

4.22 Comparison of the performance statistics obtained when the correlation-based method was applied to one random split of the GSE4922 dataset into training and test sets, with and without the use of a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50. Complete linkage was used in the construction of the clusters and the genes were standardised to $\mu = 0$, sd $= 1$. The classifiers were constructed using the first two principal components of the top ranked clusters as explanatory variables. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 152

4.23 Comparison of the performance statistics obtained when the correlation-based method was applied to one random split of the GSE4922 dataset into training and test sets, with the classifier built using binary explanatory variables created using the means of the first principal component and then the first two principal components of the top ranked clusters in the training set. The complete linkage method was used in the construction of the clusters, with the genes standardised to $\mu = 0$, sd $= 1$ and a threshold used at the point of classification to ensure that the sensitivity in the training set was greater than 0.50. Approximate 95% asymmetric confidence intervals for the classification rates are shown in parentheses. 152

4.24 Summary of the median performance statistics obtained when the first principal component of the top ranked clusters from the correlation-based method were used to construct a classifier for each of the three datasets. Complete linkage was used to define the clusters and the genes were standardised to $\mu = 0$, sd $= 1$. The performance statistics were based on 100 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 153

4.25 Summary of the median performance statistics obtained when the first principal component of the top ranked clusters from the correlation-based method were used to construct a classifier, which incorporated a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50, for each of the three datasets. Complete linkage was used to define the clusters and the genes were standardised to $\mu = 0$, sd $= 1$. The performance statistics were based on 100 random splits of each dataset into training and test sets. 95% empirical confidence intervals for the median performance statistics are shown in parentheses. 153

4.26 Summary of the 11 fold cross validation performance statistics obtained when the first principal component of the top ranked clusters from the correlation-based method were used to construct a classifier, which incorporated a threshold at the point of classification to ensure that the sensitivity in the training set was greater than 0.50, for each of the three datasets. Complete linkage was used to define the clusters and the genes were standardised to $\mu = 0$, sd $= 1$. Approximate 95% asymmetric confidence intervals for the performance statistics are shown in parentheses. 153