Feature Selection and Modelling Methods for Microarray Data from Acute Coronary Syndrome

A thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

Adrian Alecu

Department of Automatic Control and Systems Engineering

February 2015

To my family

Acknowledgements

I would like to thank my supervisors Professor Daniel Coca and Doctor Timothy Chico for the guidance offered throughout my years of research in Sheffield. I also would like thank Doctor Marta Milo for providing the data for this project.

Many thanks go to my colleagues in Room 316 who made the PhD experience more enjoyable and rewarding. In particular, I would like to thank Krishnanathan Kirubhakaran, Dorian Florescu and Andrew Hills for all the work-related discussions as well as for being great friends in time of need.

A massive thanks to all my friends from Sheffield and abroad who helped me write some adventure chapters in my book of days. In particular I would like to thank Ines for always keeping in touch in spite of my countless attempts at taking refuge in work to escape her enthusiastic randomness. A big thank you goes to Alina and Hibo for all the philosophical debates, to Irina for the occasional social dances and to Krish and Mihai for all the sport activities.

Finally and most importantly, I would like to thank my mother, my sister, Caterina, and my brother, Marius, for the constant support offered in the years I have been away from home.

Abstract

Acute coronary syndrome (ACS) represents a leading cause of mortality and morbidity worldwide. Providing better diagnostic solutions and developing therapeutic strategies customized to the individual patient represent societal and economical urgencies. Progressive improvement in diagnosis and treatment procedures require a thorough understanding of the underlying genetic mechanisms of the disease. Recent advances in microarray technologies together with the decreasing costs of the specialized equipment enabled affordable harvesting of time-course gene expression data. The high-dimensional data generated demands for computational tools able to extract the underlying biological knowledge.

This thesis is concerned with developing new methods for analysing time-course microarray gene expression data, focused on identifying differentially expressed genes, deconvolving heterogeneous gene expression measurements and inferring dynamic gene regulatory interactions. The main contributions include: a novel multi-stage feature selection method, a new deconvolution approach for estimating cell-type specific signatures and quantifying the contribution of each cell type to the variance of the gene expression patters, a novel approach to identify the cellular sources of differential gene expression, a new approach to model gene expression dynamics using sums of exponentials and a novel method to estimate stable linear dynamical systems from noisy and unequally spaced time series data.

The performance of the proposed methods was demonstrated on a time-course dataset consisting of microarray gene expression levels collected from the blood samples of patients with ACS and associated blood count measurements. The results of the feature selection study are of significant biological relevance. For the first time is was reported high diagnostic performance of the ACS subtypes up to three months after hospital admission. The deconvolution study exposed features of within and between groups variation in expression measurements and identified potential cell type markers and cellular sources of differential gene expression. It was shown that the dynamics of post-admission gene expression data can be accurately modelled using sums of exponentials, suggesting that gene expression levels undergo a transient response to the ACS events before returning to equilibrium. The linear dynamical models capturing the gene regulatory interactions exhibit high predictive performance and can serve as platforms for system-level analysis, numerical simulations and intervention studies.

Contents

Nomenclature **xiii**

Acronyms **xv**

1 Introduction **1**

1.1 Background . . . 1

1.2 Motivation . . . 2

1.3 Overview of the thesis . . . 4

2 Feature selection methods for microarray gene expression data **6**

2.1 Introduction to microarray technology . . . 7

2.1.1 Glass slide cDNA microarrays . . . 7

2.1.2 High-density oligonucleotide microarrays . . . 8

2.2 Gene expression profiling using microarrays . . . 9

2.2.1 Experimental design . . . 9

2.2.2 Sample preparation . . . 9

2.2.3 Hybridization and washing . . . 10

2.2.4 Image acquisition . . . 10

2.2.5 Data pre-processing . . . 10

2.2.6 Data analysis . . . 13

2.3 Statistical hypothesis testing in microarray experiments . . . 13

2.3.1 Single hypothesis testing . . . 14

2.3.2 Multiple hypothesis testing . . . 15

2.3.3 Statistical significance and biological relevance . . . 18

2.4 Feature selection in machine learning . . . 19

2.4.1 Unsupervised filtering approaches for microarray data . . . 21

2.4.2 Filters based on information theory . . . 24

2.4.3 Wrapper methods . . . 27

2.4.4 Embedded methods based on SVM . . . 30

2.5 Conclusions . . . 31

3 Modelling methods for microarray gene expression data 33

3.1 Deconvolution of microarray gene expression data from heterogeneous tissues 33

3.1.1 The linear deconvolution model 34

3.1.2 Estimation of cell type-specific proportions 35

3.1.3 Estimation of cell type-specific expression levels 36

3.1.4 Joint estimation of cell type-specific proportions and expression levels 37

3.2 Modelling dynamic GRNs 39

3.2.1 General properties of GRNs 40

3.2.2 Boolean and probabilistic Boolean networks 42

3.2.3 Dynamic Bayesian networks 44

3.2.4 Linear state-space models 45

3.3 Conclusions 49

4 A novel multi-stage feature selection method for microarray gene expression data 50

4.1 Stage I: Filtering uninformative genes 52

4.2 Stage II: Gene subset selection using statistical tests 53

4.2.1 The Welch's t-test 53

4.2.2 The Wilcoxon's rank-sum test 54

4.2.3 pFDR control and q-value estimation 55

4.3 Stage III: Gene subset selection using mRMR 56

4.3.1 The CACC discretization algorithm 56

4.4 Stage IV: Gene subset selection using SVMs 58

4.5 Model selection and reliable performance evaluation 60

4.5.1 The nested cross-validation design 60

4.5.2 The multi-stage feature selection method 62

4.6 Gene selection for ACS classification 66

4.6.1 The ACS dataset 66

4.6.2 Selection of differentially expressed genes between MI and UA 67

4.6.3 Selection of differentially expressed genes between NSTEMI and STEMI 74

4.7 Conclusions 79

5 A novel deconvolution method for microarray gene expression data 82

5.1 An orthogonal forward regression approach for microarray data deconvolution 83

5.2 Testing for cell type-specific differential expression 86

5.3 Expression deconvolution of the genes differentiating between the ACS subtypes . . . 89
5.3.1 The CBC dataset . . . 89
5.3.2 Expression deconvolution of the genes differentiating MI from UA . . . 90
5.3.3 Expression deconvolution of the genes differentiating NSTEMI from STEMI . . . 96
5.4 Conclusions . . . 101

6 A novel approach for modelling stable GRNs 103
6.1 Nonlinear approximation of gene expression dynamics by sums of exponentials . . . 104
6.1.1 The gene expression model . . . 104
6.1.2 Parameter estimation using non-linear least squares . . . 106
6.1.3 Selecting the regularization parameter using the Morozov's discrepancy principle . . . 108
6.2 Modelling GRNs using linear dynamical systems . . . 109
6.2.1 Parameter estimation using non-linear least squares . . . 109
6.2.2 Selecting the regularization parameter by cross-validation . . . 111
6.3 Estimating sparse GRNs . . . 112
6.4 Modelling GRNs for the diagnostic groups of ACS . . . 113
6.4.1 Modelling the regulatory interactions between the genes differentiating MI from UA . . . 113
6.4.2 Modelling the regulatory interactions between the genes differentiating NSTEMI from STEMI . . . 117
6.5 Conclusions . . . 121

7 Conclusions 123
7.1 Summary and Conclusions . . . 123
7.2 Future work . . . 125

A Fundamentals of genetics 127

B Feature selection for ACS classification 132

C Deconvolution of microarray gene expression data 144

D Reconstruction of GRNs for the ACS subtypes 152

Bibliography 165

List of Figures

4.1 Hierarchical organization of the stage-specific forms of relevance . . 51

4.2 Nested cross-validation design of the multi-stage feature selection algorithm. 65

4.3 Expression levels across arrays for (a) a non-informative probe set and (b) an informative probe set. The bars denote the technical variance of each measurement. 68

4.4 Mean densities (dotted lines) of the Hedges' g scores for the subsets of gene selected using pFDR (Stage II) and mRMR (Stage III) in the external cross-validation loop. The continuous lines denote one standard deviation around the mean densities. 68

4.5 Hierarchical clustering of the group-specific expression averages of the genes selected at Stage III in one fold of the external cross-validation loop. 69

4.6 Principal component analysis of the combined expression data of $\mathcal{S}_f$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space. 71

4.7 Group-specific expression averages across visits for (a) X_1 (WASH1), (b) X_3 (C17orf103) and (c) X_6 (OSBP2). The bars denote one standard deviation around the mean expression levels. 71

4.8 Principal component analysis of the combined expression data of $\mathcal{S}_f^*$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space. 74

4.9 Expression level across arrays for (a) a non-informative probe set and (b) an informative probe set. The bars denote the technical variance of each measurement 75

4.10 Mean density (dotted line) of the Hedges' g scores for the subsets of gene selected using pFDR (Stage II) in the external cross-validation loop. The continuous lines denote one standard deviation around the mean density. 75

4.11 Hierarchical clustering of the group-specific expression averages of the genes selected at Stage II in one fold of the external cross-validation loop. 76

4.12 Principal component analysis of the combined expression data of $\mathcal{S}_f$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space. 77

4.13 Group-specific expression averages across visits for (a) X_1 (HLA-DQB1), (b) X_2 (MAPK8Ip1) and (c) X_{21} (LRRC37A). The bars denote one standard deviation around the mean expression levels. 77

4.14 Principal component analysis of the combined expression data of $\mathcal{S}_f^*$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space. 79

5.1 Distributions of the CODs across visits for the models fitted on data from the MI and UA groups using (a) the constrained OFR and (b) the unconstrained OFR. The lower and upper edge of the box represent the 25th and 75th percentiles, respectively, while the whiskers extend to the most extreme points not considered outliers. 93

5.2 Cell type-specific contributions to the expression level of genes X_1, X_3 and X_6 in the MI and UA groups. 94

5.3 Distributions of the blood cell counts in the MI and UA groups . . . 95

5.4 Distributions of the CODs across visits for the models fitted on data from NSTEMI and STEMI groups using (a) the constrained OFR method and (b) the unconstrained OFR method. 98

5.5 Cell type-specific contributions to the expression level of genes X_1, X_2 and X_3 in the NSTEMI and STEMI groups. 99

5.6 Distributions of the blood cell counts in the NSTEMI and STEMI groups 100

6.1 Estimated dynamic profiles for (a) X_1 (WASH1), (b) X_3 (C17orf103) and (c) X_6 (OSBP2). The bars denote one standard deviation around the group-specific gene expression averages. 114

6.2 Distributions of the prediction errors for the families of GRNs. . . . 116

6.3 Sparsity filters for the parameters of the GRNs associated with NSTEMI, STEMI and UA 117

6.4 Regulatory map for genes X_1-X_5 in the: (a) NSTEMI network, (b) STEMI network and (c) UA network. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene. 118

6.5 Estimated dynamic profiles for (a) X_1 (HLA-DQB1), (b) X_2 (MAPK8IP1) and (c) X_{21} (LRRC37A). The bars denote one standard deviation around the group-specific gene expression averages. 119

6.6 Distributions of the prediction errors for the families of GRNs 120

6.7 Sparsity filters for the parameters of the GRNs associated with NSTEMI and STEMI 121

6.8 Regulatory map for genes X_1-X_5 in the: (a) NSTEMI network and (b) STEMI network. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it 122

A.1 DNA double helix 127

A.2 RNA hairpin loop 128

A.3 The fundamental stages of protein biosynthesis 128

A.4 Schematic representation of a GRN. Gene1 and Gene4 jointly regulate the expression level (abundance of ribonucleic acid (RNA) and protein) of Gene5 through the protein complex assembled for their individual translation products (Protein1 and Protein4). The amount of Protein1 is regulated by Gene2 through RNA2 that binds to molecules of RNA1, preventing further protein translation. 129

A.5 Affymetrix GeneChip design 130

A.6 Example of a gene (X_1) with three inward connections (in-degree) and eleven outward connections (out-degree). X_1 is regulated by X_2, X_3 and X_4 and regulates the expression level of genes $X_5 - X_{15}$. 131

B.1 Group-specific expression averages across visits for genes $X_1 - X_{18}$ in the MI vs. UA study. 135

B.2 Group-specific expression averages across visits for genes $X_{19} - X_{36}$ in the MI vs. UA study. 136

B.3 Group-specific expression averages across visits for genes $X_1 - X_{15}$ in the NSTEMI vs. STEMI study. 141

B.4 Group-specific expression averages across visits for genes $X_{16} - X_{21}$ in the NSTEMI vs. STEMI study. 142

D.1 Estimated dynamic profiles for genes $X_1 - X_{12}$ in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages 153

D.2 Estimated dynamic profiles for genes $X_{13} - X_{20}$ in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages 154

D.3 Dynamic profiles of the combined expression levels for genes $X_1^* - X_{12}^*$ selected by l_1-StaR in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages 156

D.4 Dynamic profiles of the combined expression levels for genes $X_{13}^* - X_{20}^*$ selected by l_1-StaR in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages 157

D.5 Gene regulatory network for genes X_1-X_{20} in the NSTEMI group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene. 158

D.6 Gene regulatory network for genes X_1-X_20 in the STEMI group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene. 159

D.7 Gene regulatory network for genes X_1-X_{20} in the UA group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene. 160

D.8 Estimated dynamic profiles for genes X_1-X_{12} in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages 161

D.9 Estimated dynamic profiles for genes X_{13}-X_{21} in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages 162

D.10 Dynamic profiles of the combined expression levels for genes $X_1^* - X_{11}^*$ selected by l_1-StaR in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages. 164

List of Tables

2.1 Possible outcomes from m hypothesis tests 15

4.1 Quanta matrix for feature X_j 57
4.2 Amount of patients by visits and diagnostic groups 67
4.3 Classification performance of the subsets of genes $\mathcal{S}^{IV}$ 70
4.4 l_1-STaR parameters 72
4.5 l_1-STaR performance and number of selected genes for each η 72
4.6 Classification performance of the subsets of genes selected by l_1-StaR 73
4.7 Classification performance for the subsets of genes $\mathcal{S}^{IV}$ 76
4.8 l_1-STaR performance and number of selected genes for each η 78
4.9 Classification performance of the subsets of genes selected by l_1-StaR 78

5.1 Amount of arrays associated with CBC data 90
5.2 Average count for each blood cell in the diagnostic groups of ACS . 90
5.3 Gene differentially expressed between NSTEMI and STEMI 91
5.4 Correlation between blood cell counts in the MI and UA groups . . 92
5.5 Gene differentially expressed between NSTEMI and STEMI 96
5.6 Pairwise correlations between blood cell in the NSTEMI and STEMI groups 97
5.7 Gene differentially expressed between NSTEMI and STEMI 101

6.1 Eigenvalues of the GRNs associated with the ACS subtypes 115
6.2 Eigenvalues of the GRNs associated with NSTEMI and STEMI . . . 120

B.1 Sample of estimated gene expression levels 133
B.2 Standard errors of the estimated gene expression levels listed in Table B.1 134
B.3 Differentially expressed genes between the MI and UA groups selected by the multi-stage feature selection method 137

B.4 Optimal classifier parameters for the MI vs. UA study 138
B.5 Differentially expressed genes between the MI and UA groups selected by l_1-StaR 139
B.6 Differentially expressed probe sets between the NSTEMI and STEMI groups selected by the multi-stage method 140
B.7 Optimal classifier parameters for the NSTEMI vs. STEMI study . . . 140
B.8 Differentially expressed genes between the NSTEMI and STEMI groups selected by l_1-StaR 143

C.1 Cell type-specific contributions to the variance of the genes $X_1 - X_{18}$ in the MI group 145
C.2 Cell type-specific contributions to the variance of the genes $X_{19} - X_{36}$ in the MI group 146
C.3 Cell type-specific contributions to the variance of the genes $X_1 - X_{18}$ in the UA group 147
C.4 Cell type-specific contributions to the variance of the genes $X_{19} - X_{36}$ in the UA group 148
C.5 Cell type-specific contributions to the variance of the genes $X_1 - X_{21}$ in the NSTEMI group 149
C.6 Cell type-specific contributions to the variance of the genes $X_1 - X_{21}$ in the STEMI group 150
C.7 Amount of genes expressed in each cell type in the MI and UA groups151
C.8 Amount of genes expressed in each cell type in the NSTEMI and STEMI groups 151

D.1 Goodness of fit for the genes differentiating MI from UA 155
D.2 Goodness of fit for the genes differentiating NSTEMI from STEMI . 163

Nomenclature

A list of the variables and notation used in this thesis is defined below. The definitions and conventions set here will be observed throughout unless otherwise stated. For a list of acronyms, please consult page xv.

$\#(S)$	cardinality of set S
α	significance level for hypothesis testing
$\cap$	set intersection
$\cup$	set union
Δt	sampling time
$\det$	determinant of a matrix
$\dim V$	dimension of the vector space V
$\emptyset$	empty set
η	regularization parameter
γ	rejection region in statistical hypothesis testing
$\kappa(\cdot,\cdot)$	kernel function for the SVM classifier
$\ker V$	kernel of the vector space V
λ	eigenvalue
$[x]_+$	maximum of x and 0
$\mathbb{C}$	complex numbers
$\mathbb{R}_+$	positive real numbers
$\mathbb{Z}$	integer numbers

$\mathcal{N}(\mu, \Sigma)$	normal distribution with mean μ and covariance Σ
$\mathcal{U}[a, b]$	uniform distribution with support $[a, b]$
$\setminus$	set difference
$\mathrm{var}(x)$	variance of the random variable x
$\top$	matrix transpose
ϱ	tolerance parameter for the OFR algorithm
$\mathbf{I}$	identity matrix
C	box constraint for the SVM classifier
$E[\cdot]$	expected value operator
p	p-value
$P(x)$	probability density function of the random variable x
t	time

Acronyms

A/M/P Absent/Marginal/Present. 22, 23

ACS acute coronary syndrome. ii, 1–3, 5, 52, 66, 80, 81, 89, 90, 101, 113, 114, 121, 123–126

AIC Akaike information criterion. 37

ANOVA analysis of variance. 125

BDe Bayesian Dirichlet equivalent. 45

BIC Bayesian information criterion. 45

BNRC Bayesian network and nonparametric regression criterion. 45

CACC class-attribute contingency coefficient. 56–58

CBC complete blood count. 89–91, 97

cDNA complementary DNA. 7–10, 31, 129

CMIM conditional mutual information maximization. 26

COD coefficient of determination. 43, 85, 92, 97, 101, 102

cRNA complementary RNA. 10, 129

csSAM cell type-specific significance analysis of microarrays. 37

DISR double input symmetrical relevance. 27

DNA deoxyribonucleic acid. 127–129

DrSVM doubly regularized SVM. 31

ECG electrocardiogram. 2

EM expectation maximization. 45–47

FARMS factor analysis for robust microarray summarization. 12, 13, 23, 53

FDR false discovery rate. 16–18, 87

FWER family-wise error rate. 15–18

GCV generalized cross-validation. 112

GEO Gene Expression Omnibus. 35

GRN gene regulatory network. 3–5, 33, 39–41, 44, 47–49, 103, 104, 109–113, 115–119, 121, 124, 126, 129

I/NI informative/non-informative. 23, 24, 53

IM ideal mismatch. 12

JMI joint mutual information. 27, 32

LASSO least absolute shrinkage and selection operator. 40

LCM laser capture micro-dissection. 34, 35

LDA linear discriminant analysis. 21, 28

LOOCV leave-one-out-CV. 61

MAS 5.0 Affymetrix microarray suite 5.0. 11, 12, 22

MCMC Markov chain Monte Carlo. 38

MI myocardial infarction. 2, 3, 66–70, 78, 82, 89–92, 94, 96, 113, 114, 118, 125

MID mutual information difference. 26, 56, 80

MIFS mutual information based feature selection. 26

MIQ mutual information quotient. 26, 56, 80

MM mismatch. 8, 11–13, 22

mRMR minimum redundancy - maximum relevance. 26, 27, 32, 50, 56, 80, 125

mRNA messenger ribonucleic acid. 6, 9, 10, 23, 34, 128, 129

multi-mgMOS multi-chip modified gamma Model for Oligonucleotide Signal. 12, 13, 52, 66, 79, 108

NB naive Bayes. 21, 27

NSTEMI non-ST elevation MI. 2, 3, 66, 69, 74–78, 82, 89, 91, 96, 97, 100, 113, 115, 116, 119

OFR orthogonal forward regression. 4, 82, 83, 85–87, 89, 92, 97, 98, 101, 124

OLS orthogonal least squares. 83, 84, 87, 88

PCR polymerase chain reaction. 7, 8

pFDR positive false discovery rate. 17, 18, 32, 53, 55, 80, 89

PM perfect match. 8, 11–13, 22, 23

PPLR probability of positive log ratio. 66

RFE recursive feature elimination. 30, 50

RMA robust multi-array average. 11–13

RMSPE root mean square percentage error. 116, 119

RNA ribonucleic acid. viii, 9, 66, 92, 101, 127–129

RT reverse transcriptase. 9, 10

STEMI ST elevation MI. 2, 3, 66, 69, 74–78, 82, 89, 91, 96, 97, 100, 113, 115, 116, 119

SVM support vector machine. 21, 28–32, 50, 58, 59, 72, 80, 125

SVM-RFE SVM with recursive feature elimination. 30, 50, 51

UA unstable angina. 2, 66–70, 78, 82, 89–92, 94, 96, 113–116, 118, 125

Chapter 1

Introduction

1.1 Background

The cellular pathways are regulated by complex functional interactions between genes. Abnormal levels of gene expression can indicate irregularities in cell functioning (e.g. diseases) induced by changes in gene regulatory interactions (Emilsson et al., 2008, Schadt et al., 2005). Gene expression profiling and analysis can provide a deeper understanding of current diseases and assist personalized medicine to develop more efficient treatments tailored to the individual patients that account for their unique genetic variations.

Gene expression profiling is performed using high-throughput screening technologies such as microarrays, which allow for a genome-wide interrogation of the cell's transcriptional activity. Depending on the objectives of the experiment, the high dimensionality of the microarray data can raise significant computational and theoretical challenges in terms of extracting the underlying biological knowledge. Responses to these challenges emerged from various disciplines such as applied mathematics, computers science, statistics and engineering which recently defined the multidisciplinary field of computational biology. This field became the working ground for uncovering the dynamics of diseases that eluded the traditional medical understanding.

One such disease and a leading cause of mortality worldwide is acute coronary syndrome (ACS), which accounts for more than 2.5 million hospitalisations each year (Grech and Ramsdale, 2003). ACS is caused by the rupture of an atherosclerotic plaque (accumulation of fatty and calcium substances on an artery wall), resulting in a complete, partial or intermittent obstruction of blood supply to the heart (Libby, 2001). This occurs when inflammatory reactions caused by the interaction between macrophages and low density lipoproteins (molecules in charge with the transport of cholesterol) trigger plague erosion or disruption (Davies, 2000). The resulting thrombosis (blood clot inside a blood vessel) decreases the oxygen supply to the heart and leads to chest pain or to myocardial infarction (MI). Major factors influencing the risk of ACS consists of age, sex, family history, smoking, alcohol consumption and type II diabetes (Overbaugh, 2009).

ACS is usually diagnosed by performing an electrocardiogram (ECG) test. This test allows the doctor to evaluate the heart's electrical activity. Abnormal changes in the electrical activity patterns can be used to identify the three subtypes of ACS: non-ST elevation MI (NSTEMI), ST elevation MI (STEMI) and unstable angina (UA), which differ with respect to severity, duration and treatment (Overbaugh, 2009). The ECG test is often insufficient to accurately diagnose MI due to other medical conditions presenting ST deviations (Ahmad and Sharma, 2012). To assist and improve the diagnostic accuracy of MI, cardiac markers such as troponin, creatine kinase-MB and myoglobin are often measured. The markers rise in response to the ischemic event and return to baseline in 24 hours (myoglobin), 72 hours (creatine kinase-MB) or are cleared from circulation up to 14 days (troponin) after infarction (Ahmad and Sharma, 2012).

The symptoms of the disease, ranging from chest pain (with or without radiation to the left arm, back or neck), nausea, shortness of breath and fatigue, emerge when the atherosclerotic process is already well-developed and often result in critical coronary events. Recent figures show that in UK only, every four minutes someone suffering from ACS is admitted to hospital and over 90 people die every day from a heart attack (Associates, 2011), a large proportion of deaths occurring before the patients reach a hospital. Treatment procedures consisting of drug therapy, percutaneous cardiac intervention or coronary artery bypass grafting, induce substantial healthcare expenditure and economic losses, amounting to £3.6 billion in 2009-10 (Associates, 2011). The high mortality rate and the current economic burden of the disease highlight the urgency to acquire a better understanding of the genetics of ACS in order to improve diagnosis and treatment solutions.

1.2 Motivation

The increasing affordability of gene expression profiling services empowered time-course studies of ACS (Kiliszek et al., 2012, Silbiger et al., 2013). These studies focused on identifying genes serving as new biomarkers for the early stages of the disease and for monitoring cardiac ischemic recovery. Extending the temporal range of the gene expression studies together with the spectrum of biological questions addressed using the resulted time-course data could greatly advance our understanding of the disease.

This thesis addresses three fundamental challenges related to the genetics of ACS from a computational biology perspective. The first challenge consists of identifying genes differentiating between MI (NSTEMI and STEMI) and UA patients, and between NSTEMI and STEMI patients, up to three months after hospital admission given time-course microarray data measured from blood samples. These findings could reveal novel cardiac markers for long term diagnosis or indicate genes explaining the genetic predisposition to ACS.

The second challenge consists of identifying the blood cells expressing the genes discriminating between the ACS subtypes and quantifying their contribution to the variability and abundance of the total gene expression measured. These findings could indicate sources of interindividual variability in the gene expression patterns, reveal the cellular sources of differential gene expression or pinpoint cell type-specific markers (gene uniquely expressed only in particular cell types).

The third challenge consists of inferring the regulatory pathways between the genes distinguishing between the ACS subtypes. Models describing the dynamics of the gene regulatory networks (GRNs) could be used to make qualitative and quantitative predictions about the network's behaviour under different conditions. Numerical simulations and intervention studies based on these networks could expose the underlying mechanisms of the disease and assist the development of more efficient treatments.

To address the first challenge, a novel feature selection method consisting of four stages is proposed. In the first stage, a new unsupervised filter is used to remove noisy and uninformative genes based on their biological and technical variance. The second stage uses standard test statistics to select differentially expressed genes while accounting for the problems arising when simultaneously testing multiple hypotheses. The third stage adopts an information theoretic based criterion operating on discretized data to select genes highly correlated with the ACS subtypes and minimally correlated with each other. Data discretization is performed using a state-of-the-art discretization algorithm that accounts for the unequal number of patients in each diagnostic group of ACS. The final stage combines a search strategy with a state-of-the-art classifier to select a minimal subset of genes with the highest diagnostic performance. To provide an unbiased estimate of the diagnostic performance and avoid the sources of bias incurred in feature selection and parameter estimation, the stages are embedded in a nested cross-validation framework.

To address the second challenge, a novel deconvolution method for microarray gene expression data is proposed. This method combines non-negative least square optimization with the orthogonal forward regression (OFR) approach proposed by Billings et al. (1988) to estimate positive cell type-specific expression levels and quantify their contribution to the variance of the gene expression patterns. To identify the cellular sources of differential gene expression, an approach for comparing the coefficients of two regression models is adopted from the econometrics literature. This approach consists of introducing interaction terms between the regressing variables and the covariates (diagnostic groups) into the deconvolution model and testing the significance of their coefficients using a Wald test (Wald, 1943).

To address the third challenge, a novel method to estimate stable linear dynamical systems from time course gene expression data is proposed. The approach exploits the particular form of the state transition matrix when the dynamical system has single eigenvalues. The parameters of the state transition matrix are estimated using a regularized nonlinear optimization approach incorporating stability constraints. To generate enough gene expression data for reconstructing GRNs of arbitrarily large sizes, a new approach based on modelling gene expression dynamics using sums of exponentials is proposed. This approach can handle unequally sampled time series data and accounts for the measurement noise.

1.3 Overview of the thesis

This thesis is organized as follows:

- Chapter 2 provides an in-depth review of feature selection methods for microarray data widely used in computational biology. The chapter begins with an introduction to microarray technology followed by an overview of the central steps of a microarray experiment. Emphasis is put on the Affymetrix GeneChip format and associated data pre-processing methods. The chapter continues with an introduction to statistical hypothesis testing in microarray experiments and a review of the multiple comparison procedures. The relation between statistical significance and biological relevance is further discussed. An introduction to feature selection in machine learning is presented setting the scene for an extensive review of the state-of-the-art methods used to remove noisy and uninformative genes and select subsets of highly relevant features.

- Chapter 3 provides a review of modelling methods for microarray data focused on gene expression deconvolution and reconstruction of GRNs. The review of deconvolution methods highlights the biases induced by sample heterogeneity in differential expression studies and discusses computational tools addressing different deconvolution problems. The review of GRN reconstruction methods lists the biological properties of GRNs together with strategies to incorporate biological constraints into the modelling process. Three major modelling formalism are reviewed and their strengths and limitations are highlighted.

- Chapter 4 presents the novel multi-stage feature selection method for microarray data. A comparative study of the performance of the novel method and another multi-stage approach is conducted on a time-course microarray dataset collected from patients suffering from ACS. The results report the comparable performance of the two approaches expressed in terms of discriminatory power of the selected genes and highlight the appropriateness of the novel method for biomarker discovery in time-course microarray studies.

- Chapter 5 introduces the novel deconvolution method together with the approach to conduct cell type-specific differential expression analysis. These methods are applied on the genes discriminating between the subtypes of ACS to identify the cellular sources of differential expression and measure the increments to the proportion of explained gene expression variance associated with each cell type. Results from this study show high deconvolution performance for most of the genes and argue that features of interindividual variability specific to microarray data collected from blood samples are responsible for the low variability captured by the deconvolution model in the case of some genes. The need to supplement cell type-specific differential expression analysis with information regarding the deconvolution performance is also formulated.

- Chapter 6 presents the novel approaches to model gene expression dynamics and reconstruct stable GRNs. Additionally, a method to obtain sparse topologies for the estimated GRNs is discussed. Results show that modelling time course gene expression data using sums of exponentials adequately capture the dynamics of the genes differentiating between the ACS subtypes and that the novel method for GRN reconstruction estimates stable dynamical systems whose trajectories match extremely well the trajectories approximated using sums of exponentials. Sparse topological representations for the estimated GRNs are derived and briefly discussed.

- Chapter 7 concludes the work done in this thesis and provides suggestions for future directions of research

Chapter 2

Feature selection methods for microarray gene expression data

A major challenge in molecular biology is to uncover disease-specific genes that can be used as targets for treatment or as biomarkers for diagnosis. This challenge is traditionally approached through differential gene expression studies comparing the transcriptome abundance between case and control groups. The microarrays technology is instrumental for this task, allowing the expression levels (abundance of messenger ribonucleic acid (mRNA) molecules) of tens of thousands of genes to be measured simultaneously. The high-dimensional data thus generated demands for computational tools able to remove genes that generate uninformative signals and to select the genes that best discriminate between groups.

This chapter presents an introduction to microarray technology in Section 2.1 and an overview of the fundamental steps of a gene expression profiling experiment in Section 2.2. This is followed by a review of the statistical methods used to select differentially expressed genes in Section 2.3 and a survey of the supervised and unsupervised feature selection methods widely used in machine learning and bioinformatics in Section 2.4. Concluding remarks are given in Section 2.5. Fundamental concepts of genetics introduced in this chapter and used throughout the thesis are discussed in more detail in Appendix A.

The supervised feature selection methods presented in this chapter address the problem of class comparison while the unsupervised approaches serve the task of removing noisy features. Unsupervised feature selection methods searching for the subspaces where data points cluster (class discovery) are outside the scope of this thesis. Comprehensive reviews of feature selection methods for clustering are provided by a number of authors including Parsons et al. (2004), Kriegel et al. (2009), Dash and Koot (2009) and Alelyani et al. (2013).

2.1 Introduction to microarray technology

The first step towards understanding the dynamics governing gene regulation was made by single-gene studies (Guan et al., 2010). The purpose of these studies was to characterize the selective activation and the functional role of individual genes. Recently, genome-wide-scale studies of genetic regulation became possible thanks to technological advances that enabled a global assessment of the cell's transcriptional activity (Fryer et al., 2002). In particular, the microarray technology allowed for the expression of thousands of genes to be measured simultaneously using high-density and massively-parallel chips (Heller, 2002).

The microarray consists of oligonucleotides or polymerase chain reaction (PCR) products generated from purified complementary DNA (cDNA) templates, immobilized on a solid surface in an ordered arrangement (Falciani, 2007). Gene expression profiling using microarrays relies on nucleic acid hybridization – the ability of single stranded nucleic acids to bind to complementary sequences (Strachan and Read, 2011). Complementarity between the nucleic acids immobilized on the array (probes) and the fluorescently labelled nucleic acids in the experimental sample (targets) allows for parallel quantification of the transcript abundance. Probes are incorporated on the array based on their sensitivity (strong complementarity with the target sequence) and specificity (absence of near-complementarity with non-target sequences) (Draghici et al., 2006, Kane et al., 2000). Variation in sensitivity and specificity among microarrays can be tracked down to fabrication technology.

Microarrays are manufactured using two different technologies: spotting PCR products of cDNA templates on a glass surface (glass slide cDNA microarrays) (Duggan et al., 1999) or in situ synthesis of oligonucleotides (high-density oligonucleotide microarrays) (Lipshutz et al., 1999). The first manufacturing method is usually adopted within single facilities or laboratories; the second method is adopted by commercial companies. In what follows, the fabrication protocols of the glass slide cDNA microarrays and high-density oligonucleotide microarrays are presented with emphasis on the Affymetrix GeneChip format. Advantages and disadvantages of each technology are further discussed.

2.1.1 Glass slide cDNA microarrays

The first step in cDNA microarrays fabrication consists of selecting the templates for the probes (cDNA fragments of hundreds to thousands bases long) from genomic libraries. The templates are cloned and amplified using PCR. In the next step, the glass slide is coated with chemicals that restrict the spread of spotted probes and facilitate their immobilization to the surface. Purified PCR products representing specific genes are later deposited in a matrix format. This is accomplished through highly accurate contact printing methods (physical contact between the metal pin of a robotic arm and the slide surface) or non-contact printing methods (ink-jet array printers). Finally, the glass slide is dried and kept at room temperature.

Glass slide cDNA microarrays are particularly appealing for their reduced fabrication costs. They are also highly versatile platforms for gene expression profiling studies, allowing customization in terms of the probes that are arrayed on the slide. cDNA microarrays have high sensitivity but low specificity. Their reproducibility is affected by problems inherent to PCR product concentration and quality of PCR clones (Järvinen et al., 2004).

2.1.2 High-density oligonucleotide microarrays

The first step in high-density oligonucleotide microarrays fabrication consists of attaching covalent linker molecules with photolabile protecting groups on a quartz wafer. Light directed through a photolithographic mask produces deprotection at specific locations on the quartz. Next, a solution of single protected nucleotides is incubated with the quartz and chemical coupling occurs at the deprotected sites. Uncoupled nucleotides are washed away and the process is repeated by changing the photolithographic mask and the solution of nucleotides until oligonucleotides with desired lengths are synthesized (Pease et al., 1994).

The Affymatrix GeneChip format uses 25-mer oligonucleotides (25 bases in length) representing unique sequences of genes. Target abundance is measured using a collection of 11-20 probe pairs, called a probe-set. Each probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe which is obtained from the PM probe by changing the middle nucleotide. The MM probes quantify background and non-specific hybridization thus allowing PM values to be corrected for hybridization artefacts. Multiple probe sets can interrogate the same gene. This level of redundancy enables false-positives detection and improves signal-to-noise ratios (Lipshutz et al., 1999). A schematic representation of the Affymetrix GeneChip is shown in Figure A.5 of Appendix A.

High-density oligonucleotide microarrays have high reproducibility. Consistency in fabrication protocols guarantees that the same gene profiling platform is used across experiments. Oligonucleotides have high specificity but low sensitivity. However, Affymetrix GeneChips achieve an optimal balance between high sensitivity and specificity through MM control probes. This additional level of information improves the quality of the gene expression measurements thus increasing the robustness of genome-wide expression studies results (Affymetrix, 2001).

One disadvantage of this technology is affordability – oligonucleotide microarrays are expensive platforms. The costs of the specialized equipment necessary for conducting a gene expression profiling experiment make this technology unavailable for many average size laboratories.

2.2 Gene expression profiling using microarrays

Gene expression profiling experiments typically involve the following steps:

1. Experimental design
2. Sample preparation
3. Hybridization and washing
4. Image acquisition (scanning)
5. Data pre-processing
6. Data analysis

These steps are discussed in more detail below.

2.2.1 Experimental design

Careful formulation of the questions to be addressed or hypothesis to be tested lies at the heart of properly conducted gene expression studies (Yang and Speed, 2002). The aims of the experiment influence the choice of the microarray platform (cDNA arrays, oligonucleotides arrays), the number of technical replicates (arrays hybridized with the same sample) or biological replicates (arrays hybridized with different individual samples from the population being studied) needed, and the tools necessary for data pre-processing and analysis. Crafting the experimental setup around clear objectives maximizes the information leveraged from the data whilst minimizing the efforts and the costs (Stekel et al., 2003).

2.2.2 Sample preparation

Sample preparation consists of extracting mRNA from frozen tissue (freezing prevents further degradation or production of RNA). Isolation of mRNA can be performed directly from the cells or through purification of total RNA extracted. After isolation, mRNA is converted to cDNA using reverse transcriptase (RT) and labelled with a fluorescent dye. Sample preparation for Affymetrix GeneChips involves additional steps: isolated mRNA is linearly amplified before the RT reaction and the resulted cDNA is used to produce anti-sense complementary RNA (cRNA) which is fluorescently labelled. Detailed technical specifications of sample preparation are provided by a number of authors including Mahadevappa and Warrington (1999) and Hegde et al. (2000).

2.2.3 Hybridization and washing

The labelled sample is poured onto the microarray and incubation follows (at high salt concentrations, presence of formamide and temperature of 42° C, or presence of an aqueous solution and temperature of 65°C) to promote probe-target hybridization. After the microarray is removed from the incubation chamber, stringent washes are performed (at lower salt concentrations and room temperature) to remove non-specific and weak bindings. For high-quality gene expression profiling, the hybridization and washing solution must be evenly mixed and spread on the microarray (Freeman et al., 2000).

2.2.4 Image acquisition

The microarray is scanned to produce an image containing the fluorescence intensity of each probe. This stage consists of exciting the dye using a laser beam thus generating signals dependent on the quantity of target sample hybridized with each probe on the chip. Ideally, fluorescent signals should come only from targets that hybridized to their complementary probes. In practice, unwashed targets adhering to the slide, non-specific bindings, and other chemicals used for array preparation, hybridization and washing also generate fluorescent signals (Scharpf et al., 2007). These residual signals (background noise) need to be accounted for as they affect probe value calculations.

2.2.5 Data pre-processing

Processing of the raw intensities generated during scanning consists of the following steps:

1. Background correction
2. Normalisation
3. Summarization (Affymetrix GeneChips only)

These steps are discussed in more detail below.

Background correction

The purpose of background correction is to adjust the probe intensities by removing the background noise. Methods for estimating and subtracting the background noise are microarray technology dependent.

For cDNA microarrays, an unbiased estimate of the probe-specific background can be calculated from the local neighbourhood surrounding the probe by taking the mean of the pixel intensity values. The standard correction approach consists of subtracting the local background from the raw intensity values (Yin et al., 2005). To correct for negative adjusted intensities resulting from this approach, Edwards (2003) proposed a method that uses subtraction only when the difference exceeds a certain threshold; otherwise a smooth monotonic function that is linear on the log-scale with respect to the raw probe intensities is used. Yang et al. (2002) proposed sampling the probe-specific background at the probe nominal location from a background image estimated using morphological operations. This approach results in less variable background adjusted intensities. Ritchie et al. (2007) provides a comparison of background correction methods for cDNA microarrays and their effect on identifying differentially expressed genes.

For oligonucleotide microarrays, the high-density of the chip prohibits the use of local neighbourhood methods. Instead, the MM control probes, which are adjacent to their PM probes, can be used for background correction. Two widely used background correction methods are integral parts of Affymetrix microarray suite 5.0 (MAS 5.0) (Affymetrix, 2002) and robust multi-array average (RMA) (Irizarry et al., 2003).

The MAS 5.0 background correction method splits the chip into 16 rectangular zones of equal size. Zone-specific background and noise are taken as the mean and standard deviation of the lowest 2% probe intensities. For each probe, background and noise are calculated as weighted averages of the zone-specific background and noise signals, respectively. Probe intensities are adjusted by subtracting their background unless the difference leads to a value less than the probe-specific noise, in which case the raw intensity is replaced by the noise.

The RMA background correction assumes the observed PM values within a probe set follow a linear additive model consisting of an exponentially distributed true signal and a normally distributed background. The adjustment procedure consists of replacing the PM intensity with the expected value of the true signal (which depends on the particular PM being adjusted, PM values above their mode and MM below their mode).

Normalization

The purpose of normalization is to adjust for systematic biases in sample collection (Fentz et al., 2004), hybridization (Han et al., 2006), labelling (Gregory Cox et al., 2004) and scanning (Bengtsson et al., 2004) so that unbiased comparisons can be made across technical and biological replicates. Robust and widely used normalization methods vary from global normalization (Stafford, 2012) to quantile normalization (Bolstad et al., 2003) and lowess normalization (Smyth and Speed, 2003).

Global normalization consists of multiplying the probe intensities of each array with an array-specific scaling factor to adjust the mean or the median intensity values of the arrays; quantile normalization consists of transforming the distributions of the array-specific intensities to have identical statistical properties; lowess normalization consists of removing the dependency of the $\log_2$ ratio values on the intensity using locally weighted linear regression analysis. Normalization is a growing field which generated a consistent amount of literature. Comprehensive reviews are provided by a number of authors including Quackenbush (2002) and Bilban et al. (2002).

Summarization

The purpose of summarization is to estimate probe set expression levels from associated probe-pair intensities. Summarization methods are divided into two major categories: single-chip methods (summarization is performed for each array independently) and multi-chip methods (summarization is performed by borrowing information across arrays).

One of the most widely used single-chip methods is MAS 5.0 which provides for each probe set a robust average of the differences between the PM values and the associated ideal mismatch (IM) values. The IM value is equal to the MM value if MM<PM and is slightly less than the PM value (according to some adjustment formula) if MM>PM. Subtraction of MM values from PM values can induce bias in probe set level summarization (Irizarry et al., 2003) as MM probes also measure specific hybridization (Wu et al., 2004b) and can often exceed the values of the PM probes (effect observed in about one third of the probe-sets) (Naef et al., 2002). In addition, the PM and MM intensities vary in probe-specific ways (Li and Wong, 2001). These factors confounding accurate probe-set level summarization are not accounted for by MAS 5.0.

Multi-chip methods such as RMA, factor analysis for robust microarray summarization (FARMS) (Hochreiter et al., 2006) and multi-chip modified gamma

Model for Oligonucleotide Signal (multi-mgMOS) (Liu et al., 2005) adjust for the systematic differences in probe hybridization affinities. RMA assumes the $\log_2$ transformed PM intensities follow a linear additive model consisting of the true probe set level, probe pair affinity effect and independent and identically distributed error term. FARMS assumes the $\log_2$ transformed PM intensities follow a factor analysis model consisting of the normally distributed true signal and normally distributed noise. Finally, multi-mgMOS uses a probabilistic model with gamma distributed PM and MM values that additionally accounts for the MM probes measuring specific hybridization. This method returns the estimated probe set levels and the uncertainties (standard errors) around these estimates, which are particularly useful for downstream analyses adopting a Bayesian framework.

2.2.6 Data analysis

The pre-processed data is analysed to extract biologically-relevant information. Analysis methods are purpose-specific and the choice of the algorithms influences the quality of the results which are usually based on a list of selected genes. These genes need to be enriched with functional information (molecular functions and biological processes associated with the genes) in order to strengthen the understanding of the biological process being studied.

Functional annotation can be performed using Gene Ontology (Consortium et al., 2004) which unifies functional information in highly interconnected structures and under a common nomenclature. If valuable knowledge was gained by addressing the questions or testing the hypothesis of the experiment, the data is published following carefully defined standards (Brazma et al., 2001) into a public repository (Gardiner-Garden and Littlejohn, 2001).

2.3 Statistical hypothesis testing in microarray experiments

Statistical hypothesis testing is concerned with making inference about a statistical population given information from a random statistical sample (measured data). In the arena of differential gene expression studies, statistical hypothesis testing consists of assessing the quality of the evidence provided by the data against the claims of no differential expression.

Section 2.3.1 introduces the terminology of statistical testing and presents the single hypothesis testing scenario. This will lay the foundations for a discussion of multiple hypothesis testing and a review of multiple comparison correction methods in Section 2.3.2. Finally, a discussion on the relation between statistical significance and biological relevance is presented in Section 2.3.3, highlighting the need to supplement statistical findings with information from effect size measures.

2.3.1 Single hypothesis testing

In a single hypothesis testing scenario, given a statistic that quantifies an effect (or difference) in the sample data through a numerical summary T, one tests a null hypothesis (absence of an effect) against an alternative hypothesis (presence of the effect) at a chosen significance level α (commonly 0.05 or 0.01). If the p-value i.e. the probability (under the null hypothesis) of observing a statistic at least as extreme as T is less than α, the null hypothesis is rejected in favour of the alternative hypothesis; otherwise, one fails to reject the null hypothesis or equivalently concludes the data provides no reliable evidence against the null hypothesis.

Statistics commonly applied in differentially gene expression studies are divided into two major categories: parametric methods (distributional assumptions are made about the data) and non-parametric methods (no distributional assumptions are made). The Welch t-test (Welch, 1947) and the Wilcoxon rank-sum test (Hollander et al., 2013) are two widely used parametric and non-parametric methods, respectively. These methods are discussed in more detail in Chapter 4.

The known sampling distribution of these statistics under the null hypothesis allows for direct p-values calculation. When the sampling distribution is not known a priori, it can be estimated using random permutation methods (Dudoit et al., 2002b, Pan, 2003). These methods are robust against outliers and represent powerful alternatives to using known sampling distributions in studies with small sample size (Saeys et al., 2007). Chen et al. (2005) and Pan (2002) provide comprehensive reviews of the test statistics used in microarray studies.

There are two types of errors that can be committed when performing statistical hypothesis testing: type I error and type II error. A type I error (false positive) is committed when rejecting a true null hypothesis; a type II error (false negative) is committed when failing to reject a false null hypothesis. In the context of differential gene expression, type I error corresponds to erroneously calling a gene differentially expressed; a type II error corresponds to calling a gene non-differentially expressed when in fact it is differentially expressed. The false positive rate is the rate at which true null hypotheses are called significant and equals the significance level α. The false negative rate is the rate at which true alternative hypotheses are called null. The single hypothesis testing scenario described above aims to control the false positive rate conditional on maximizing the power (1-false negative rate) of detecting an effect.

2.3.2 Multiple hypothesis testing

In a typical microarray experiment tens of thousands of genes are tested for differential expression. Controlling the false positive rate at a common significance level α when simultaneously testing multiple hypotheses increases the chance of committing type I errors (Dudoit et al., 2003). The significance level α multiplied by the number of hypothesis tested represents a conservative upper bound for the expected number of false positives. This bound is too large for microarray data.

By increasing the stringency of individual tests (lowering α) in an attempt to decrease the expected number of Type I errors one instead decreases the power of the tests thus lowering the chances of detecting true differentially expressed genes. Since control of the false positive rate in the single hypothesis testing sense is unsatisfactory for multiple hypotheses testing, several generalizations of the Type I error rate were proposed together with procedures that control these error rates at a desired level α. These compound error measures and associated controlling procedures are summarized within the setup of a multiple comparison experiment presented below.

Consider testing simultaneously m null hypotheses H_i, $i = 1, \ldots, m$, at a common significance threshold α. Assume the hypotheses are ordered in ascending order of their p-values p_i, which are associated with the independent test statistics T_i. The possible outcomes of the multiple comparison tests are listed in Table 1. Specifically, V represents the number of false positives, U represents the number of true positives, T represents the number of false negatives, S represents the number of true negatives and R represents the number of rejected null hypotheses. The quantities V, U, T and S are unobservable random variable; R is an observable random variable; the number of true null hypotheses m_0 and the number of true alternative hypotheses $m_1 = m - m_0$ are unknown parameters.

Table 2.1: Possible outcomes from m hypothesis tests

Hypothesis	Called significant	Called non-significant	Total
Null true	V	S	m_0
Alternative true	U	T	m_1
Total	R	$m - R$	m

Historically, the first type I error rate proposed for multiple hypothesis testing was the family-wise error rate (FWER) (Shaffer, 1995):

$$FWER = P(V \geq 1) \tag{2.1}$$

FWER represents the probability of rejecting at least one true null hypothesis out of all hypotheses tested. Control of the FWER at a significance level α can be achieved using various sequential p-value procedures reviewed by Dudoit et al. (2003) and Nichols and Hayasaka (2003). Presented below, in increasing order of their power, are three widely known representatives.

- The Bonferroni procedure (Simes, 1986) controls the FWER as follows:

$$\text{reject } H_j \text{ for } j = 1, \ldots, \max \left\{ i | p_i \leq \frac{\alpha}{m} \right\}) \tag{2.2}$$

- The Holm procedure (Holm, 1979) controls the FWER as follows:

$$\text{reject } H_j \text{ for } j = 1, \ldots, \min \left\{ i | p_i \geq \frac{\alpha}{m - i + 1} \right\} - 1 \tag{2.3}$$

  If there is no i for which the p-value inequality is satisfied, all hypotheses are rejected.

- The Hochberg procedure (Hochberg, 1988) controls the FWER as follows:

$$\text{reject } H_j \text{ for } j = 1, \ldots, \max \left\{ i | p_i \leq \frac{\alpha}{m - i + 1} \right\} \tag{2.4}$$

  If there is no i for which the p-value inequality is satisfied, mo hypotheses are rejected.

Controlling the FWER is concerned with guarding against the occurrence of any false positives. While this strict criterion is appropriate for studies where few true alternatives are expected, it is far too stringent for differential gene expression studies where the presence of an effect is manifested through more than one or two genes. In general, when controlling the FWER the power decreases with increasing number of hypothesis being tested (Storey, 2002).

Instead of controlling the probability of making at least one Type I error, one may be interested in making as many significant findings as possible while controlling the proportion of false positives. This problem was address by Benjamini and Hochberg (1995) who proposed controlling the false discovery rate (FDR):

$$FDR = E\left(\frac{V}{R} \middle| R > 0 \right) P(R > 0) \tag{2.5}$$

as follows:

$$\text{reject } H_j \text{ for } j = 1, \ldots, \max \left\{ i | p_i \leq \frac{i}{m} \alpha \right\} \tag{2.6}$$

The FDR represents the expected proportion of Type I errors among the rejected null hypotheses times the probability of making at least one rejection. Controlling the FDR leads to a substantial gain in power over Bonferroni and Hochberg procedures for FWER control. This gain increases when the number of tested hypothesis or the number for true alternative hypotheses also increase (Benjamini and Hochberg, 1995). Storey (2002) signals that caution must be taken when controlling the FDR at a significance level α as in the case when null hypothesis are rejected, one really controls the FDR at level $\alpha / P(R > 0)$.

The previous methods relied on setting the significance level α prior to estimating the rejection region for the p-values. The appropriateness of this approach is conditional upon the number of rejected hypothesis. Additionally, none of these methods uses information about the number of true null hypotheses m_0 in the data. To address these issues, Storey (2002) proposed fixing the rejection region γ and then estimating the significance level α through control of the positive false discovery rate (pFDR):

$$pFDR(\gamma) = E\left(\frac{V(\gamma)}{R(\gamma)}\middle| R(\gamma) > 0\right) \tag{2.7}$$

The pFDR cannot be controlled using sequential p-value approaches but needs to be estimated for the particular rejection region γ as follows:

$$\widehat{pFDR}_{\delta}(\gamma) = \frac{m\hat{\pi}_0(\delta)\gamma}{\#\{p_i \leq \gamma\}} \tag{2.8}$$

where

$$\hat{\pi}_0(\delta) = \frac{\#\{p_i > \delta\}}{(1-\delta)m} \tag{2.9}$$

with $\delta \in [0,1]$ representing a tuning parameter used to determine the estimated fraction of true null hypotheses $\hat{\pi}_0(\delta)$. The bias of $\hat{\pi}_0(\delta)$ is the smallest when $\delta \to 1$ (Storey, 2002). Equation (2.9) shows that pFDR control takes into account the information stored in the observed p-values to estimate m_0/m.

Within the pFDR framework, Storey (2003) proposed the q-value as an individual measure of significance analogous to the p-value. The q-value of the statistic T_i, given a set of nested rejection regions Γ containing T_i is:

$$q(T_i) = \inf_{\{\Gamma : T_i \in \Gamma\}} pFDR(\Gamma) \tag{2.10}$$

The q-value of a statistic gives the pFDR obtained when rejecting any statistic at least as extreme as the one observed among all the hypotheses; the p-value gives the false positive rate when rejecting any statistic at least as extreme as the observed statistic. A fundamental difference between the q-value and the p-value is that the former quantity accounts for the multiple comparisons carried out in parallel while the latter quantity does not. Thus the p-value carries no information about the set of more extreme statistics; the q-value directly provides a measure of strength for this set.

If the hypotheses are assumed to be independent, the q-value of a statistic T_i can be expressed in terms of the associated p_i as follows:

$$q(p_i) = \inf_{\gamma \geq p_i} pFDR(\gamma) \tag{2.11}$$

This allows for the q-values to be estimated as:

$$\hat{q}(p_i) = \inf_{\gamma \geq p_i} \widehat{pFDR}(\gamma) \tag{2.12}$$

Storey et al. (2004) proved the simultaneous conservative consistency of these estimates. This means that the estimated q-values are simultaneously greater than or equal to the true q-values for any rejection region. Thus, by calling significant statistics with q-value less than α one obtains a conservative point estimate for the pFDR at significance level α. When compared with the FDR control method of Benjamini and Hochberg (1995), pFDR control provided greater power (Storey, 2002).

The multiple adjustment procedures listed above rely on the independence between the statistics T_i. Although microarray data violate this simplifying assumption through correlated gene expression levels, positive results were reported in differential gene expression studies adopting the multiple adjustment framework (Dudoit et al., 2002b, 2003, Efron and Tibshirani, 2002, Storey and Tibshirani, 2003). Methods that account for the dependence structure among statistics were proposed by Tusher et al. (2001) for FWER control and by Benjamini and Yekutieli (2001) and Storey et al. (2004) for FDR control. However, the area of multiple hypotheses testing exploiting the distribution of the statistics is still in its infancy and much research is needed to lay solid theoretical foundations.

2.3.3 Statistical significance and biological relevance

A common criticism addressed to statistical hypothesis testing is that statistical significance doesn't necessarily translate into biological relevance (Lovell, 2013, Martínez-Abraín, 2008). The p-values don't offer a quantitative measure of the magnitude of an effect but a qualitative measure of evidence against the null hypothesis. Thus, genes statistically significant can have a biologically irrelevant effect size.

This raises the question of how to define biological relevance or important effect. There is no agreed upon definition and expert opinion is required to make the right judgement for the experiment at hand. Thompson (2001) argues that using benchmark values to define biological relevance would be no different than setting the significance level for statistical hypothesis testing. Nakagawa and Cuthill (2007) and Hentschke and Stüttgen (2011) recommend using measures of effect size as a supplement or alternative to hypothesis tests. A widely used measure of effect size is the Hedges' g metric (Hedges, 1981) defined below.

Let $\mathbf{y}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{y}_2 \in \mathbb{R}^{n_2}$ denote independent samples from two populations. The Hedges' g metric is defined as:

$$g(\mathbf{y}_1, \mathbf{y}_2) = \frac{\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2}{\sigma_*} \tag{2.13}$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are the means of $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, while the pooled standard deviation σ_* is defined as:

$$\sigma_* = \sqrt{\frac{(n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\,\sigma_2^2}{n_1 + n_2 - 2}} \tag{2.14}$$

where σ_1^2 and σ_2^2 denote the variances the two samples. The metric (2.13) represents a biased estimator for the standardized mean difference between two populations (Nakagawa and Cuthill, 2007). An unbiased estimator is given by:

$$g_* = g\left(1 - \frac{3}{4\,(n_1 + n_2 - 2) - 1}\right) \tag{2.15}$$

Intuitively, the Hedges' g metric quantifies the number of standard deviations necessary to move one distribution over the other such that their means match.

2.4 Feature selection in machine learning

Feature selection methods are based on two approaches: individual evaluation (Guyon and Elisseeff, 2003) and subset evaluation (Dash and Liu, 1997).

Individual evaluation (variable ranking) consists of ranking features according to independent criteria such as correlation and mutual information (Guyon and Elisseeff, 2003). From the top of the ranking list, a subset of user-supplied cardinality is selected for further analysis. Computationally, this approach is very appealing for its linear time complexity. However, it focuses on feature relevance only and disregards feature redundancy. Guyon and Elisseeff (2003) point out that redundant features provide no additional information in classification problems while Yu and Liu (2004) and Saeys et al. (2007) suggest that redundant features may decrease classification performance.

Subset evaluation (feature subset selection) consists of identifying the optimal subset of features according a suitable criterion function that accounts for both feature relevance and feature redundancy (Blum and Langley, 1997). Subset methods consist of three major components:

- a generation procedure
- an evaluation function
- a stopping criterion

The generation procedure is essentially a search strategy that supplies candidate feature subsets to the evaluation function. There are three major families of search strategies: exhaustive search, sequential search and random search. Exhaustive search methods generate all the possible subsets; sequential search methods (Somol et al., 2010) add features iteratively to an initial empty set (sequential forward selection), remove features iteratively from the initial full set of features (sequential backward selection) or alternate between adding and removing features (sequential forward floating selection, sequential forward backward selection); random search methods (Brassard and Bratley, 1996, Liu and Setiono, 1996) generate subsets in a completely random manner. Subset methods adopting the exhaustive search identify the optimal subset according to the evaluation function but are computationally prohibitive for large datasets; subset methods adopting sequential or random search trade optimality for computational efficiency.

The evaluation function uses data intrinsic measures (distance, consistency, dependence and information) (Molina et al., 2002) or the classifier error rate (Kohavi and John, 1997) to measure the goodness of the feature subsets produced by the generation procedure. A subset replaces the previous best if it has a better score according to the evaluation function. Subset methods adopting data intrinsic measures are called filters while subset methods adopting the classifier error rate are called wrappers. Methods combining the characteristics of filters and wrappers are known as embedded methods.

The stopping criterion defines when the generation procedure stops producing candidate feature subsets. Stopping criteria include: bounding the number of selected features or finding a subset whose score cannot be improved by adding ot removing features. Subset methods adopting the first stopping criterion select feature subsets of user-specified cardinality; subset methods adopting the second stopping criterion optimize the number of features for the given evaluation function.

Section 2.4.1 proposes a review of variable ranking methods used in bioinformatics to remove noisy and non-informative genes. Note that within the bioinformatics community these methods are referred to as unsupervised filters, although they operate based on individual evaluation than subset evaluation. By abuse of terminology, for the remaining of the thesis, these methods will be called unsupervised filters even though they don't satisfy the definitions of filters proposed in machine learning.

The overview of the unsupervised filters will be followed by comprehensive reviews of the state-of-the-art filters (Section 2.4.2), wrappers (Section 2.4.3) and embedded methods (Section 2.4.4) widely used in machine learning and pattern recognition. Specifically, the survey of the filter methods will focus on approaches adopting information theoretic measures; the survey of wrapper methods will focus on approaches based on three widely used classifiers: naive Bayes (NB) (Duda and Richard, 1973), linear discriminant analysis (LDA) (Webb, 2003) and support vector machine (SVM) (Vapnik, 2000). Special attention will be given to the SVM classifier for two main reasons. Firstly, the SVM classifier outperforms other state-of-the-art classification tools (including NB and LDA) on microarray data (Lee et al., 2005). Secondly, its mathematical formulation will be used to present the embedded methods (based on SVM) within a unified optimization framework.

2.4.1 Unsupervised filtering approaches for microarray data

Affymetrix GeneChips can measure the transcriptional activity of more than 33,000 genes simultaneously. Practically, only 10,000 – 15,000 genes are actively expressed in most tissues (Tang et al., 2004). Most of these genes are housekeeping genes i.e. genes that code for proteins necessary for the cell and that are expressed at relatively constant levels under any condition (Gohlmann and Talloen, 2010). This leaves only a small fraction of genes potentially differentially expressed between conditions (Gohlmann and Talloen, 2010).

The genes that are not expressed in a tissue can still generate fluorescent signals due to non-specific bindings. Inclusion of these noisy genes and housekeeping genes in differential expression studies increases the risk of reporting false positives (Dudoit et al., 2003) and reduces the power of multiple testing adjustment procedures (Talloen et al., 2007). Conversely, discarding these genes increases substantially the number of discoveries (rejected null-hypothesis) and detection power (Bourgon et al., 2010a).

The next three sections provide an overview of the key representatives from two classes of unsupervised gene filtering methods. The methods in the first class (overall variance and overall mean) work with probe set expression levels and are compatible with most microarray platforms; the methods in the second class (Absent/Marginal/Presents calls and Informative/non-informative calls) work with probe pair intensities and are particularly suited for Affymetrix GeneChip microarrays. In the fourth section, a discussion of the marginal independence between the filter and the test statistics used to select differentially expressed genes is presented.

Overall variance and overall mean filtering

Filtering by overall variance consists of removing a user-specified percentage of the genes with the lowest variance, measured across all arrays. This simplistic approach has powerful features: Hackstadt and Hess (2009) show that filtering by overall variance increases the number of discoveries and the statistical power to detect differentially expressed genes while Bourgon et al. (2010b) argue that removing genes with the lowest 50% variance provides a powerful alternative to platform-specific filtering techniques. An adaptive method for selecting the stringency of the filter was proposed by Marczyk et al. (2013).

Filtering by overall mean consists of removing a user-specified percentage of the genes with the lowest mean, measured across all arrays. Compared to overall variance, the overall mean filter is less effective leading to substantially fewer detections of differentially expressed genes at higher stringencies (Bourgon et al., 2010a). This effect is due to the filter depleting the set of differentially expressed genes with small overall mean. Additionally, genes that are expressed in one group and unexpressed in another can also have low overall mean and therefore risk to be erroneously removed.

Absent/Marginal/Present calls

The Absent/Marginal/Present (A/M/P) calls (Mei et al., 2002) is an integral part of MAS 5.0 summarization algorithm. Each probe set is declared absent, marginally present or present based on the *p*-value of a non-parametric statistical test performed on the PM and MM intensity values. Specifically, for the *i*th probe set, a discrimination score is computed for each probe pair *j* as follows:

$$R_{ij} = \frac{PM_{ij} - MM_{ij}}{PM_{ij} + MM_{ij}} \tag{2.16}$$

A one sided Wilcoxon signed-rank test is conducted with the null and alternative hypotheses:

$$\begin{cases} H_0 : \text{median}\left(R_{ij}\right) = \tau \\ H_1 : \text{median}\left(R_{ij}\right) > \tau \end{cases} \tag{2.17}$$

where the user-adjustable parameter τ is set to the default value of 0.015. The p-value of the test statistic is compared against two significance levels α_1 and α_2 with default values of 0.4 and 0.6, respectively. The probe set is declared present if $p < \alpha_1$, marginally present if $\alpha_1 < p < \alpha_2$, and absent if $p > \alpha_2$.

The A/M/P calls can be used to filter probe sets across arrays. A typical setting is to retain for further analysis probe sets that where called present in at least one array (Mei et al., 2002). However, this approach decreases the statistical power to detect differentially expressed genes (Talloen et al., 2007). McClintick and Edenberg (2006) proposed removing probe sets that are not called present in at least half of the arrays in any of the treatment groups. Bourgon et al. (2010a) remarks that since this method uses information about the class labels of the arrays, it can lead to overly optimistic results when assessing differential gene expression using statistical methods.

Informative/non-informative calls

The informative/non-informative (I/NI) calls (Talloen et al., 2007) expands upon the FARMS summarization method to remove probe sets with technical variation exceeding its biological variation (measured across arrays). Specifically, for the ith probe set, FARMS assumes the zero mean normalized vector of $\log_2$ PM intensities $\mathbf{x}_i$ depends on the true log-concentration of mRNA z_i via:

$$\mathbf{x}_i = \boldsymbol{\nu}_i z_i + \boldsymbol{\epsilon}_i \tag{2.18}$$

with

$$z_i \sim \mathcal{N}\left(0, 1\right), \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_i\right), \boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{I}_n \tag{2.19}$$

where $\boldsymbol{\nu}_i \in \mathbb{R}^n$ models the probe-specific factors, $\boldsymbol{\epsilon}_i \in \mathbb{R}^n$ models the independent homoscedastic probe-specific noise, and n is the number of probes in a probe set. The I/NI calls accesses the variance of z_i given the measurements x_i and removes the ith probe set if

$$\text{var}\left(z_i | \mathbf{x}_i\right) = \boldsymbol{\nu}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\nu}_i < 1 \tag{2.20}$$

Note that $\boldsymbol{\nu}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\nu}_i$ represents the ratio between the biological signal of the ith probe set measured by $\boldsymbol{\nu}_i^\top \boldsymbol{\nu}_i$ and its technical noise measured by $\boldsymbol{\Sigma}_i$ through σ_i^2.

The I/NI call for a probe set relates to the consistency of the corresponding probes to measure the same biological signal. If the probe measurements are highly correlated across arrays, the probe set is called informative and kept for further analyses. If more than 11 probes show no consistent probe behaviour, the probe set is called non-informative.

The I/NI calls filter has the advantage of being independent of user-supplied parameters. Additionally, it increases the statistical power to detect differentially expressed genes (Talloen et al., 2007).

The marginal independence criterion

A necessary (but not sufficient) condition to obtain an increase in detection power is the marginal independence under the null hypothesis between the filter and test statistic used to assess differential gene expression (Bourgon et al., 2010a). This criterion requires the unconditional null distributions of the test statistic before filtering and the conditional null distributions after filtering to match. Violations of this criterion occurs when using supervised filtering prior to the statistical test (McClintick and Edenberg, 2006) or moderated statistic after unsupervised filtering (Scholtens and Von Heydebreck, 2005) . These combinations results in overly optimistic p-values and reduced detection power.

The overall variance, overall mean and I/NI calls are marginally independent of t-test and Wilcoxon rank sum test statistics. While marginal independence preserves the correct size of the p-values, their correlation structure can change which in turn may impact the performance of the multiple adjustment procedures. For microarray data, these effects are negligible (Bourgon et al., 2010a).

2.4.2 Filters based on information theory

Information theory has been a fruitful ground for the development of subset selection methods (Brown et al., 2012). Two central concepts in information theory used to define compound measures of relevance and redundancy are: *entropy* and *mutual information*. These quantities together with their conditional formulation are defined below.

Definition 1. The entropy of a discrete random variable X with support $\mathcal{X}$ and probability density function $P(X)$ is:

$$H(X) = -\sum_{x\in\mathcal{X}} P(x)\log P(x) \tag{2.21}$$

$H(x)$ measures the uncertainty associated with X.

Definition 2. The conditional entropy of discrete random variable X with support $\mathcal{X}$ and probability density function $P(X)$ given the discrete random variable Y with support $\mathcal{Y}$ and probability density function $P(Y)$ is:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} P(y) \sum_{x \in \mathcal{X}} P(x|y) \log P(x|y) \tag{2.22}$$

$H(X|Y)$ measures the uncertainty around X once we learned about Y.

Definition 3. The mutual information between the discrete random variables X and Y with joint probability density function $P(X,Y)$ is:

$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{2.23}$$

Alternatively, $I(X;Y)$ can be expressed using the entropy of X and conditional entropy of X given Y as follows:

$$I(X;Y) = H(X) - H(X|Y) \tag{2.24}$$

$I(X;Y)$ measures the amount of information shared by X and Y.

Definition 4. The conditional mutual information between the discrete random variables X and Y given the discrete random variable Z is:

$$I(X;Y|Z) = \sum_{z \in \mathcal{Z}} P(z) \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P(x,y|z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)} \tag{2.25}$$

Alternatively, $I(X;Y|Z)$ can be expressed using the conditional entropy of X given Y and Z as follows:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{2.26}$$

$I(X;Y|Z)$ measures the amount of information shared by X and Y once we learned about Z.

In a feature selection problem, given the set of features $\mathcal{S}_m = \{X_1, \ldots, X_m\}$ and the class variable $\mathcal{C}$, one is faced with the task of selecting an optimal subset $\mathcal{S}_k$ of cardinality k according to some evaluation function. In terms of mutual information,

S_k can be found by solving the optimization problem:

$$\arg\max_{\mathcal{S}_k \in \mathbb{S}_K} I(\mathcal{S}_k; \mathcal{C}) \tag{2.27}$$

where $\mathbb{S}_K$ is the space of feature subsets of cardinality k and $I(\mathcal{S}_k; \mathcal{C})$ is the multivariate (joint) mutual information between $\mathcal{S}_k$ and $\mathcal{C}$. Estimation of $I(\mathcal{S}_k; \mathcal{C})$ can be approached using non-parametric methods such as histograms and kernel methods (Beirlant et al., 1997). For datasets with large number of features exceeding the number of observations, these estimation methods become impractical. To avoid this problem, several iterative approximations of (2.27) were proposed. Specifically, given $\mathcal{S}_{i-1}$, the ith feature can be selected from $\mathcal{S}_m \setminus \mathcal{S}_{i-1}$ according to one of the criteria:

- Mutual information based feature selection (MIFS) Battiti (1994):

  $$\max_{X_i \in \{\mathcal{S}_m \setminus \mathcal{S}_{i-1}\}} \left[I(X_i; \mathcal{C}) - \beta \sum_{X_j \in \mathcal{S}_{i-1}} I(X_i; X_j) \right] \tag{2.28}$$

  where β is a user-supplied parameter.

- Minimum redundancy - maximum relevance (mRMR) (Ding and Peng, 2005) with mutual information quotient (MIQ) formulation:

  $$\max_{X_i \in \{\mathcal{S}_m \setminus \mathcal{S}_{i-1}\}} \left[I(X_i; \mathcal{C}) / \left(\frac{1}{i-1} \sum_{X_j \in \mathcal{S}_{i-1}} I(X_i; X_j) \right) \right] \tag{2.29}$$

  which is equivalent to the MIFS criterion where β was set to $1/(i-1)$.

- MRMR (Ding and Peng, 2005) with mutual information difference (MID) formulation:

  $$\max_{X_i \in \{\mathcal{S}_m \setminus \mathcal{S}_{i-1}\}} \left[I(X_i; \mathcal{C}) - \frac{1}{i-1} \sum_{X_j \in \mathcal{S}_{i-1}} I(X_i; X_j) \right] \tag{2.30}$$

- Conditional mutual information maximization (CMIM) (Fleuret, 2004):

  $$\max_{X_i \in \{\mathcal{S}_m \setminus \mathcal{S}_{i-1}\}} \left[I(X_i; \mathcal{C}) - \max_{X_j \in \mathcal{S}_{i-1}} \left(I(X_i; X_j) - I(X_i; X_j | \mathcal{C}) \right) \right] \tag{2.31}$$

  which accounts for the conditional redundancy $I(X_i; X_j | \mathcal{C})$

- Joint mutual information (JMI) (Brown, 2009, Yang and Moody, 1999):

$$\max_{X_i \in \{\mathcal{S}_m \backslash \mathcal{S}_{i-1}\}} \left[I(X_i;\mathcal{C}) - \frac{1}{i-1} \sum_{X_j \in \mathcal{S}_{i-1}} \big(I(X_i;X_j) - I(X_i;X_j|\mathcal{C}) \big) \right] \tag{2.32}$$

  which extends the mRMR criterion to account for conditional redundancy

- Double input symmetrical relevance (DISR) (Meyer et al., 2008):

$$\max_{X_i \in \{\mathcal{S}_m \backslash \mathcal{S}_{i-1}\}} \left[\sum_{X_j \in \mathcal{S}_{i-1}} \frac{I(X_iX_j;\mathcal{C})}{H(X_iX_j\mathcal{C})} \right] \tag{2.33}$$

  where X_iX_j denotes the joint random variable of X_i and X_j .

These criteria balance relevance measured by $I(X_i;\mathcal{C})$ with redundancy measured by $I(X_i;X_j)$ and conditional redundancy measured by $I(X_i;X_j|\mathcal{C})$. Meyer et al. (2008) and Brown et al. (2012) show on various datasets that mRMR and JMI outperform the remaining criteria.

2.4.3 Wrapper methods

The NB classifier uses the Bayes theorem to compute the probability of each class given a multi-dimensional test instance, under the assumption that the features (dimensions) are conditionally independent given the class. This assumption states that each feature independently contributes to the probability of the test instance belonging to a certain class. In spite of this stringent and unrealistic assumption, the NB classifier competes in terms of predictive performance with more complex classification tools (Friedman et al., 1997a).

The wrapper approach based on NB was introduced by Langley and Sage (1994) who showed that the classifier guided by a forward selection search leads to improved classification performance and argued that this scheme is beneficial for applications involving correlated features. Kohavi and Sommerfield (1995) proposed a wrapper approach based on NB that uses compound operators to dynamically change the search topology. These compound operators were also used by Kohavi and John (1997), who additionally experimented with best-fit search and hill-climbing search to generate candidate feature subsets evaluated by the NB classifier. Cortizo and Giraldez (2006) proposed an improved NB wrapper approach that uses information about the correlation between features to guide the search for the optimal subset. Inza et al. (2002) used a NB wrapper approach with sequential forward selection for gene subset selection.

The LDA classifier performs classification in a one-dimensional space by projecting the multidimensional test instance on the direction that maximizes the separation between the two classes i.e. maximizes the distance between the means of the classes while minimizing the variance within the classes. The test instance is assigned to the class associated with the smallest distance between the projected test instance and the projected class mean.

Xiong et al. (2001a) used a LDA wrapper approach with sequential forward search and sequential forward floating search for biomarker discovery. Xiong et al. (2001b) used LDA with step-wise selection and Monte Carlo methods to select genes for tumour classification. Yue et al. (2007) proposed a novel wrapper-based method using LDA which recursively removes redundant genes while Huerta et al. (2010) used a genetic algorithm to generate candidate feature subsets evaluated by a LDA classifier. Peng et al. (2005) used a LDA wrapper approach with backward elimination and sequential forward selection to select relevant genes while Pique-Regi et al. (2005) proposed a novel sequential diagonal LDA (Dudoit et al., 2002a) for microarray classification.

The SVM is a powerful classification tool widely used in gene expression studies (Lee et al., 2005, Lee and Lee, 2003, Statnikov et al., 2008). The SVM maps the data into a high dimensional space (possibly infinite) and searches for the optimal hyperplane that maximizes the margin (the smallest distance between any of the data points and the decision boundary) in order to minimize the generalization error.

Consider a binary classification problem with nonlinear decision boundary. Given a set of training examples $\{\mathbf{x}_i\}_{i=1}^{n} \in \mathbb{R}^m$ with associated class labels $\{c_i\}_{i=1}^{n} \in \{-1, 1\}$ and the mapping $\boldsymbol{\phi} : \mathbb{R}^m \longrightarrow \mathcal{H}$ that maps the training examples to a higher dimensional space, the SVM algorithm constructs the maximum margin hyper-plane:

$$f(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) + b \tag{2.34}$$

by solving the quadratic optimization problem:

$$\min_{\mathbf{w},\xi,b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \tag{2.35}$$

subject to:

$$\begin{aligned} c_i(\mathbf{w}_i^\top \boldsymbol{\phi}(\mathbf{x}) + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0, \forall i = 1, \ldots, n \end{aligned} \tag{2.36}$$

where C is user supplied parameter that balances model complexity with training error minimization while $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ is a vector of slack variables (one slack variable for each $\mathbf{x}_i$) measuring the the degree of misclassification for the training examples.

The optimization problem (2.35) with constraints (2.36) admits the following dual formulation (Bishop et al., 2006):

$$\max_{\mathbf{a}} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j c_i c_j \, \kappa \left(\mathbf{x}_i, \mathbf{x}_j \right) \tag{2.37}$$

subject to:

$$\begin{aligned} \sum_{i=1}^{n} a_i c_i = 0 \\ 0 \leq a_i \leq C, \forall i = 1, \ldots, n \end{aligned} \tag{2.38}$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ is a vector of Lagrange multipliers while $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function defined as:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j) \tag{2.39}$$

By solving (2.37) with constraints (2.38), $\mathbf{w}$ and b can be recovered using:

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{n} a_i c_i \boldsymbol{\phi}(\mathbf{x}_i) \\ b &= \frac{1}{\#(\mathcal{V})} \sum_{i \in \mathcal{V}} \left(c_i - \sum_{j \in \mathcal{V}} a_j c_j \, \kappa(\mathbf{x}_i, \mathbf{x}_j) \right) \end{aligned} \tag{2.40}$$

where $\mathcal{V}$ is the set of indices of the support vectors i.e. indices of the training examples associated with non-null Lagrance multipliers. For a test instance $\mathbf{x}$, the predicted class is given by the sign of:

$$f(\mathbf{x}) = \sum_{i=1}^{n} a_i c_i \, \kappa(\mathbf{x}_i, \mathbf{x}) + b \tag{2.41}$$

Equation (2.41) shows that the decision boundary is a linear combination of dot products in $\mathcal{H}$ between the support vectors and the test instance.

Yu and Cho (2003), Huerta et al. (2006) and Zhuo et al. (2008) proposed SVM-based wrapper approaches using genetic algorithms to identify minimal subsets of features associated with high classification performance. Maldonado and Weber (2009) proposed a wrapper approach combining SVM with sequential backward elimination while Tang et al. (2006) used least squares SVM (Suykens and Vandewalle, 1999) with sequential forward selection. SVM-based wrapper approaches using backward elimination and forward selection were also used in the work of Peng et al. (2005). Li et al. (2004) used sequential floating forward search with SVM for leukemia subtype classification.

2.4.4 Embedded methods based on SVM

Besides its merits in terms of classification, the SVM has emerged as a powerful feature selection tool. Surveys on the feature selection methods using SVM-based criteria are provided by a number of authors including Lal et al. (2006) and Rakotomamonjy (2003). This section discusses embedded method based on the linear SVM:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{2.42}$$

These methods will be presented within a unified optimization framework by noting that problem (2.35) with constraints (2.36) can be expresses as a Tikhonov regularization problem (Hastie et al., 2004):

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} \left[1 - c_i \left(b + \mathbf{w}^\top \mathbf{x}_i\right)\right]_+ + \frac{\eta}{2} \|\mathbf{w}\|_2^2 \tag{2.43}$$

where η corresponds to $1/C$ in (2.35). Formulation (2.42) with optimization problem (2.43) will be referred to as L_2-norm SVM.

Guyon et al. (2002) proposed an embedded method based on the L_2-norm SVM and recursive feature elimination (RFE) , termed SVM-RFE. The method starts with the full set of features and computes at each iteration the vector of weights $\mathbf{w}$ of a L_2-norm SVM trained on the training data. The feature associated with the component with the smallest contribution to $\|\mathbf{w}\|_2^2$ is removed and the process is repeated until the subset of features with the highest prediction performance of the test data is found. Duan et al. (2005) improved the performance on SVM-RFE by stabilizing feature ranking at each step of the recursive procedure using data resampling methods.

Bradley and Mangasarian (1998) proposed an embedded method based on the L_1-norm SVM:

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} \left[1 - c_i \left(b + \mathbf{w}^\top \mathbf{x}_i\right)\right]_+ + \eta \|\mathbf{w}\|_1 \tag{2.44}$$

The L_1-norm induces sparsity on $\mathbf{w}$ for sufficiently large values of the tuning parameter η (Tibshirani, 1996), allowing features to be removed simultaneously rather than iteratively. Zhu et al. (2004) supplemented the feature selection method with adaptive selection of the tuning parameter η.

Wang et al. (2006a) noted two major limitations of the L_1-norm SVM approach to feature selection. Firstly, the L_1-norm penalty selects few representatives from groups of highly correlated features that are relevant for class separation and disregards the rest. Secondly, the upper bound for the number of features selected using the L_1-norm SVM is equal to the number of training examples. For microarray data with small sample size (where the number of differentially expressed genes can exceed the number of biological replicates available), feature selection using L_1-norm SVM can omit interesting genes. To compensate for these shortcomings, they proposed for feature selection the doubly regularized SVM (DrSVM):

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} \left[1 - c_i \left(b + \mathbf{w}^\top \mathbf{x}_i\right)\right]_+ + \frac{\eta_2}{2}\|\mathbf{w}\|_2^2 + \eta_1\|\mathbf{w}\|_1 \tag{2.45}$$

where η_2 and η_1 are tuning parameters. DrSVM uses the elastic net penalty (Zou and Hastie, 2005) on $\mathbf{w}$ consisting of the L_1-norm penalty, which imposes sparsity, and the L_2-norm penalty, which selects groups of correlated genes. DrSVM removes features simultaneously rather than iteratively, in a manner analogous to the L_1-norm SVM. When compared to feature selection methods based on L_1-norm SVM and L_2-norm SVM, the DrSVM approach achieved superior performance (Wang et al., 2006a).

2.5 Conclusions

This chapter provided an overview of the microarray technology and of the gene expression profiling using microarrays, a review of statistical methods used to select differentially expressed genes and a survey of the feature selection methods used in bioinformatics and machine learning.

The overview of the microarray technology presented the principles and the technical aspects of two widely used gene expression profiling platforms: glass slide cDNA microarrays and high-density oligonucleotide microarrays. It was argued that the former platform has high sensitivity, low specificity and low reproducibility while the latter has low sensitivity, high specificity and high reproducibility, the exception being the Affymetrix GeneChip format which achieves high sensitivity through control probes.

The overview of gene expression profiling presented the fundamental steps of a typical microarray experiment. Emphasis was placed on the data pre-processing step where widely used methods for background correction, normalization and summarization were further discussed.

The review of the statistical methods for selecting differentially expressed genes highlighted the multiplicity problems occurring when simultaneously testing multiple hypothesis and presented appropriate strategies for controlling different Type I error rates. It was argued that pFDR control provides greater power among the reviewed methods. The relation between statistical significance and biological relevance was also discussed and the need to complement statistical findings with measures of effect size was presented.

The survey of feature selection methods presented unsupervised filters for removing noisy and non-informative genes, supervised filters based on mutual information, wrapper approaches and embedded methods based on SVMs. It was argued that when the marginal independence criterion holds, unsupervised filtering can increase the power of multiple testing procedures. Additionally, it was claimed that mRMR and JMI outperform the other supervised filters while SVM-based wrapper and embedded approaches excel due to the performance of the SVM classifier.

Chapter 3

Modelling methods for microarray gene expression data

Various modelling formalisms for analysing and characterising microarray gene expression data, which can help answer important biological questions, have been developed. This chapter focuses on deconvolution of heterogeneous gene expression data and reconstruction of GRNs. Specifically, Section 3.1 presents sample heterogeneity as a major confounding factor of differential gene expression studies and provides a review of the methods used to deconvolve gene expression signals from mixture of cells into cell type-specific signatures weighted by cell type-specific proportions. Section 3.2 provides an overview of the biological properties of the GRNs and a review of the main formalisms used to model the dynamic gene regulatory interactions. Concluding remarks are given in Section 3.3.

3.1 Deconvolution of microarray gene expression data from heterogeneous tissues

Microarray gene expression data is usually measured from tissue samples containing multiple cell types (heterogeneous sample) in varying degrees of abundance (Liebner et al., 2013) and operating according to different programs of transcription (Su et al., 2002). Thus, gene expression patterns represent a convolution of the cell type-specific signals weighted by the associated cell type frequencies. This convolution masks the underlying sources of measurement variability and hinders the biological interpretation of the results from downstream analyses (Erkkilä et al., 2010). In particular, it limits the conclusions derived from differential gene expression studies as the observed differences in gene expression cannot be attributed to differences in cell type composition or differences in cellular contributions to the total mRNA pooled from the mixture (Shen-Orr et al., 2010, Wang et al., 2006b).

Cell separation techniques such as laser capture micro-dissection (LCM) can be used to address the problems underlying sample heterogeneity by extracting purified cell type signatures. However this approach is costly, time consuming, labour intensive and can introduce artefacts in gene expression measurements (Lähdesmäki et al., 2005, Zuckerman et al., 2013). A powerful alternative to physical separation methods has emerged in the form of computational tools that perform gene expression deconvolution (Gaujoux and Seoighe, 2013). These methods are based on a linear deconvolution model and depend on the type of data available.

Section 3.1.1 discusses the deconvolution model and presents the problems that can be addressed within its framework. Solutions to these problems are reviewed in Section 3.1.2, Section 3.1.3 and Section 3.1.4.

3.1.1 The linear deconvolution model

Considering that gene expression is often measured from complex mixtures of cell types present in various proportions and exhibiting different transcriptional programs, a sensible approach to expression deconvolution is to model the transcript abundance of a gene as a linear combination of cell type-specific expression levels weighted by the cell type-specific proportions (Shen-Orr et al., 2010). Formally, given the gene expression matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where n denotes the number of samples (biological or technical replicates) and m denotes the number of genes, the deconvolution model is defined as:

$$x_{ij} = \sum_{k=1}^{K} f_{ik} g_{kj} \tag{3.1}$$

with

$$\sum_{k=1}^{K} f_{ik} = 1, \forall i \tag{3.2}$$

where $x_{ij} \geq 0$ represents the expression level of gene j in sample i, $f_{ik} \geq 0$ represents the proportion of cell type k in sample i, $g_{kj} \geq 0$ represents the expression level of gene j in cell type k, and K is the total number of cell types in the mixture. The measurements x_{ij} can represent raw or log-transformed data. The benefits of performing expression deconvolution using raw and log-transformed data will be discussed later in this chapter.

Being arguably the most widespread formalism for gene expression deconvolution (Zhao and Simon, 2010), this model faithfully describes how cell types in a mixture collectively contribute to the overall transcript abundance (Shen-Orr et al., 2010). Depending on the type of data available, the linear model (3.1)-(3.2) can be used to address different deconvolution problems:

1. Estimation of the cell type-specific proportions f_{ik} given heterogeneous gene expression data x_{ij} and cell type-specific expression levels g_{kj}.
2. Estimation of the cell type-specific expression levels g_{kj} given heterogeneous gene expression data x_{ij} and cell type-specific proportions f_{ik}
3. Joint estimation of cell type-specific proportions f_{ik} and cell type-specific expression levels g_{kj} given heterogeneous gene expression data x_{ij}.

Methods for solving these problems are discussed in the following sections.

3.1.2 Estimation of cell type-specific proportions

In order to estimate the cell type proportions from heterogeneous gene expression data, the purified expression levels of the cell types in the mixture (g_{kj}) are required. These quantities can be obtained prior to gene expression profiling using LCM or can be substituted with reference expression signatures (Abbas et al., 2009) available at Gene Expression Omnibus (GEO) (Barrett et al., 2009). Given these measurements, estimation of the cell type-specific proportions associated with the ith sample reduces to solving the linear systems of equations:

$$x_{ij} = \sum_{k=1}^{K} f_{ik} g_{kj}, \; j = 1 \ldots m \tag{3.3}$$

for the unknowns $f_{ik} \geq 0$ satisfying (3.2).

Many genes exhibit the same transcriptional program across cell types (Mellick et al., 2002) and therefore have little utility in estimating the mixing proportions (Gong et al., 2011). Only genes relatively specific for the cell types in the mixture (cell type markers) can be used as reliable basis for deconvolution (Gong et al., 2011). Methods for estimating the cell type-specific proportions consist of two steps: (i) selecting an appropriate subset of basis genes for deconvolution, (ii) solving the linear deconvolution problem using an optimization algorithm.

To select the deconvolution basis, Wang et al. (2010) proposed an approach that consists of ranking genes by their F-statistic. Subsets of increasing numbers of highest ranked genes were evaluated using cross-validation on test data to identify the subset with the lowest average prediction error rate. Lu et al. (2003) manually selected the genes associated with the transcriptional program of interest. Abbas et al. (2009) ranked genes based on their degree of differential expression between cell types using Student's t-test. Matrices containing the reference expression signatures for subsets of increasing numbers of top genes were evaluated by their condition number. The genes contained in the matrix with the smallest condition number (lowest sensitivity to small perturbations in the data) were selected as basis for deconvolution. Gong et al. (2011) adopted the same approach based on minimizing the condition number of matrices containing reference expression signatures but selected cell type-specific markers using linear modelling and empirical Bayes methods (Smyth, 2004). Finally, Wang et al. (2006b) used stepwise discriminant analysis to identify the optimal subset of genes for deconvolution.

To solve the linear deconvolution problem, Lu et al. (2003) and Wang et al. (2006b) used an approach based on simulated annealing (Kirkpatrick, 1984) while Abbas et al. (2009) applied a linear least squares algorithm iteratively until positive mixing proportions were estimated. Gong et al. (2011) noted that the former method can be trapped in local minima while the later requires small negative parameters to be removed at each step. To avoid these problems, they proposed a quadratic programming technique that identifies the globally optimal set of positive mixing parameters (in the least squares sense).

Deconvolution of the cell type-specific proportions can provide a valuable insight into the biology of the system being studied by revealing the dynamics of cell populations and identifying cell growth defects (Lu et al., 2003). Additionally, it can enhance the understanding of human diseases (Abbas et al., 2009, Wang et al., 2010) and increase the statistical power to detect differentially expressed genes (Wang et al., 2006b).

3.1.3 Estimation of cell type-specific expression levels

In order to estimate the cell type-specific expression levels from heterogeneous gene expression data, the cell type-specific fractions for each sample are required. These quantities can be obtained using automated methods such as flow cytometry (Shapiro, 2005). Given these measurements, estimation of the cell type-specific signatures associated with the jth gene reduces to solving the system of linear equations:

$$x_{ij} = \sum_{k=1}^{K} f_{ik} g_{kj}, \; i = 1 \ldots n \tag{3.4}$$

for the unknowns $g_{kj} \geq 0$.

Stuart et al. (2004) solved the system of equations using standard linear least-squares regression to identify cell type-specific patterns in prostate cancer. The same approach was used by Shen-Orr et al. (2010) to identify cell type-specific genes that are differentially expressed between groups of kidney transplant recipients. Their method, termed cell type-specific significance analysis of microarrays (csSAM), sets to zero negative estimates of g_{kj}.

Zhong and Liu (2011) noted that csSAM underestimates true expression signals when log-transformed data is used and showed that better deconvolution performance can be obtained in the linear space (raw data). On the other hand, deconvolution of log-transformed data increases the performance of statistical methods to detect cell type-specific differences between groups compared to the case when raw data is used (Shen-Orr et al., 2011).

Liebner et al. (2013) addressed the problem of over-fitting the model parameters when individual cell types are present at low frequencies in the mixture or the number of cell types exceeds the number of samples. Their method partitions the cell types into subsets and assumes that within a subset the cell types share common expression signatures. These subsets are evaluated using the Akaike information criterion (AIC) (Akaike, 1974) to identify the subset best supported by the data. Parameter optimization is performed using standard least squares.

While the methods discussed above provide estimates of the cell type-specific levels, they don't quantify the contribution of each cell type to the variance of the heterogeneous gene expression measurements. Ranking cell types in terms of their increment to the observed variance can help identify features of interindividual variation in gene expression patterns. This problem is addressed in Chapter 5 where a new deconvolution method that additionally identifies the number of cell type sources of gene expression is proposed.

Deconvolution of the cell type-specific expression levels defines the framework where problems related to cell type-specific markers discovery and identification of cell type-specific transcriptional programs across different experimental conditions can be adequately tackled using efficient computational tools.

3.1.4 Joint estimation of cell type-specific proportions and expression levels

Most gene expression profiling studies rarely measure the proportions of cells in the mixture or isolate cell populations to extract purified expression signatures (Zuckerman et al., 2013). The previous methods relied on purified cell type signatures or mixing percentages to solve different deconvolution problems. In the absence of such data, these problems can be addressed within the unified framework of blind deconvolution by solving the system of equations:

$$\mathbf{X} = \mathbf{F} * \mathbf{G} \tag{3.5}$$

for the unknowns $\mathbf{F} = [f_{ik}]$ and $\mathbf{G} = [g_{kj}]$ with $f_{ik} \geq 0$ and $g_{kj} \geq 0$ subject to the constraints (3.2).

Venet et al. (2001) estimated the model parameters in (3.5) using a two-step approach. Starting from an initial guess of **G**, the positive entries in **F** are estimated using a non-negative least squares algorithm. In the second step, the estimated **F** is used to optimize **G** in the non-negative least squares sense using the additional constraints:

$$\sum_{j=1}^{m} g_{kj} = m, \forall k \tag{3.6}$$

These constraints ensure that **X** and **G** have the same normalization. The two steps are repeated until convergence to a local or global optimum. To find a unique solution, the authors proposed a transformation that makes the rows of **G** uncorrelated at the end of each second step. Using this transformation, the convergence of the algorithm to a local or global optimum is no longer guaranteed (Venet et al., 2001).

Lähdesmäki et al. (2005) proposed a two-step approach similar to Venet et al. (2001). In the first step, given an initial guess for **F**, the matrix **G** is estimated using standard least squares. This estimate is used in the second step to optimize **F** using the same optimization algorithm. The two steps are repeated until convergence to a local or global optimum. For the case when the number of cell types in the mixture is unknown, the authors proposed an extension based on cross-validation to perform model selection.

Erkkilä et al. (2010) proposed a probabilistic approach to simultaneously estimate cell type proportions and cell type-specific signatures. The method assigns prior distributions to the model parameters in (3.5) and constructs their posterior density given heterogeneous gene expression data. The optimal parameters are estimated in a Markov chain Monte Carlo (MCMC) fashion from the posterior density. For the case when the number of cell types in the mixture is unknown, the authors suggest using cross-validation or reversible-jump MCMC (Green, 1995) in order to estimate a suitable value for K.

Zuckerman et al. (2013) proposed a three-step approach for blind deconvolution that additionally identifies the number of cell types in the mixture and their identities. In the first step, an initial estimate of the matrix **G** is obtained using non-negative matrix factorization (Piper et al., 2004). This step requires an initial guess of the number of cell types in the analysed tissue together with their reference signatures. These signatures, which need not be study-specific or reflective of pathological state of the tissue (healthy, disease), can be taken from public databases (Gardiner-Garden and Littlejohn, 2001). In the second step, the symmetric Kullback-Leibler divergence (Cover and Thomas, 2012) between the reference signatures and the estimated signatures is used to identify the true number of cell types in the mixture. This step removes redundant rows from **G** allowing the final estimate to be constituted only from the relevant rows. In the third stage, the matrix **F** is estimated in the non-negative least squares sense.

Joint deconvolution of cell type-specific proportions and expression levels borrows from the advantages of the constituent deconvolution frameworks. Specifically, it can reveal the dynamics of cell populations, identify cell growth defects, uncover cell type-specific markers and expose differences in cell type-specific gene expression between different classes or experimental conditions.

3.2 Modelling dynamic GRNs

Understanding the principles underlying the cellular functions, requires knowledge of the underpinning GRNs (see Appendiz A for a definition). Mathematical models of the gene regulatory pathways represents powerful tools for system-level analysis, numerical simulation and intervention studies. Such studies could predict new structures or functionalities and ultimately uncover effective therapeutic strategies for correcting human diseases.

The advent of high-throughput genomic technologies inspired considerable scientific research toward efficient GRN reconstruction. In particular, the time-course interrogation of gene expression abundance using microarrays provided a genome-wide survey of the cell's temporal activity. The microarray technology signalled the departure from the limited single-gene approaches attempting to explain the cellular processes and advanced the global perspective of genetic regulation by harnessing the collective power of genes (to elucidate the systemic properties of organisms).

The ability to harvest large amounts of time series gene expression data empowered the development of computational and formal tools to capture the topological organization of the GRN, to describe the dynamic nature of the regulatory pathways and to make qualitative and quantitative predictions about the network's behaviour under different conditions. These tools are rooted in a wide spectrum of modelling formalisms that use experimental data to solve the inverse problem of GRN reconstruction (reverse-engineering).

The following sections provide an overview of the general properties of GRNs used to assist the modelling process and a review of three major classes of modelling formalisms: Boolean and probabilistic Boolean networks, dynamic Bayesian networks and state-space models . The review is not exhaustive with respect to the different formalism for modelling GRNs. Specifically, approaches based on Petri nets (Steggles et al., 2007), non-linear differential equations (Qian et al., 2008) and stochastic equations (McAdams and Arkin, 1999) are omitted. Comprehensive reviews are provided by a number of authors including De Jong (2002), Schlitt and Brazma (2007), Lee and Tzou (2009) and Hecker et al. (2009).

3.2.1 General properties of GRNs

Efficient reconstruction of large-scale GRNs requires a considerable amount of accurate time-course gene expression data (low measurement noise). This amount increases approximately logarithmically with the number of genes in the network (Hecker et al., 2009). To address this dimensionality problem, some authors increased the amount of data using spline interpolation (Bansal et al., 2006, D'haeseleer et al., 1999). An alternative approach is to incorporate biological constraints into the modelling process to reduce the space of candidate topologies and range of network parameters. These constraints are summarized below.

Sparsity

Sparsity is the most frequently used biological constraint in GRN reconstruction and relates to genes having a small number of regulators (Arnone and Davidson, 1997). From a topological point of view, genes have small in-degree (number of inward connections) but large out-degree (number of outward connections). A pictorial representation of these notions is shown in Figure A.6 of Appendix A.

The sparsity constraint is particularly appealing for modelling as it reduces the number of parameters to be estimated. The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) has been largely used to enforce topological sparsity (Fujita et al., 2007, Gustafsson et al., 2005, van Someren et al., 2006). Other methods rely on removing regulatory connections whose associated parameters are either below a certain threshold (Chen et al., 2011, d'Alché Buc et al., 2005) or have negligible sensitivity (De Hoon et al., 2002).

Scale-free property

Scale-free models (Barabási and Albert, 1999) represent a blueprint for the topological organization of large-scale biological networks (Albert, 2005, Jeong et al., 2000, Jordan et al., 2004). According to these models, the probability distribution of the node-specific connections k follows the power law $P(k) = k^{-\nu}$, where ν is a network-specific constant. This distribution implies that the topology of scale-free networks contains few central nodes (hubs) which are highly connected with the remaining, peripheral nodes. Hubs emerging in biological networks are believed to provide the link between groups of nodes (modules) responsible for different metabolic functions (Jeong et al., 2000). In GRNs, hubs are represented by genes that dominate the overall regulation.

To capture the scale-free topology of the gene regulatory pathways, Chen et al. (2008) proposed a method that constructs the network connectivity using the correlation between nodes. This approach requires the central and peripheral nodes to be specified a priori. Wildenhain and Crampin (2006) proposed two sequential methods that consists of removing and adding connections until an optimal topology is found (according to some scoring function).

Robustness

GRNs are robust to fluctuations in molecular concentrations, noisy expression of the constituent genes and even knock-out of individual genes (Ciliberti et al., 2007a, MacNeil and Walhout, 2011). Highly robust networks are relatively insensitive to small structural perturbations (robustness to mutations) and to variation of their internal parameters and changes in the environment (robustness to noise) (Ciliberti et al., 2007b, Kitano, 2002). These forms of insensitivity are essential to ensure consistency of cell reproduction and persistence of functioning under the inherent stochasticity of biochemical reactions (Davidson, 2001). It has also been argued that only certain topological arrangements result in robust GRNs (Ingolia, 2004, Jeong et al., 2000).

Methods for modelling robust GRNs focused on capturing the robustness to noise as opposed to robustness to mutations. These methods rely on conditioning the model to be asymptotically stable under external perturbations and variations of initial conditions of the system (Chen et al., 2011, 1999, Koh et al., 2009).

Modularity

Cellular networks consist of functionally related components clustered in densely connected modules, which in turn are sparsely connected with each other (Hartwell et al., 1999). Departures from this topological arrangement consisting of overlapping modules were observed in protein-protein networks (Han et al., 2004). In terms of GRNs, topological modularity reflects the scale-free organization of the network. Strongly co-expressed genes or genes that share functional roles are clustered in modules connected through hubs (Jeong et al., 2000).

To capture the modular structure of the GRNs Segal et al. (2003) proposed an iterative method that reassigns gene to modules until a certain criterion is optimized while Hirose et al. (2008) used a state-space formulation of the network that automatically identifies genes with similar expression patterns.

3.2.2 Boolean and probabilistic Boolean networks

Boolean networks (Kauffman, 1969) are rule-based dynamical systems that model the qualitative (logical) interactions between genes. These models are grounded in the biological knowledge that cells exhibit switch-like transitions between functional states associated with growth or patterns of response to external stimuli. To capture this switching behaviour, Boolean networks assume each gene can take two possible values, ON (expressed) and OFF (unexpressed), and that the functional interactions between genes are represented by logical rules. Binarized gene expression data retain substantial biological information (Shmulevich and Zhang, 2002, Shmulevich et al., 2002b) which allows Boolean networks to capture the generic properties of gene regulatory networks (Lähdesmäki et al., 2003).

Formally, the Boolean network can be represented as a directed graph $G(\mathcal{V}, \mathcal{F})$ defined by a set of nodes (genes) $\mathcal{V} = \{x_1 \ldots, , x_m\}$ with $x_i \in \{0, 1\}$ and a set of logic functions $\mathcal{F} = \{f_1, \ldots, f_m\}$. The ith gene is expressed if $x_i = 1$ and unexpressed if $x_i = 0$. Each x_i is associated with a Boolean function $f_i(x_{i_1}, \ldots, x_{i_{k(i)}})$ with $\left\{i_1, \ldots, i_{k(i)}\right\} \subseteq \{1, \ldots, m\}$, where $k(i)$ represents the in-degree (number of parent nodes) for the node x_i. At a given time t, the state of the Boolean network $G(\mathcal{V}, \mathcal{F})$ is given by the binary vector $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))$. The transition from $\mathbf{x}(t)$ to $\mathbf{x}(t+1)$ is determined by the synchronous update of the nodes according to their Boolean functions:

$$x_i(t+1) = f_i(x_{i_1}(t), \ldots, x_{i_{k(i)}}(t)) \tag{3.7}$$

Although the gene regulation operates according to asynchronous dynamics, the artificial synchrony of the Bayesian networks preserves the qualitative properties of the regulatory pathways while simplifying the computation associated with network estimation and simulation (Shmulevich et al., 2002a).

Methods for estimating the underlying network topology consists of searching for each node the space of Boolean functions with varying number of inputs that are consistent with the state transitions. Liang et al. (1998) reduced the search space using mutual information to identify the regulators for each node whereas Akutsu et al. (1999) opted for an exhaustive search. The later approach is more general as it solves the problems of identifying all the networks that are consistent with the data. Having the complete set of solutions allows for model validation and selection based on unseen (testing) data. The previous methods work for any $k(i) \leq m$; in practice $k(i)$ is set to small values for biological considerations (network sparsity) and computational considerations (avoid over-fitting, speed-up estimation).

The deterministic directionality of Boolean networks coupled with the finite cardinality of the state space guarantees that the system cycles infinitely often through a finite set of states. These stable states, called attractors, capture the long-term behaviour of the network (Somogyi and Sniegoski, 1996). The set of states that leads to an attractor constitutes its basin of attraction (Wuensche, 1998). The size of the basin of attraction dictates the likelihood that the system will cycle through or settle on the corresponding attractor given random initial (starting) conditions. Attractors of Boolean networks were associated with cellular phenotypes (Huang, 1999). These findings suggest than questions related to cell proliferation can be tackled using the Boolean formalism.

One major criticism addressed to Boolean networks is that they represent closed systems grounded in biological determinism. These modelling assumptions preclude the use of the inherent stochasticity associated with biological processes and the uncertainty in the gene expression measurements (aggregated at each step of the gene expression profiling). To address these issues, Shmulevich et al. (2002a) proposed the probabilistic Boolean networks.

Formally, a probabilistic Boolean network can be represented as a directed graph $G(\mathcal{V}, \mathcal{F})$ defined by the set of nodes (genes) $\mathcal{V} = \{x_1, \ldots, x_m\}$ with $x_i \in \{0, 1\}$ and a set of logic functions $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_m\}$ where $\mathcal{F}_i = \left\{f_1^{(i)}, \ldots, f_{l(i)}^{(i)}\right\}$ and $l(i)$ is the number of functions for x_i. The sets of functions $\mathcal{F}_i$ are determined using the coefficient of determination (COD) (Dougherty et al., 2000). Each function $f_j^{(i)}, j = 1, \ldots, l(i)$, has a probability $p_j^{(i)}$ of being chosen to predict x_i given by:

$$p_j^{(i)} = \frac{d_j^{(i)}}{\sum_{k=1}^{l(i)} d_k^{(i)}} \tag{3.8}$$

where $d_j^{(i)}$ is the COD for x_i with respect to the set of genes used by $f_j^{(i)}$ which need not be the same across j. At any point in time, the probabilistic Boolean network is defined by the a realization $\mathbf{f}_k = \left\{f_{k_1}^1, \ldots, f_{k_m}^m\right\}$ with $f_{k_i}^{(i)} \in \mathcal{F}_i$, $1 \leq k_i \leq l(i)$, and $k = 1, \ldots, N$ where N denotes the number of possible realizations. Thus, state transitions are operated at each instant of time by one of the N possible Boolean networks.

The attractors of a probabilistic Boolean network consists of the union of the sets of attractors associated with each Boolean network that can operate at a given time (Xiao, 2009). These attractors have steady-state distributions characterising the long-term behaviour of the system independent of the initial state. Since attractors can be associated with cellular phenotypes, the steady-state distributions reflect the likelihood of observing these phenotypes. Methods for steering the long-time behaviour of the network towards desirable states without changing the steady-state distribution of the attractors were proposed by Shmulevich et al. (2002d). In some cases, it would be advantageous to alter the steady state probabilities in order to decreases the changes of the network ending up in an undesirable attractor. This issue was addressed in the work of Shmulevich et al. (2002c) where a methodology based on minimal structural modifications of the network dependencies was proposed.

Probabilistic Boolean networks supplement the strengths of Boolean networks with the ability to incorporate biological knowledge into the network topology by constraining the space of candidate Boolean functions. Additionally, absorbing the sources of uncertainties into the set of functions $\mathcal{F}_i$ allows for a more general modelling of the qualitative rules of genetic regulation. Probabilistic Boolean networks are intimately related to dynamic Bayesian networks (Lähdesmäki et al., 2006, Xiao, 2009). This class of methods for inferring GRNs is discussed in more detail in the next section.

3.2.3 Dynamic Bayesian networks

Dynamic Bayesian networks (Murphy et al., 1999) are quantitative causal models that capture the non-linear stochastic regulatory interactions between genes. They represent a temporal extension to Bayesian networks (Friedman et al., 2000) to account for cyclic regulation (feedback loops).

Formally, a Bayesian network is defined by a pair $B = (G, \Omega)$ where G is a directed acyclic graph defined by the set of nodes $\mathcal{V} = \{x_1, \ldots, x_m\}$ and Ω denotes a set of network parameters. The random variable x_i associated with the ith gene is drawn from the conditional probability distribution $P(x_i|Pa(x_i))$ parameterized by Ω, where $Pa(x_i)$ denotes the set of parents of x_i. The graph G encodes the Markovian assumption that each x_i is independent of its non-descendants. This allows for the joint probability distribution induced by B over all x_i to be written as:

$$P(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} P(x_i|Pa(x_i)) \tag{3.9}$$

This representation, which captures the static interactions between genes, was extended by dynamic Bayesian networks to model temporal processes. Formally, a dynamic Bayesian network is defined by a pair $DB = (B_0, B_{\rightarrow})$ where the prior Bayesian network $B_0 = (G_0, \Omega_0)$ specifies the distribution of the initial state

$\mathbf{x}(0) = (x_1(0), \ldots, x_m(0))$ and the transition Bayesian network $B_{\rightarrow}(G, \Omega)$ specifies the transition probabilities $P(\mathbf{x}(t)|\mathbf{x}(t-1))$. The joint distribution over $N+1$ sequences of observations $\mathbf{x}(0), \ldots, \mathbf{x}(N)$ can be expressed as:

$$P(\mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(N)) = P(\mathbf{x}(0)) \prod_{i=1}^{N} P(\mathbf{x}(t)|\mathbf{x}(t-1)) \tag{3.10}$$

Using the factorization in (3.9), the joint distribution (3.10) can be written as:

$$P(\mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(N)) = \prod_{i=1}^{m} P(x_i(0)|Pa(x_i(0))) \prod_{t=1}^{N} \prod_{j=1}^{m} P(x_j|Pa(x_j)) \tag{3.11}$$

Methods from estimating the topology and the parameters of the dynamic Bayesian network model (3.11) from discrete and continuous data were proposed by Friedman et al. (1998) and Kim et al. (2004). These methods consist of a search strategy and a scoring function. The search strategy explores the space of models using hill-climbing (Johnson et al., 1988) or simulated annealing (Kirkpatrick, 1984) to provide candidate topologies for evaluation. The scoring function combines structure selection with parameter optimization to identify the network that is most consistent with the data. Examples of scoring functions used are: the Bayesian information criterion (BIC) (BIC) (Schwarz et al., 1978), Bayesian Dirichlet equivalent (BDe) (Heckerman et al., 1995) and the Bayesian network and nonparametric regression criterion (BNRC) (Kim et al., 2004). Parameter optimization can be performed using maximum likelihood estimation (Marx and Larsen, 2006) when complete data is available or structural expectation maximization (EM) (Friedman et al., 1997b) when incomplete data is available (missing values or hidden nodes).

Although dynamic Bayesian networks model the temporal interactions between genes, their structure is time-invariant. Song et al. (2009) proposed an extension exploiting time-varying topologies to capture the systematic rewiring of gene networks. Overall, dynamic Bayesian networks represent a powerful modelling tool that accounts for the inherent stochasticity of the gene expression data and accommodates missing variables. In the next section, linear state-space models, which represent a subclass of Dynamic Bayesian networks, are discussed.

3.2.4 Linear state-space models

Linear state-space models describe a dynamical system by a set of first order differential (continuous-time representation) or difference (discrete-time representation) equations using unobserved internal variables also known as state variables, representing independent energy storage elements of the system. In continuous-time, the state-space description of a dynamical system with input $\mathbf{u}(t) \in \mathbb{R}^m$ and output $\mathbf{y}(t) \in \mathbb{R}^p$ takes the form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{3.12}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{3.13}$$

where $t \in \mathbb{R}$, $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^\top$ is the state vector, n is the model order and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ are matrices with adequate dimensions. Equation (3.12) describes the rate of change of the system's energy given past realizations of the input and state vector. Equation (3.13) gives the output signals for the current realizations of the input and state vector.

The continuous-time representation can be discretized to:

$$\mathbf{x}(t+1) = \tilde{\mathbf{A}}\mathbf{x}(t) + \tilde{\mathbf{B}}\mathbf{u}(t) \tag{3.14}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{3.15}$$

where $t \in \mathbb{Z}$ and the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ are given by:

$$\tilde{\mathbf{A}} = e^{\mathbf{A}\Delta t} \tag{3.16}$$

$$\tilde{\mathbf{B}} = \int_0^{\Delta t} e^{\mathbf{A}\tau}\mathbf{B}\mathrm{d}\tau \tag{3.17}$$

with Δt denoting the sampling time. An important property for monitoring the system's response to external perturbations and changes in the initial condition is stability. Letting $\sigma(\mathbf{A}) = \{z \in \mathbb{C} \ |\det(z\mathbf{I} - \mathbf{A}) = 0\}$ denote the spectrum of $\mathbf{A}$, the continuous-time system (3.12)-(3.13) is asymptotically stable (Hurwitz stable) if $\sigma(\mathbf{A}) \subseteq \mathbb{C}_H$, where $\mathbb{C}_H = \{z \in \mathbb{C} | \mathrm{Re}(z) < 0\}$. The discrete-time system (3.14)-(3.15) is asymptotically stable (Schur stable) if $\sigma(\tilde{\mathbf{A}}) \subseteq \mathbb{C}_S$, where $\mathbb{C}_S = \{z \in \mathbb{C} || z| < 1\}$. These stability conditions guarantee that given bounded input signals, the output signals will also be bounded.

Methods for estimating stable state-space models from input-output observations consist of two steps: structure selection and parameter optimization. Structure selection is concerned with identifying the dimensionality of the state vector and can be approached using cross-validation (Rangel et al., 2004), BIC or AIC. Parameter optimization is concerned with identifying the system's matrices together with the initial state and can be approached using non-iterative procedures based on subspace methods (Van Overschee and De Moor, 1994) or iterative procedures based on prediction error minimization (Ljung, 1998), EM (Dempster et al., 1977) or variational Bayesian EM (Beal et al., 2005).

The state-space formalism has recently penetrated the field of GRN reconstruction. Rangel et al. (2004) used state-space models to reverse engineer the transcriptional network of genes involved in T-cell activation,Yamaguchi and Higuchi (2006) modelled the regulatory pathways between genes involved in the cell cycle regulation of yeast, Yamaguchi et al. (2007) estimated regulatory circuits between gene-modules while Koh et al. (2009) used state-space models to infer gene regulatory relationships with time delays.

These studies used time series gene expression data as input and output signals for the state-space model, leaving the state variables to model unobserved dynamics related to regulatory proteins or mRNA effects. Another widely used approach is to let the state variables model the observed gene expression dynamics (changes in gene expression levels over time (hours)) using the simplified state-space description:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \tag{3.18}$$

where $t \in \mathbb{R}$ and $\mathbf{A} = [a_{ij}]$ denotes the connectivity matrix. The weight a_{ij} measures the strength of the regulatory effect that $x_j(t)$ has on $x_i(t)$. A negative value for a_{ij} indicates that gene j inhibits gene i while a positive value for a_{ij} indicates that gene j activates gene i; a value of zeros indicates that gene j has no regulatory effect on gene i. The system of equations (3.18) can be approximated by the system of difference equations:

$$\mathbf{x}(t+1) = \tilde{\mathbf{A}}\mathbf{x}(t) \tag{3.19}$$

where $t \in \mathbb{Z}$ and $\tilde{\mathbf{A}}$ is given by (3.16)

Given the time series gene expression data $\{\mathbf{x}(t_k)\}_{k=1}^{N}$, the problem of modelling the gene regulatory interactions using (3.18) or (3.19) reduces to estimating the parameters of the matrices $\tilde{\mathbf{A}}$ or $\mathbf{A}$. For the discrete-time representation, when $N > n$, parameter estimation can be carried out using standard methods of linear algebra such as ordinary least squares. In practice, the number of temporal observations is smaller than the number of measured genes. This makes the inference problem ill-conditioned and the parameter estimation procedure prone to over-fitting (Aluru, 2005).

To avoid the curse of dimensionality, Yeung et al. (2002) proposed a method that searches among all the networks consistent with the data for the solution with the smallest number of connections (sparsity); Wang et al. (2008) proposed an EM-based algorithm to handle data shortage while D'haeseleer et al. (1999) and Bansal et al. (2006) sampled cubic spline interpolants of the time series data to generate enough samples to avoid ill-conditioning.

Another problem with estimating $\tilde{\mathbf{A}}$ comes from the irregular sampling of the data. Biological experiments usually produce unequally spaced time-series measurements. Sampling asynchrony is often disregarded in studies modelling the regulatory interactions using the discrete-time representation, the exception being (Holter et al., 2001). Additionally, discrete-time models exhibit sensitivity issues with respect to parameters in comparison to continuous-time models (Unbehauen and Rao, 1998).

Gene regulatory pathways operate according to continuous dynamics. The problem of estimating $\mathbf{A}$ was partially addressed in Bansal et al. (2006) were the discrete model was converted to continuous form using a bilinear transformation. Chen et al. (1999) directly estimated $\mathbf{A}$ using a Fourier series decomposition of the time-series gene expression data. This method is restricted to genes driven by cyclic (periodic) dynamics. Wang et al. (2006c) proposed approximating the derivatives in (3.18) using first-order finite differences prior to parameter estimation while Gustafsson et al. (2005) estimated derivatives from spline interpolation of the data. However, taking derivatives poses the risk of increasing the noise level in the data (Bansal et al., 2006, Unbehauen and Rao, 1998).

Stability of gene regulatory pathways has been largely disregarded as a biological constraint when modelling the connectivity between genes with linear differential equations. In the absence of this fundamental feature, monitoring the network's response to specific interventions becomes impractical. Chen et al. (1999) suggested setting the eigenvalues of the continuous-time dynamical system to integer multiples of the cell cycle frequency followed by the estimation of the linear parameters accounting for the connectivity between genes. Zavlanos et al. (2011) used Gersgorin's circle theorem and Lyapunov's theorem to estimate stable regulatory networks from steady-state gene expression data. Improvements and extensions to this work are proposed in (Kulkarni et al., 2012) where stability constraints are formulated using Perron-Frobenius diagonal dominance conditions. Finally, Chen et al. (2011) used Gergorin's theorem to reverse-engineer discrete-time gene regulatory networks from multiple time-course datasets.

The problem of estimating large scale stable GRNs with continuous-time representation (3.18) from time-course measurements without using derivatives was not addressed in the literature. This problem is solved in Chapter 6 where a novel method for inferring stable GRNs is presented.

3.3 Conclusions

Microarray gene expression data can provide a valuable insight into the biology of the system being studied. To harness this power, a wide spectrum of modelling formalism and computational tools were developed. This chapter provided a review of the methods for deconvolving heterogeneous gene expression data and for inferring GRNs models from time-course microarray measurements.

The survey of deconvolution methods discussed the limitations imposed by sample heterogeneity with respect to the accurate interpretation of the results arising for differential expression studies and presented different formalism to extract the missing biological information which can take the form of cell type-specific proportions and/or cell type-specific expression levels. Within the framework of estimating the cell type-specific proportions it was argued that only cell type markers should be used in the deconvolution basis. Methods for selecting an appropriate subset of basis genes were also discussed. When estimating the cell type-specific expression levels it was claimed that microarray deconvolution leads to more accurate results when it is performed in the linear space as opposed to the logarithmic space, the latter being more appropriate for cell type-specific differential expression analyses. The lack of deconvolution methods that estimate the cell type-specific contributions to the variance of gene expression patterns was also discussed. The typical deconvolution scenario when only heterogeneous microarray data is available was tackled within the framework of joint estimation of cell type-specific proportions and cell type-specific expression levels. Deconvolution methods within this class are more general as they address the challenges of the constituent deconvolution frameworks.

The survey of GRN reconstruction methods presented data scarcity as the major obstacle for reliable inference of the gene regulatory pathways and argued that biological constraints should be used to narrow the space of topologies that fit the data equally well. Additionally, three major modelling formalism were presented and their strengths were further discussed. Specifically, the Boolean and probabilistic Boolean networks were shown to be appropriate for biological systems exhibiting switching behaviour between functional states; dynamic Bayesian networks stand out by accounting for the inherent stochasticity of microarray data while state-space models can capture the hidden factors driving the gene regulatory interactions. Within the framework of state-space models, in was pointed out the lack of methods to model stable GRNs from time-course gene expression data using systems of differential equations.

Chapter 4

A novel multi-stage feature selection method for microarray gene expression data

Multi-stage feature selection methods have emerged as a powerful alternative to standalone methods (Bins and Draper, 2001, Tang et al., 2005). Peng et al. (2005) proposed a two-stage method consisting of mRMR filter followed by a wrapper approach operating in either forward or backward selection fashion.

Tang et al. (2007) stabilized the SVM-RFE algorithm using a pre-filtering step that removes noisy and redundant genes. The resulting two-stage SVM-RFE method is computationally more efficient, more accurate and more reliable than the standard SVM-RFE.

Peng et al. (2006) used a two-stage feature selection method for biomarker discovery. The subset of relevant genes selected in the first stage using the Fisher's discriminant ratio filter (Wang et al., 2011) is further narrowed in the second stage by a wrapper approach operating in sequential forward selection fashion.

Ahsen et al. (2012) proposed a two-stage approach that combines the L_1-norm SVM with the standard t-test and RFE. The method, called l_1-STaR, uses randomized splits of the data into training and test partitions to stabilize feature ranking at each step of the recursive procedure.

Bins and Draper (2001) proposed a three-stage feature selection method. The first stage uses a modified Relief algorithm (Kononenko, 1994) to filter out irrelevant features. In the second stage, redundant genes are removed using a K-means algorithm (MacQueen, 1967) while in the third stage relevant genes are selected using the Mahalanobis distance (Mahalanobis, 1936) inside various sequential search strategies.

Du et al. (2013) developed a three-stage feature selection method for microarray gene expression data. The first stage uses an improved version of the signal-to-noise ratio proposed by Golub et al. (1999) to remove irrelevant genes. In the second stage, gene clusters obtained using a support vector clustering algorithm (Sun et al., 2008) are ranked by a recursive cluster elimination method. Low-ranking clusters are removed and a SVM-RFE algorithm is used to rank the genes in each on the remaining clusters. The second stage ends with redundant (low-ranking) genes being removed. The third stage uses the standard SVM-RFE algorithm to select a final subset of genes.

This chapter proposes a new multi-stage feature selection method that combines some of the approaches reviewed in Chapter 2 according to a hierarchical structuring of different notions of relevance. Specifically, the method consists of four feature selection stages imposing stage-specific levels of stringency, as shown in Figure 4.1.

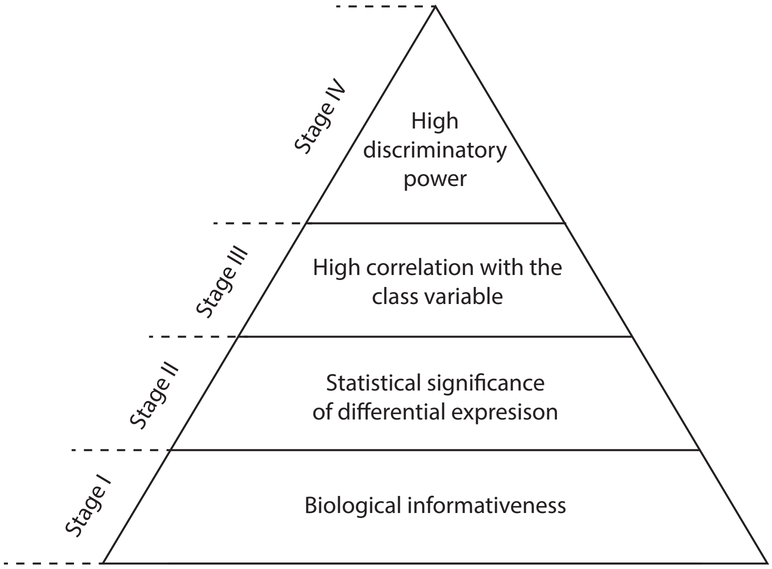
Stage IV
High discriminatory power
Stage III
High correlation with the class variable
Stage II
Statistical significance of differential expresison
Stage I
Biological informativeness

Figure 4.1: Hierarchical organization of the stage-specific forms of relevance

Section 4.1 (Stage I) proposes a new unsupervised filter that selects biological informative genes. In Section 4.2 (Stage II) this set of genes is further screened for statistically significant differentially expressed genes using standard test statistics. In Section 4.3 (Stage III) genes highly correlated with the class variable and minimally correlated with each other are selected from the set of genes selected in Stage II. Section 4.4 presents the final stage (Stage IV) that narrows the space of relevant genes to the subset with the highest discriminatory power. These stages are combined within a methodological framework allowing for unbiased parameter estimation and performance evaluation in Section 4.5. The resulting novel multi-stage feature selection method is compared with the l_1-StaR algorithm on a time-course microarray data from patients suffering from ACS in Section 4.6. Concluding remarks are given in Section 4.7.

4.1 Stage I: Filtering uninformative genes

The first stage of the multi-stage feature selection method consists of removing uninformative genes. In Chapter 2 it was shown that many genes either generate noisy signals or show little or no variation across conditions. Inappropriate filtering of these uninformative genes can induce bias in statistical analyses of differential gene expression (Bourgon et al., 2010a). Conversely, appropriate combinations of unsupervised filters and test statistics (see the marginal independence criterion in Section 2.4.1 of Chapter 2) can substantially increase the power to detect true differences (Bourgon et al., 2010a). This section proposes a novel unsupervised filtering method for microarray data summarized with multi-mgMOS (Liu et al., 2005).

Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ denote the gene expression matrix, where n represents the number of arrays and m represents the cardinality of the set of genes $\mathcal{S}^0 = \{X_1, X_2, \ldots, X_m\}$. The entries x_{ij} of $\mathbf{X}$ represent gene expression measurements summarized with multi-mgMOS. For each x_{ij}, multi-mgMOS additionally estimates the technical variance σ_{ij}^2 representing the uncertainty of x_{ij}. The aim of the unsupervised filter proposed here is to remove genes whose technical variance (the variance of the sampling distribution of the summarized gene expression levels) exceeds their biological variance (the variance of the gene expression levels across biological replicates). The technical variance (σ_t^2) of gene j across n arrays was approximated with:

$$\sigma_t^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_{ij}^2 \tag{4.1}$$

while the biological variance (σ_b^2) was approximated with:

$$\sigma_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \tag{4.2}$$

where

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{4.3}$$

The gene j was rejected if :

$$\frac{\sigma_b^2}{\sigma_t^2} \leq 1 \tag{4.4}$$

This procedure reduces $\mathcal{S}^0$ to $\mathcal{S}^I$, where the latter contains the gene satisfying (4.4). If the technical variance σ_t^2 is the same across genes then the signal-to-noise ratio (4.4) is equivalent to the overall variance filter. Note the similarity with the I/NI calls filter presented in Chapter 2, which also represents a ratio between the biological variance and the technical variance of a gene, where the latter quantity is estimated internally during gene expression level summarization using FARMS.

The signal-to-noise ratio proposed here is equivalent to applying a variance filter using a cut-off threshold specific to each probe set. For permutation invariant and t-test statistics, the variance filter is marginally independent of the statistics (Bourgon et al., 2010a).

4.2 Stage II: Gene subset selection using statistical tests

The second stage of the multi-stage feature selection method consists of selecting from $\mathcal{S}^I$ a subset of genes $\mathcal{S}^{II}$ that are statistically significant differentially expressed between case and control groups. The Welch's t-test and the Wilcoxon's rank-sum test are used to asses differential gene expression. To correct for the multiplicity problems arising when testing multiple hypotheses, control of the pFDR was used following the implementation guidelines proposed by Storey and Tibshirani (2003). The statistical tests together with the pFDR controlling procedure are presented in the following sections.

4.2.1 The Welch's *t*-test

The Welch's t-test (Welch, 1947) represents a variant of the standard t-test for equality of two population means. Like the standard t-test, it is a parametric test that assumes the data are independently sampled from normal distributions. Unlike the standard t-test, it assumes that the population distributions have unequal variances.

Let $\mathcal{C} = \{c_i | c_i \in \{1,2\} \text{ for } i = 1 \ldots n\}$ denote the class variable for the gene expression matrix $\mathbf{X}$, where c_i represents the class label for the ith array. Also, let $\mathcal{C}_1 = \{i | c_i = 1\}$ with cardinality k_1 and $\mathcal{C}_2 = \{i | c_i = 2\}$ with cardinality k_2 denote the sets of array indices for each class. The Welch's t-statistic for X_j is:

$$T_j = \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}}{\sqrt{\frac{\sigma^2_{(1)j}}{k_1} + \frac{\sigma^2_{(2)j}}{k_2}}} \tag{4.5}$$

where

$$\bar{x}_j^{(l)} = \frac{1}{k_l} \sum_{i \in C_l} x_{ij} \tag{4.6}$$

and

$$\sigma^2_{(l)j} = \frac{1}{k_l - 1} \sum_{i \in C_l} (x_{ij} - \bar{x}_j^{(l)})^2 \tag{4.7}$$

for $l \in \{1,2\}$. Under the null hypothesis of no difference between the means of the two populations, the sampling distribution of the Welch's t-statistic is approximated with a Student's t distribution with degrees of freedom:

$$d_j = \frac{\left(\frac{\sigma^2_{(1)j}}{k_1} + \frac{\sigma^2_{(2)j}}{k_2}\right)^2}{\frac{1}{k_1-1}\left(\frac{\sigma^2_{(1)j}}{k_1}\right)^2 + \frac{1}{k_2-1}\left(\frac{\sigma^2_{(2)j}}{k_2}\right)^2} \tag{4.8}$$

The Welch's t-test was used in this work to test the null hypothesis that the gene expression means of the case and control populations are equal against the alternative hypothesis that the means are unequal (with no particular concern on the directionality of the difference).

4.2.2 The Wilcoxon's rank-sum test

The Wilcoxon's rank-sum test (Hollander et al., 2013) is a nonparametric test for equality of two population medians. To compute the rank-sum statistic for X_j, the measurements x_{ij} are sorted in ascending order and ranked based on their position in the series. The rank given to equal measurements is the rank that each individual measurement would have gotten in the absence of ties, divided by the number of competing measurements for that rank. Let w_i denote the rank given to x_{ij}. The Wilcoxon's rank-sum statistic is:

$$W_j = \sum_{i \in C_l} w_i - k_1 k_2 - \frac{k_l\,(k_l + 1)}{2} \tag{4.9}$$

where l corresponds to the class with the smallest number of samples. When the compared groups have equal number of samples, either group can be used for computing W_j. Under the null hypothesis, the sampling distribution of the Wilcoxon's rank-sum tests tends asymptotically to the normal distribution $\mathcal{N}\left(0, k_1 k_2 \left(k_1 + k_2 + 1\right)/12\right)$ as $\min\left(k_1, k_2\right)$ approaches infinity.

The Wilcoxon's rank-sum test was used in this work to test the null hypothesis that the gene expression medians of the case and control populations are equal against the alternative hypothesis that the medians are different (with no particular concern on the directionality of the difference).

4.2.3 pFDR control and q-value estimation

Chapter 2 introduced the q-value of a statistic (associated with a feature) as the expected proportions of false positives among features with more extreme statistics. Calling significant features (from $\mathcal{S}^I$) with q-values less that a predefined threshold α produces a subset of features $\mathcal{S}^{II}$ containing an expected proportion of false positives equal to α. When the number of hypotheses tested is large, as is often the case for gene expression data, this procedure is equivalent to controlling the pFDR at level α (Storey and Tibshirani, 2003). The following algorithm proposed by Storey and Tibshirani (2003) provides conservative estimates for the true q-values.

Step 1. Let $p_1 \leq p_2 \leq \cdots \leq p_m$ denote the ordered set of p-values resulted from testing m hypotheses (one for each gene in the gene expression matrix $\mathbf{X}$).

Step 2. Calculate the proportion of true null hypotheses

$$\hat{\pi}_0\left(\delta\right) = \frac{\#\left\{p_j > \delta\right\}}{m\left(1 - \delta\right)}$$

for a range of δ taken from $[0, 1]$.

Step 3. Fit a cubic spline $\hat{f}$ with 3 degrees of freedom to $\hat{\pi}_0\left(\delta\right)$ on δ.

Step 4. Set the estimate of π_0 to be

$$\hat{\pi}_0 = \hat{f}\left(1\right)$$

representing the least biased estimate of the proportion of true null hypotheses ($\delta = 1$).

Step 5. Calculate

$$\hat{q}(p_m) = \hat{\pi}_0 \cdot p_m$$

Step 6. Calculate

$$\hat{q}(p_i) = \min\left(\frac{\hat{\pi}_0 m \cdot p_i}{i}, \hat{q}(p_{i+1})\right)$$

The resulting estimates $\hat{q}(p_1) \leq \hat{q}(p_1) \leq \cdots \leq \hat{q}(p_m)$ are simultaneously less than or equal to the true q-values (Storey et al., 2004). This conservative property prevents the proportion of false positives from being underestimated.

4.3 Stage III: Gene subset selection using mRMR

The third stage of the multi-stage feature selection method consists of selecting from $\mathcal{S}^{II}$ a subset of genes $\mathcal{S}^{III}$ maximally correlated with the class variable $\mathcal{C}$ and minimally correlated with each other. This step reduces the dimensionality of the features set selected at stage two by choosing highly relevant genes from clusters of co-regulated genes.

To this end the mRMR algorithm presented in Chapter 2 with the MIQ formulation is used to select a subset $\mathcal{S}^{III}$ of r relevant genes. The choice of the MIQ formulation is supported by experimental results showing that mRMR performs better on discretized data than on continuous data and that the MIQ formulation selects subsets leading to superior prediction accuracy than the MID formulation when discretized data is used (Ding and Peng, 2005).

In this work, gene expression data is discretized using the class-attribute contingency coefficient (CACC) algorithm (Tsai et al., 2008) presented below.

4.3.1 The CACC discretization algorithm

Before presenting the steps of the algorithm, the standard teminology and notation for data discretization will be introduced. Let $\mathbf{x}_j$ denote the column j of the gene expression matrix $\mathbf{X}$ and consider the more general case when the class variable $\mathcal{C} = \{c_i | c_i \in \{1, \ldots, c\} \text{ for } i = 1 \ldots n\}$ has $c \geq 2$ classes. The continuous attribute $\mathbf{x}_j$ can be partitioned into a finite set of k adjacent intervals using the discretization scheme:

$$D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{k-1}, d_k]\} \tag{4.10}$$

where d_0 and d_k represents the minimum and the maximum values of $\mathbf{x}_j$, respectively, and $d_1, \ldots d_{k-1}$ denote the cut-points. The class variable $\mathcal{C}$ and the discretization scheme D define a quanta matrix (contingency table) (Ching et al., 1995) for feature X_j. This structure is shown in Table 4.1, where q_{il} represents the number of samples belonging to class i and interval $(d_{l-1}, d_l]$, n_{+l} represents the total number of samples in the interval $(d_{l-1}, d_l]$ and n_{i+} represents the number of measurements in class i.

Table 4.1: Quanta matrix for feature X_j

Class	Interval					Sum of class
	$[d_0, d_1]$	$\dots$	$(d_{l-1}, d_l]$	$\dots$	$(d_{k-1}, d_k]$	
0	q_{11}	$\dots$	q_{1l}	$\dots$	q_{1k}	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	q_{i1}	$\dots$	q_{il}	$\dots$	q_{ik}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c	q_{c1}	$\dots$	q_{cl}	$\dots$	q_{ck}	n_{c+}
Total	n_{+1}	$\dots$	n_{+l}	$\dots$	n_{+k}	n

A good discretization algorithm outputs a discretization scheme D that ensures a high degree of interdependence between the discrete vector $\mathbf{x}_j^D$ and the class variable $\mathcal{C}$ thus preserving the original distribution of the data (Su and Hsu, 2005). Data discretization methods can be divided into two major classes (Liu et al., 2002): top-down and bottom-up.

- Top-down methods operate by splitting the range of $\mathbf{x}_j$. Starting with a discretization scheme containing an empty set of cut-points, candidate cut-points are evaluated and successively added to the discretization scheme until a stopping criterion is met.

- Bottom-up discretization methods, on the other hand, operate by merging intervals. Starting with a discretization scheme containing the complete list of distinct values of $\mathbf{x}_j$ as cut-points, adjacent intervals are merged until a stopping criterion is met. Bottom-up methods usually have larger computational complexity than top-down methods (Tsai et al., 2008).

The CACC algorithm is a top-down method that measures the interdependency between $\mathbf{x}_j^D$ and the class variable $\mathcal{C}$ using the criterion:

$$cacc = \sqrt{\frac{y'}{y' + n}} \tag{4.11}$$

where

$$y' = \frac{n}{\log k} \sum_{i=1}^{c} \sum_{l=1}^{k} \frac{q_{il}^2}{n_{i+} n_{+l}} \tag{4.12}$$

Good discretization schemes are associated with high *cacc* values (Tsai et al., 2008).

The CACC algorithm searches for D maximizing (4.11) using the following computational procedure.

Step 1. Find the minimum d_0 and the maximum d_k of $\mathbf{x}_j$

Step 2. Sort the distinct elements of $\mathbf{x}_j$ in ascending order

Step 3. Calculate the midpoints of all adjacent pairs in the set of ordered elements

Step 4. Initialize all possible cut-points B with the midpoints calculated at Step 3.

Step 5. Set the initial discretization scheme $D : \{[d_0, d_k]\}$ and $cacc_{Global} = 0$.

Step 6. Initialize $k = 1$

Step 7. Add each cut-point from $B \setminus D$ to D and compute the *cacc* value

Step 8. Select the discretization scheme D' with the highest *cacc* value.

Step 9. If $(cacc > cacc_{Global})$ or $(k < c)$

1. Replace D with D'
2. Set $cacc_{Global} = cacc$
3. Set $k = k + 1$
4. Go to Step 7

Step 10. Else, set $D' = D$ and terminate.

The CACC algorithm compensates for the shortcomings of class-attribute interdependence maximization (CAIM) (Kurgan and Cios, 2004), the state-of-the-art top-down discretization method. Specifically, CACC is not biased towards the class with the highest number of samples and avoids outputting a discretization scheme in which the number of intervals is very close to the number of classes.

4.4 Stage IV: Gene subset selection using SVMs

The last stage of the multi-stage feature selection algorithm consists of selecting from $\mathcal{S}^{III}$ a minimal subset of genes $\mathcal{S}^{IV}$ that best discriminate between groups. To this end, a SVM-based wrapper approach was used to identify the combination of genes with the highest prediction accuracy. Specifically the discriminatory power of genes was assessed during k_{svm}-fold cross-validation runs of the SVM classifier:

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b \tag{4.13}$$

with Gaussian kernel:

$$\kappa\left(\mathbf{x},\mathbf{x}'\right) = \phi(\mathbf{x})^\top \phi(\mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma^2}} \tag{4.14}$$

where k_{svm} is a user specified parameter, the vector of weights $\mathbf{w}$ and the bias parameter b are optimized during the training stage while the scaling factor σ is tuned before training. Methods for tuning σ and the SVM regularization parameter C are discussed in the next section.

To generate candidate subsets of features for evaluation, sequential forward selection is used. This search strategy is preferred to backward elimination due to its simplicity and computational efficiency. Additionally, forward selection has the advantage of identifying the gene with the highest discriminatory power as opposed to backward selection which normally ignores it (Guyon and Elisseeff, 2003). The following procedure combines sequential forward selection with k_{svm}-fold cross validation to select the genes that maximize the prediction performance of the SVM classifier (4.13):

Required The set of optimal classifier parameters $\mathcal{V}_{opt} = \left\{\sigma_{opt}, C_{opt}\right\}$

Step 1. Set $\mathcal{S}^{IV} = \emptyset$ and the associated accuracy $Acc_{Global} = 0$

Step 2. For each feature in $\mathcal{S}^{III} \setminus \mathcal{S}^{IV}$

1. Add it to $\mathcal{S}^{IV}$
2. Split the rows of $\mathbf{X}$ (samples) with columns $\mathcal{S}^{IV}$ into k_{svm} parts
3. Train the SVM classifier using $\mathcal{V}_{opt}$ on $k_{svm} - 1$ parts
4. Test the SVM classifier on the hold-out part
5. Repeat line (3) and (4) for all k_{svm} possible choices of the held-out part
6. Count N_{hit}, the total number of correctly classified instances
7. Set $Acc_{Local} = \frac{N_{hit}}{n}$

Step 3. Select the subset $\mathcal{S}^{IV}$ with maximum Acc_{Local}

Step 4. If $(Acc_{Local} > Acc_{Global})$

1. Set $Acc_{Global} = Acc_{Local}$
2. Go to Step 2 with the new $\mathcal{S}^{IV}$

Else output $\mathcal{S}^{IV}$ and Acc_{Global} from Step 2 and terminate.

4.5 Model selection and reliable performance evaluation

The previous sections introduced the stages of the multi-stage feature selection method. If these stages are sequentially applied on the whole data set $\mathbf{X}$, given user-supplied parameters $\mathcal{P} = \{\alpha, r, k_{svm}\}$ and optimal classifier parameters $\mathcal{V}_{opt} = \{\sigma_{opt}, C_{opt}\}$, the subset of genes with the highest discriminatory power is returned. However, the cross-validation error of the classifier will be a biased estimate of the true error incurred when testing the classifier on independent data. This feature selection bias occurs when the test instances are also used to select differentially expressed genes (Ambroise and McLachlan, 2002). Additionally, tuning the classifier parameters σ and C can induce bias in the estimation of the true error if it is performed external to the training stage of the classifier i.e. borrows information from the test instances (Varma and Simon, 2006).

Section 4.5.1 discusses the nested cross-validation design as a solution to the feature selection bias and the parameter selection bias. This methodological framework allowing for appropriate model selection and reliable performance evaluation will serve as a template for combining the feature selection stages of the multi-stage algorithm in Section 4.5.2.

4.5.1 The nested cross-validation design

The methodological framework for appropriate model selection and reliable performance evaluation consists of two nested cross-validation loops (Statnikov et al., 2005). The external loop addresses the feature selection bias and is used to obtain a conservative estimate of the true classification performance while the internal loop addresses the parameters selection bias and is used for model selection. These steps are discussed in more detail below.

The external cross-validation loop

To correct for the feature selection bias, supervised feature selection must be performed within each fold of an external cross-validation loop (Ambroise and McLachlan, 2002, Varma and Simon, 2006), as shown in the following procedure:

Required The set of optimal classifier parameters $\mathcal{V}_{opt} = \{\sigma_{opt}, C_{opt}\}$

Step 1. Split the rows of $\mathbf{X}$ into k_{ext} random parts of roughly equal sizes

Step 2. Perform feature selection on $\mathbf{X}_{Training}$ consisting of $k_{ext} - 1$ parts of $\mathbf{X}$

Step 3. Train the classifier on $\mathbf{X}_{Training}$ using $\mathcal{V}_{opt}$ and features selected at Step 2

Step 4. Test the classifier on the hold-out part $\mathbf{X}_{Test}$

Step 5. Repeat Step 2 - Step 4 for all k_{ext} possible choices of $\mathbf{X}_{Test}$

Step 6. Return the cross-validation estimate of the prediction error ϵ_{cv}

The cross-validation estimate of the prediction error representing the ratio of the number of misclassified examples to the total number of examples in $\mathbf{X}$ (Stone, 1974) will be corrected for the selection bias as the cross-validation loop was external to the feature selection process (Ambroise and McLachlan, 2002). However, this estimate will be still subject to its own bias. Specifically, the cross-validation estimate of the prediction error will be unbiased only for data subsets of $n - n/k_{ext}$ size (Bengio and Grandvalet, 2005).

The leave-one-out-CV (LOOCV) error corresponding to $k_{ext} = n$ is nearly unbiased as an estimator of the true error given the whole data $\mathbf{X}$ (Braga-Neto and Dougherty, 2004). The widely used 10-fold cross-validation setup provides a more biased but less variable estimate of the true prediction error (Ambroise and McLachlan, 2002).

Depending on the stability of the feature selection method (i.e. sensitivity to variations in the training set)(Kalousis et al., 2007), different subsets sharing at most only a few common features may be selected in Step 2 of each fold. Ranking these features in terms of their frequency of occurrence can shade light on their consistency to differentiate between groups.

The internal cross-validation loop

To correct for the parameter selection bias, tuning of the classifier parameters must be done within an internal cross-validation loop for each $\mathbf{X}_{Training}$ associated with the external cross-validation loop (Statnikov et al., 2005, Varma and Simon, 2006). The following procedure uses a grid-search approach for parameter tuning.

Required Range of classifier parameters $\mathcal{V} = \left\{ \left(\sigma_i, C_j\right) | i = 1, \ldots, s; j = 1, \ldots, q \right\}$

Step 1. Split $\mathbf{X}_{Training}$ into k_{int} random parts of roughly equal sizes

Step 2. For each pair $\left(\sigma_i, C_j\right)$

1. Train the classifier on $k_{int} - 1$ parts
2. Test the classifier on the hold-out part
3. Repeat (1) and (2) for all k_{int} possible choices of the hold-out part

Step 3. Select the pair of parameters $\mathcal{V}_{opt} = \left\{\sigma_{opt}, C_{opt}\right\}$ on $\mathbf{X}_{Training}$ minimizing the cross-validation error

Alternatively, a simplex search approach (Lagarias et al., 1998) could be used to automatically select the optimum classifier parameter, as described in the following procedure:

Required Range of random initial values for the classifier parameters: $\mathcal{V}_0 = \{(\sigma_{0,i}, C_{0,i})\}_{i=1\ldots s}$

Step 1. Split $\mathbf{X}_{Training}$ into k_{int} random parts of roughly equal sizes

Step 2. Use a simplex search algorithm that starts from $(\sigma_{0,i}, C_{0,i})$ and returns the set of parameters $\mathcal{V}_{opt} = \left\{\sigma_{opt}, C_{opt}\right\}$ minimizing the cross-validation error

Step 3. Repeat Step 2 for each pair $(\sigma_{0,i}, C_{0,i})$

Step 4. Select $\mathcal{V}_{opt}$ with the minimum cross-validation error

The optimal set of parameters $\mathcal{V}_{opt}$ resulting from either of the search procedures is later used for training the classifier on $\mathbf{X}_{train}$. Note that different sets of parameters may be selected for each $\mathbf{X}_{train}$. If the goal is to identify only one set of parameters to train the classifier on the whole data $\mathbf{X}$, then either of the two optimization procedures should be applied on $\mathbf{X}$, after am unbiased estimate of the classification performance was obtained using the nested design.

4.5.2 The multi-stage feature selection method

To select a subset of relevant genes and estimate a conservative bound of the classifier's performance on independent data, the supervised feature selection stages (Stage II, Stage III and Stage IV) were combined within a nested cross-validation design, as shown in Figure 4.2 and described in the following procedure. Note that Stage I consists of an unsupervised filter and therefore can be applied on the whole dataset without interfering with performance evaluation.

Required User-supplied parameters $\mathcal{P} = \{\alpha, r, k_{svm}, k_{ext}, k_{int}\}$, range of classifier parameters $\mathcal{V} = \left\{\left(\sigma_i, C_j\right) | i = 1, \ldots, s; j = 1, \ldots, q\right\}$

Step 1. Apply Stage I on $\mathbf{X}$ to select $\mathcal{S}^I$

Step 2. Split $\mathbf{X}$ with columns $\mathcal{S}^I$ into k_{ext} random parts of roughly equal sizes

Step 3. For each $\mathbf{X}_{Training}$ consisting of $k_{ext} - 1$ parts of $\mathbf{X}$

1. Split $\mathbf{X}_{Training}$ into k_{int} random parts of roughly equal sizes
2. For each pair (σ_i, C_j)
    - 2.1. Apply Stage II using α on $k_{int} - 1$ parts to select $\mathcal{S}^{II}$
    - 2.2. Apply Stage III using r (without data discretization) on the same $k_{int} - 1$ parts with columns $\mathcal{S}^{II}$ to select $\mathcal{S}^{III}$
    - 2.3. Train the classifier on $k_{int} - 1$ parts with columns $\mathcal{S}^{III}$
    - 2.4. Test the classifier on the hold-out part
    - 2.5. Repeat (2.1)-(2.4) for all k_{int} possible choices of the hold-out part
3. Select $\mathcal{V}_{opt} = \{\sigma_{opt}, C_{opt}\}$ minimizing the cross-validation error
4. Apply Stage II using α on $\mathbf{X}_{Training}$ to select $\mathcal{S}^{II}$
5. Apply Stage III using r (with data discretization) on $\mathbf{X}_{Training}$ with columns $\mathcal{S}^{II}$ to select $\mathcal{S}^{III}$
6. Apply Stage IV using k_{svm} and $\mathcal{V}_{opt} = \{\sigma_{opt}, C_{opt}\}$ on $\mathbf{X}_{Training}$ with columns $\mathcal{S}^{III}$ to select $\mathcal{S}^{IV}$
7. Train the classifier using $\mathcal{V}_{opt} = \{\sigma_{opt}, C_{opt}\}$ and $\mathcal{S}^{IV}$ on $\mathbf{X}_{Training}$

Step 4. Test the classifier on the hold out part $\mathbf{X}_{test}$

Step 5. Repeat Step 3 and Step 4 for all k_{ext} possible choices of $\mathbf{X}_{Test}$

Step 6. Return the cross-validation estimate of the prediction error ϵ_{cv}

The classifier parameters are tuned using a grid-search approach. Replacing this optimization method with a simplex search is straightforward. Note that feature selection (without data discretization) is also performed in the internal cross-validation loop. This dimensionality reduction step reduces the computational burden associated with training the classifier.

The final set of relevant features $\mathcal{S}_f$ (for which we have a nearly unbiased estimate of the classifier's accuracy $a_{cv} = 1 - \epsilon_{cv}$ when k_{ext} approaches n) is taken as the union of the subsets $\mathcal{S}^{IV}$ selected in the external cross-validation loop. Each feature is assigned a measure of relevance denoted by their frequency of occurrence in the subsets $\mathcal{S}^{IV}$.

Building the final prediction model using all the arrays consists of (Varma and Simon, 2006):

- Tuning the classifier parameters on $\mathbf{X}$ (see Section 3.5.1)
- Selecting relevant features using $\mathbf{X}$ (without partitioning)

- Training the classifier on $\mathbf{X}$ using the best model parameters and set of relevant features

- Estimating a conservative bound for the classifier's performance a_{cv} using the nested cross-validation design

Ambroise and McLachlan (2002) noted that the set of genes selected using all the arrays and the sets of genes selected in the external cross validation loop may share at most a few common genes. Since the estimate a_{cv} directly relates to the sets of genes selected in the external loop, the features for the final prediction model can be selected by applying Stage IV on $\mathbf{X}$ with columns defined by $\mathcal{S}_f$.

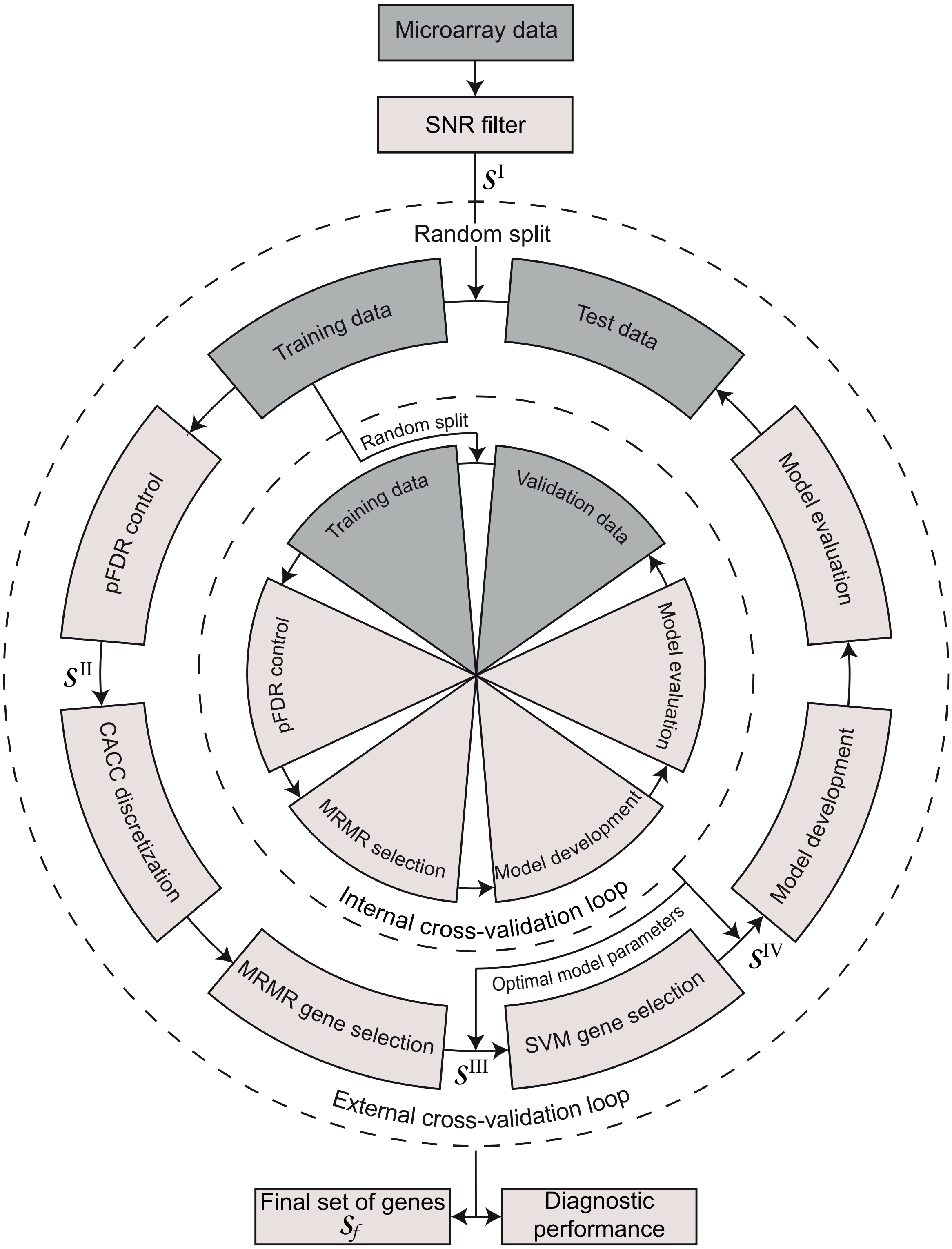

Figure 4.2: Nested cross-validation design of the multi-stage feature selection algorithm.

4.6 Gene selection for ACS classification

The multi-stage feature selection method and the l_1-STaR algorithm were applied on a time-course microarray dataset, provided by Doctor Marta Milo, containing gene expression measurements from patients with ACS to identify differentially expressed genes between the diagnostic groups of the disease. These genes could serve as new biomarkers for diagnosis or direct the identification of drug targets that may ultimately lead to clinical trials.

To this end, two genome-wide expression studies were conducted. The first study compared the aggregated cohort of NSTEMI and STEMI patients, denoted as the MI group, against the UA group. The second study compared the NSTEMI group against the STEMI group.

This section is organized as follows. Section 4.6.1 describes the time-course microarray data set. Section 4.6.2 presents a comparative analysis of the results for the first study obtained using the multi-stage feature selection methods against the findings of l_1-STaR. Section 4.6.3 presents the results for the second study within the same comparative framework. The criteria for comparison are: classification performance, average number of genes used for classification and length of the interval of differential expression.

4.6.1 The ACS dataset

The study cohort consists of 33 patients admitted to Sheffield Teaching Hospitals with chest pain. Based on presenting ECG findings and presence or absence of elevated levels of serum troponin levels, patients were diagnosed as having suffered from NSTEMI$(n = 14)$, STEMI $(n = 8)$ or UA$(n = 11)$. Peripheral blood samples were collected in Tempus tubes at day 1, day 3, day 7, day 30 and day 90 after hospital admission. Blood samples for 12 visits couldn't be obtained (see Table 4.2 for details). Total RNA was isolated from the blood samples using standard protocols. Transcriptomic abundance was measured using Affymetrix Human Genome U133 Plus 2.0 arrays containing 54675 probe sets. A total of 153 microarray experiments were thus performed.

The multi-mgMOS R-package (Pearson et al., 2009) was used to estimate expression levels and standard errors for each probe set from raw microarray data (subsets of data consisting of estimated expression levels and associated standard errors are shown in Table B.1 and Table B.2 of Appendix B). Additionally, robust group-specific gene expression averages and standard errors were obtained for each time point using the probability of positive log ratio (PPLR) (Liu et al., 2006). This later dataset, onwards referred to as the combined dataset, describes the time-course average behaviour for each gene in each of the three diagnostic groups.

Table 4.2: Amount of patients by visits and diagnostic groups

Visit	NSTEMI	STEMI	UA	Total
Day 1	14	8	11	33
Day 3	11	5	9	25
Day 7	14	8	10	32
Day 30	14	8	10	32
Day 90	13	8	10	31
Total	66	37	50	153

4.6.2 Selection of differentially expressed genes between MI and UA

The aim of this study is to identify genes that can distinguish between MI and UA in a time window of 90 days after hospital admission. The following two sections present the results obtained using the multi-stage feature selection method and the l_1-STaR algorithm, respectively.

Results using the multi-stage feature selection method

The multi-stage feature selection method was applied on the 153 arrays from MI and UA patients using the set user parameters $\mathcal{P} = \{\alpha, r, k_{svm}, k_{ext}, k_{int}\} = \{0.01, 200, 10, 10, 9\}$ and range of classifier parameters $\mathcal{V} = \mathcal{R}_\sigma \times \mathcal{R}_C$ where $\mathcal{R}_\sigma = \mathcal{R}_C = \{10^{-3}, 10^{-2.5}, \ldots, 10^{2.5}, 10^3\}$. To account for possible departures from normality of the aggregated case group, the two-tailed Wilcoxon's rank-sum test was used at Stage II. Thresholding q-values at significance level $\alpha = 0.01$ means that in $\mathcal{S}^{II}$, on average, one gene among 100 is expected to be a false positive. The cardinality of $\mathcal{S}^{III}$ was set to the conservative value $r = 200$ on the grounds that tens of genes are usually sufficient to discriminate between groups on microarray datasets (Ding and Peng, 2005, Peng et al., 2005). At Stage IV a stratified 10-fold cross-validation partitioning ($k_{svm} = 10$) of the data was used (stratified partitions contain roughly the same class proportions as in the class variable $\mathcal{C}$). A nested stratified 10-fold cross validation design (Statnikov et al., 2008) corresponding to a 9-fold stratified cross-validation design in the inner loop ($k_{int} = 9$) and a 10-fold stratified cross-validation design in the outer loop ($k_{ext} = 10$) was adopted.

Stage I of the multi-level feature selection method produced a subset $\mathcal{S}^I$ of 29050 potentially informative probe sets. Examples of non-informative and informative probe sets are shown in Figure 4.3. It can be seen that the technical variance of the non-informative probe set exceeds its biological variance. The opposite holds true for the informative probe set.

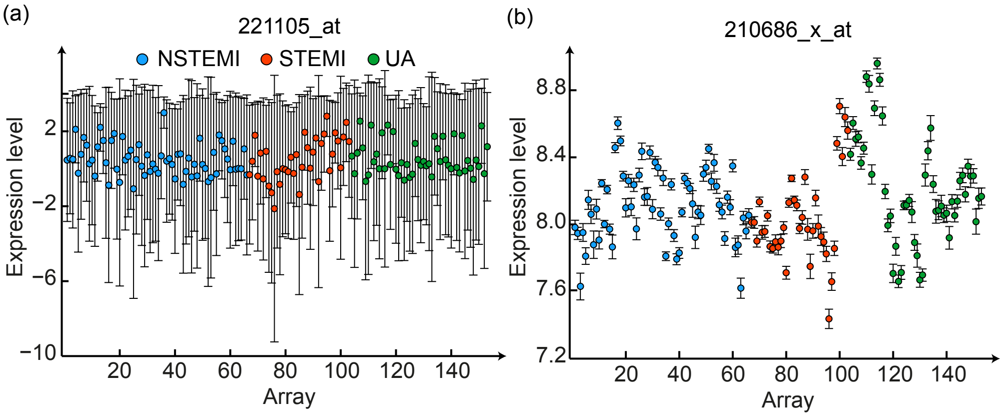

Figure 4.3: Expression levels across arrays for (a) a non-informative probe set and (b) an informative probe set. The bars denote the technical variance of each measurement.

Stage II removed genes with small effect size (Hedges' g score close to 0), as shown in Figure 4.4. The mean densities were computed by averaging the densities of the Hedge's g scores for the subsets $\mathcal{S}^{II}$ and $\mathcal{S}^{III}$ (selected in the external cross-validation loop), respectively. In the calculation of the Hedges' g scores, the MI and the UA groups were taken as the first and second population, respectively. An average of 3189 genes were selected at Stage II (standard deviation = 965.5).

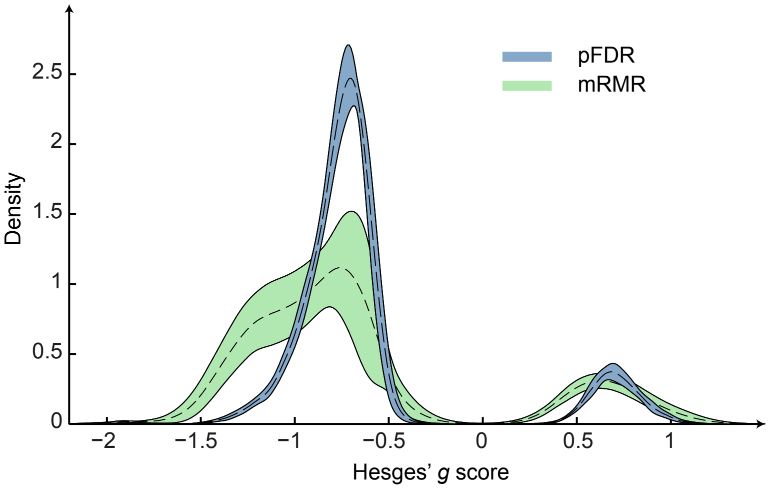

Figure 4.4: Mean densities (dotted lines) of the Hedges' g scores for the subsets of gene selected using pFDR (Stage II) and mRMR (Stage III) in the external cross-validation loop. The continuous lines denote one standard deviation around the mean densities.

The left mode of the mean density associated with the subsets $\mathcal{S}^{II}$ (pFDR) shows that the majority of differentially expressed genes are down-regulated in the MI group compared to the UA group. Also note that the left mode of the mean density associated with the subsets $\mathcal{S}^{III}$ (mRMR) is negatively skewed. This suggests the presence of genes with large effect size highly correlated with the class variable and minimally correlated with the other differentially expressed genes selected at Stage II.

The high correlation with the class variable can be visualized in Figure 4.5 which shows that the time-course expression averages of the genes in one of the subsets $\mathcal{S}^{III}$, belong to group-specific clusters. In addition, in the MI group the vectors of gene expression averages associated with each visit also cluster, indicating that the selected genes have similar time-course profiles in the NSTEMI and STEMI groups. These patterns are consistent across eight of the subsets $\mathcal{S}^{III}$ selected in the external cross-validation loop. The cosine distance (Pang-Ning et al., 2006) was used to measure the similarity between the time-course vectors of gene expression averages (columns in Figure 4.5) while one minus the Pearson correlation coefficient (Goshtasby, 2012) was used for the gene clusters. The distance between cluster trees was measured using the average linkage (Hastie et al., 2009).

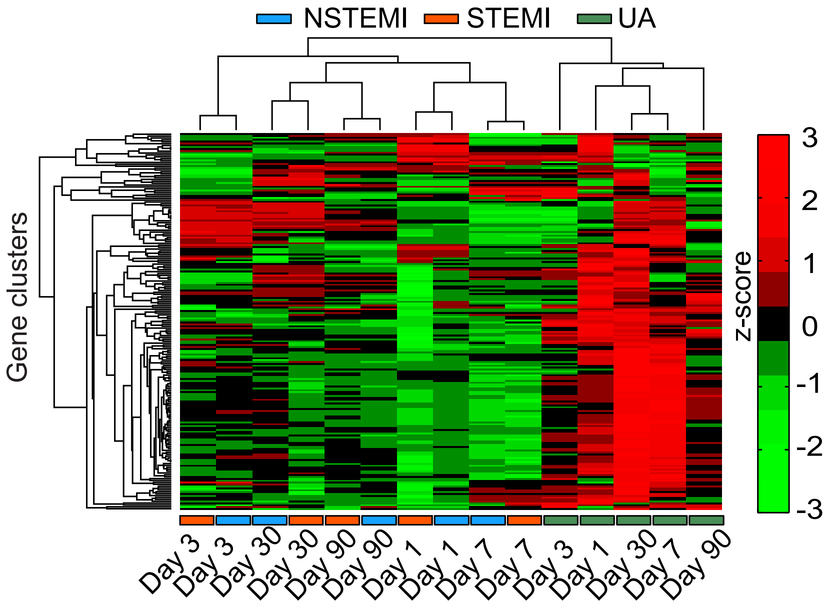

Figure 4.5: Hierarchical clustering of the group-specific expression averages of the genes selected at Stage III in one fold of the external cross-validation loop.

Stage IV selected minimal subsets of genes $\mathcal{S}^{IV}$ with high discriminatory power. These subsets together with their classification performance are listed in Table 4.3.

The union of these subsets ($\mathcal{S}_f$) consist of 36 probe sets, referred to as $X_1 - X_{36}$, corresponding to 36 unique genes. These genes are listed in Table B.3 of Appendix B together with their Hedges' g score (computed across 153 arrays) and frequency of occurrence in the subsets $\mathcal{S}^{IV}$. The optimal classifier parameters for each fold of the external cross-validation loop are listed in Table B.4 of Appendix B.

Table 4.3: Classification performance of the subsets of genes $\mathcal{S}^{IV}$

Fold	Number of arrays			Selected genes
	Training set	Test set	Predicted	
1	138	15	15	X_1, X_2, X_3, X_{11}, X_{12}, X_{16}
2	137	16	16	X_1, X_2, X_3, X_8, X_{10}, X_{17}, X_{32}
3	137	16	16	X_2, X_3, X_4, X_{13}, X_{30}, X_{35}
4	137	16	14	X_1, X_2, X_8, X_{20}, X_{21}, X_{25}, X_{26}, X_{29}
5	138	15	14	X_1, X_2, X_5, X_{19}, X_{22}
6	138	15	13	X_1, X_2, X_6, X_9, X_{27}, X_{34}
7	138	15	15	X_1, X_4, X_7, X_{15}, X_{18}, X_{24}, X_{28}, X_{31}
8	138	15	14	X_1, X_2, X_6, X_7, X_{14}, X_{23}
9	138	15	14	X_1, X_2, X_3, X_5, X_{36}
10	138	15	15	X_1, X_2, X_3, X_4, X_{33}

The classifier achieves a 95.4% accuracy on the test data (total number of correctly classified sampled over total number of test instances) using an average of 6.4 genes. These findings suggest that the classifier can efficiently distinguish between MI and UA during the three months after hospital admission. Although the classification boundary is non-linear, an insight into the discriminatory power of the genes in $\mathcal{S}_f$ can be grasped by inspecting Figure 4.6 which shows the clustering of group-specific expression averages in the linear space spanned by the first two principal components. The principal component analysis also suggests that at each time point there are genes differentially expressed between MI and UA groups. Examples of genes consistently differentially expressed across all the time points are shown in Figure 4.7. The time-course statistics of the remaining genes are shown in Figure B.1 and B.2 of Appendix B.

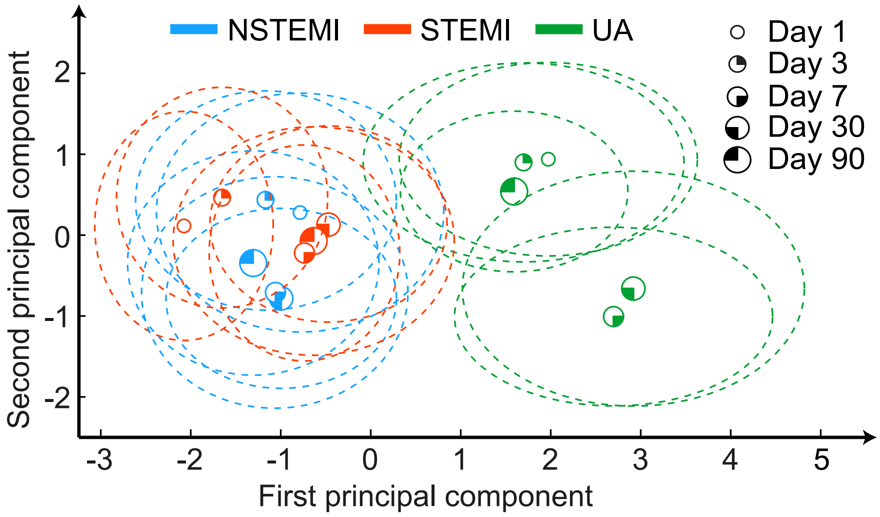

Figure 4.6: Principal component analysis of the combined expression data of $\mathcal{S}_f$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space.

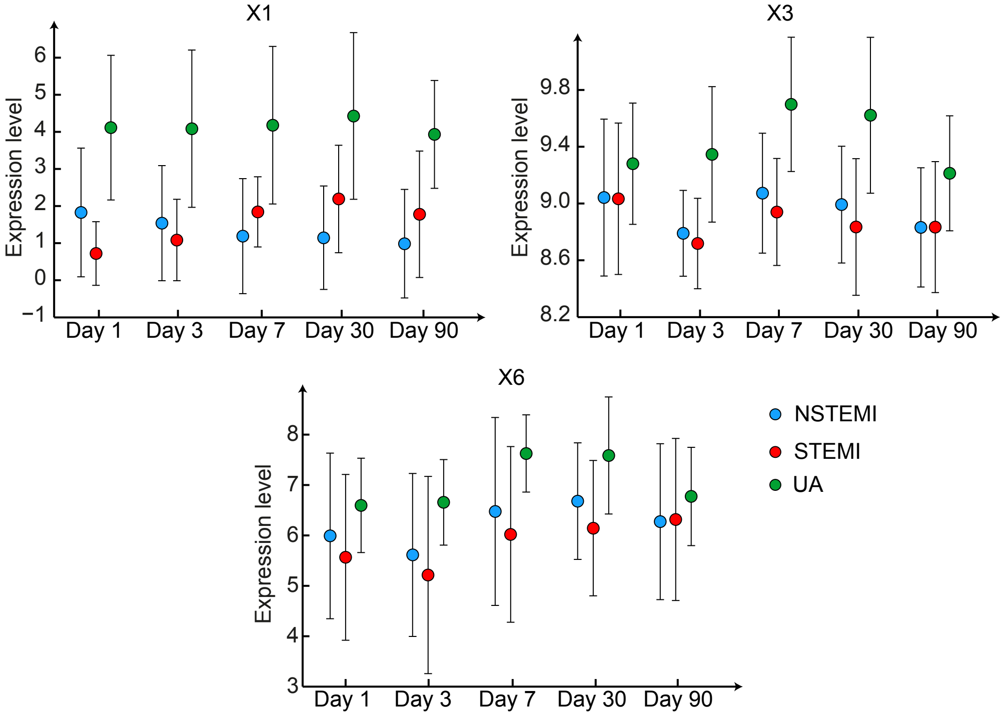

Figure 4.7: Group-specific expression averages across visits for (a) X_1 (WASH1), (b) X_3 (C17orf103) and (c) X_6 (OSBP2). The bars denote one standard deviation around the mean expression levels.

Results using l_1-STaR

The results presented in this section were obtained using the Matlab implementation of the l_1-STaR algorithm, published at (Nitin K Singh, 2013). Table 4.4 lists the parameters of the algorithm.

Table 4.4: l_1-STaR parameters

Parameter	Description	Default value
k_1	Fraction of control group data used for training	0.5
k_2	Fraction of case group data used for training	0.5
h	Trade-off between sensitivity and specificity	0.5
η	Regularization parameter for the L_1-norm SVM	0.5
α	Significance level for the t-test	0.05
N	Number of randomized data splits	80

The authors of l_1-STaR suggest setting k_1 and k_2 to values less than or equal to half the number of examples in the control group and case group, respectively. In this study, the default values for k_1 and k_2 were used (corresponding to half the number of examples). The parameter h was also used with the default values to assign equal importance to predicting the instances of each group. Given that l_1-STaR algorithm doesn't automatically select an optimal value for the regularization parameter, candidate values for η were sequentially chosen from the set $\mathcal{R}_\eta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The significance level α was used with its default value while the number of randomized data splits was set to $N = 30$ for computational reasons.

The l_1-StaR algorithm was applied for each $\eta \in \mathcal{R}_\eta$ on the same training partitions of the external cross-validation loop used for the multi-stage feature selection study and the performance was evaluated on the corresponding test partitions. In the training partitions, only the genes in $\mathcal{S}^I$ were used to prevent the algorithm from selecting relevant but noisy genes. Table 4.5 lists the classification accuracy together with the average number of selected genes for each value of regularization parameter η.

Table 4.5: l_1-STaR performance and number of selected genes for each η

	$\eta = 0.1$	$\eta = 0.3$	$\eta = 0.5$	$\eta = 0.7$	$\eta = 0.9$
Accuracy	98 %	96.7 %	98%	99.3%	98.6%
Number of genes	12.3	12.9	12.6	12.9	12.3

The highest prediction performance (99.3%) is achieved for $\eta = 0.7$. The performance of the subsets of genes selected for each training partition using this value is shown in Table 4.6. The union of these subsets ($\mathcal{S}_f^*$) consists of 34 probe sets, referred to as $X_1^* - X_{34}^*$, corresponding to 33 unique genes. These genes are listed in Table B.5 of Appendix B together with their Hedge's g score (computed across the 153 arrays) and frequency of occurrence in the subsets associated with each training partition. The principal component analysis of the combined expression data of $\mathcal{S}_f^*$ shown in Figure 4.8 provides an insight into the discriminatory power of these genes. As in the case of $\mathcal{S}_f$, the group-specific gene expression averages form clusters in the linear space spanned by the first two principal components.

Table 4.6: Classification performance of the subsets of genes selected by l_1-StaR

Fold	Number of arrays			Selected genes
	Training set	Test set	Predicted	
1	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_{10}^*, X_{11}^*, X_{17}^*, X_{26}^*
2	137	16	16	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_{11}^*, X_{13}^*, X_{14}^*, X_{18}^*
3	137	16	16	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_{10}^*, X_{11}^*, X_{12}^*, X_{25}^*
4	137	16	16	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_9^*, X_{10}^*, X_{20}^*, X_{21}^*, X_{23}^*
5	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_{10}^*,X_{12}^*,X_{18}^*, X_{27}^*,X_{30}^*,X_{31}^*
6	138	15	14	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_9^*, X_{13}^*, X_{14}^*, X_{15}^*, X_{16}^*, X_{32}^*, X_{34}^*
7	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_9^*, X_{10}^* ,X_{11}^*, X_{12}^*, X_{13}^*, X_{16}^*
8	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_8^*, X_9^*, X_{11}^*, X_{15}^*
9	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*, X_9^*, X_{19}^*, X_{24}^*, X_{28}^*, X_{33}^*
10	138	15	15	X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_9^*, X_{10}^*, X_{12}^*, X_{13}^*, X_{14}^*, X_{17}^*, X_{19}^*, X_{22}^*, X_{29}^*

The l_1-StaR algorithm uses, on average, twice the number of genes selected by the multi-stage method to increase the classification performance by 3.9%. However, most of these genes don't exhibit long-term differential expression but distinguish between the groups only at specific time-points, as shown in Figure D.3 and Figure D.4 of Appendix D, where the dynamic profiles of the genes $X_1^* - X_{20}^*$ are presented. The remaining genes exhibit similar overlapping profiles. In comparison, Figure D.1 and Figure D.2 show that the genes $X_1 - X_{20}$ selected by the multi-stage method are differentially expressed over multiple time-points which strengthen their utility as potential biomarkers. The gene expression dynamics were modelled using a technique described in Chapter 6. Thus, the l_1-StaR algo-

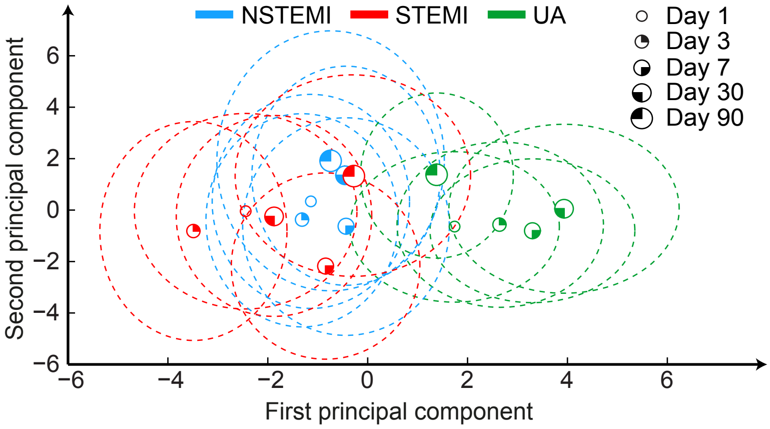

Figure 4.8: Principal component analysis of the combined expression data of $\mathcal{S}_f^*$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space.

rithm leads to better classification performance whereas the multi-stage method, whilst providing a comparable classification performance, appears to be more appropriate for biomarker discovery.

4.6.3 Selection of differentially expressed genes between NSTEMI and STEMI

The aim of this study is to identify genes that can help distinguish between NSTEMI and STEMI in a time window of 90 days after hospital admission. The following two sections present the results obtained using the multi-stage feature selection method and the l_1-STaR algorithm, respectively.

Results using the multi-stage feature selection method

The multi-stage feature selection method was applied on the 103 arrays from NSTEMI and STEMI patients with the same set of user parameters $\mathcal{P}$, range of classifier parameters $\mathcal{V}$ and settings for the nested-cross validation used for the first study. The Welch's t-test was used at Stage II to asses differential gene expression.

Stage I of the multi-level feature selection method produced a subset $\mathcal{S}^I$ of 27910 potentially informative probe sets. Figure 4.9 shows an example of a non-informative and an informative probe set, respectively.

The subsets $\mathcal{S}^{II}$ of differentially expressed genes selected in each fold of the external cross-validation loop contain, on average, only a small proportion of genes with small effect size (Hedges' g score close to 0). This can be seen by inspect-

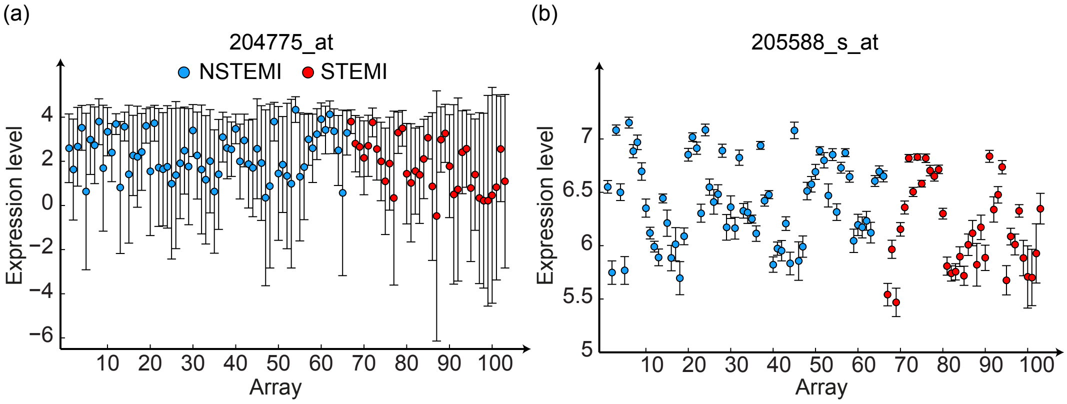

Figure 4.9: Expression level across arrays for (a) a non-informative probe set and (b) an informative probe set. The bars denote the technical variance of each measurement

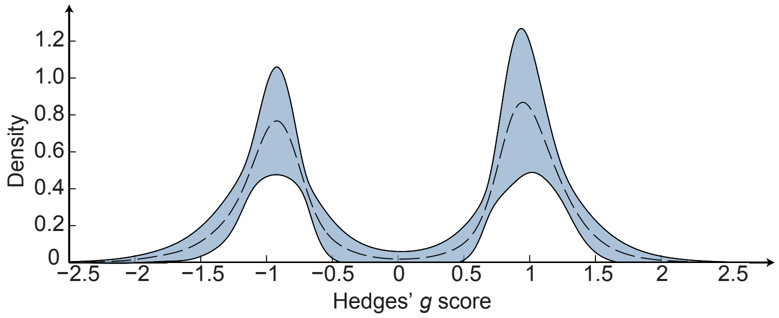

Figure 4.10: Mean density (dotted line) of the Hedges' g scores for the subsets of gene selected using pFDR (Stage II) in the external cross-validation loop. The continuous lines denote one standard deviation around the mean density.

ing Figure 4.10, which shows the mean density of the Hedges' g scores for these subsets. In the calculation of the Hedges' g scores, the NSTEMI and the STEMI groups were taken as the first and second population, respectively. An average of 187.40 genes were selected at Stage II (standard deviation = 103.85). Only subsets $\mathcal{S}^{II}$ of more than 200 genes were screened at Stage III. The remaining subsets were directly advanced to Stage IV. The subsets $\mathcal{S}^{III}$ together with the subsets $\mathcal{S}^{II}$ of cardinality less than 200 contain genes highly correlated with the class variable. Figure 4.11 shows that the time-course expression averages of the genes in one of these subsets belong to group-specific clusters. This pattern is consistent across nine of the subsets. The clustering was performed using the same similarity metrics adopted for the first study.

The subsets $\mathcal{S}^{IV}$ selected at Stage IV together with their classification performance are listed in Table 4.7. The union of these subsets ($\mathcal{S}_f$) consist of 21 probe sets, referred to as $X_1 - X_{21}$, corresponding to 20 unique genes. These probe sets

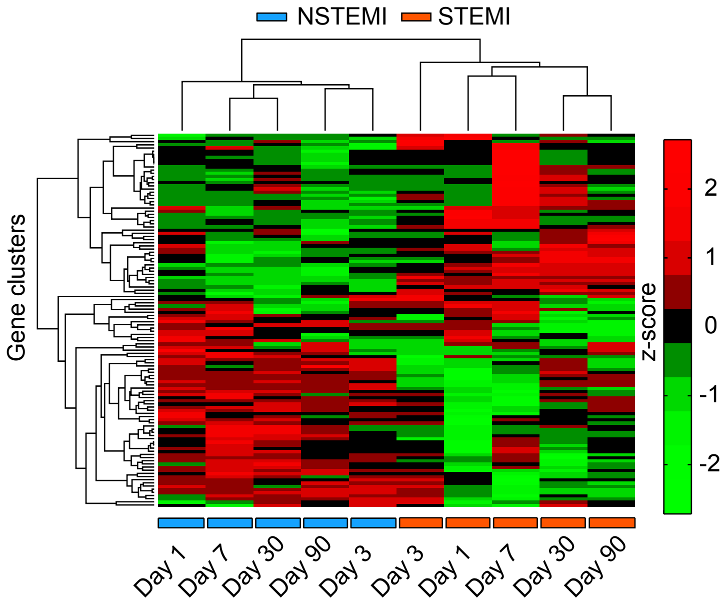

Figure 4.11: Hierarchical clustering of the group-specific expression averages of the genes selected at Stage II in one fold of the external cross-validation loop.

are listed in Table B.6 of Appendix B together with their Hedges' g score (computed across the 103 arrays) and frequency of occurrence in the fold-specific subsets. The optimal classifier parameters for each fold of the cross-validation loop are listed in Table B.7 of Appendix B.

Table 4.7: Classification performance for the subsets of genes $\mathcal{S}^{IV}$

Fold	Number of arrays			Selected genes
	Training set	Test set	Predicted	
1	93	10	10	X_1, X_4, X_6, X_8
2	92	11	11	X_1, X_9, X_{12}, X_{16}
3	92	11	11	X_1, X_2, X_3, X_{17}, X_{21}
4	92	11	11	X_1, X_2, X_3, X_6
5	93	10	9	X_4, X_7, X_8, X_{15}
6	93	10	10	X_1,X_2,X_3,X_5
7	93	10	10	X_1, X_9, X_10, X_{14}, X_{20}
8	93	10	10	X_2, X_4 ,X_5, X_6, X_{13}
9	93	10	10	X_1, X_4, X_6, X_{11}, X_{18}, X_{19}
10	93	10	10	X_1, X_2, X_3, X_5

The classifier achieves a 99% accuracy on the test data using an average of 4.5 genes. These findings suggest that the classifier can efficiently distinguish between NSTEMI and STEMI during the three months after hospital admission. Figure 4.12 shows the clustering of group-specific expression averages for all the

21 probe-sets in the linear space spanned by the first two principal components. The principal component analysis also suggest that at each time point there are genes differentially expressed between NSTEMI and STEMI groups. Examples of genes consistently differentially expressed across all the time points are shown in Figure 4.13. The time-course statistics of the remaining genes are shown in Figure B.3 and B.4 of Appendix B.

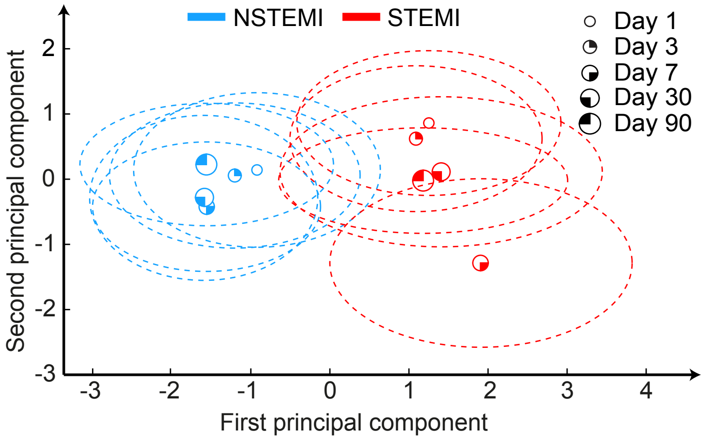

Figure 4.12: Principal component analysis of the combined expression data of $\mathcal{S}_f$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space.

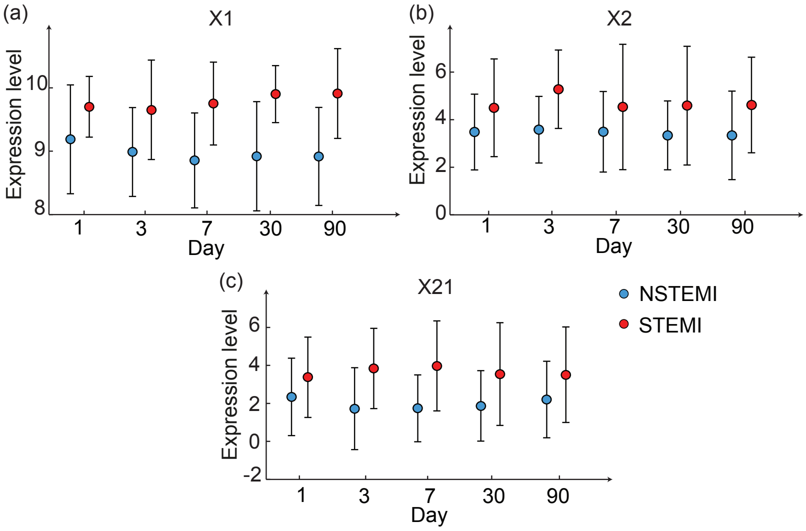

Figure 4.13: Group-specific expression averages across visits for (a) X_1 (HLA-DQB1), (b) X_2 (MAPK8Ip1) and (c) X_{21} (LRRC37A). The bars denote one standard deviation around the mean expression levels.

Results using l_1-STaR

The l_1-StaR algorithm was applied on the external training partitions of the multi-stage feature selection study using the parameters settings discussed in the MI vs. UA study. The training partitions contained only the genes in $\mathcal{S}^I$, to prevent the algorithm from selecting relevant but noisy genes. The performance of the subsets of genes selected on each training set was evaluated on the corresponding test set. Table 4.8 lists the classification accuracy together with the average number of selected genes for each value of regularization parameter η.

Table 4.8: l_1-STaR performance and number of selected genes for each η

	$\eta = 0.1$	$\eta = 0.3$	$\eta = 0.5$	$\eta = 0.7$	$\eta = 0.9$
Accuracy	97 %	99 %	99%	98%	100%
Number of genes	6	5.3	5.6	5.6	4.9

The highest prediction performance (100%) is achieved for $\eta = 0.9$. The performance of the subsets of genes selected for each training partition using this value is shown in Table 4.9. The union of these subsets ($\mathcal{S}_f^*$) consists of 11 probe sets, referred to as $X_1^* - X_{11}^*$, corresponding to 10 unique genes. These probe sets are listed in Table B.8 of Appendix B together with their Hedge's g score (computed across the 103 arrays) and frequency of occurrence in the subsets associated with each training partition. The principal component analysis shown in Figure 4.8 provides an insight into the discriminatory power of the genes in $\mathcal{S}_f^*$.

Table 4.9: Classification performance of the subsets of genes selected by l_1-StaR

Fold	Number of arrays			Selected genes
	Training set	Test set	Predicted	
1	93	10	10	$X_1^*, X_2^*, X_3^*, X_4^*, X_7^*$
2	92	11	11	$X_1^*, X_2^*, X_3^*, X_4^*, X_5^*$
3	92	11	11	$X_1^*, X_2^*, X_3^*, X_6^*$
4	92	11	11	$X_1^*, X_2^*, X_3^*, X_4^*$
5	93	10	10	$X_1^*, X_2^*, X_3^*, X_6^*, X_7^*$
6	93	10	10	$X_1^*, X_2^*, X_3^*, X_4^*, X_5^*$
7	93	10	10	$X_1^*, X_2^*, X_3^*, X_4^*, X_7^*, X_{11}^*$
8	93	10	10	$X_1^*, X_2^*, X_3^*, X_5^*, X_6^*$
9	93	10	10	$X_1^*, X_2^*, X_3^*, X_5^*$
10	93	10	10	$X_1^*, X_2^*, X_3^*, X_4^*, X_9^*, X_{10}^*$

While the multi-stage method misclassifies one array in one of the ten testing partitions, the l_1-StaR algorithm identifies subsets of genes that perfectly discriminate between the NSTEMI and STEMI groups on all testing partitions. However,

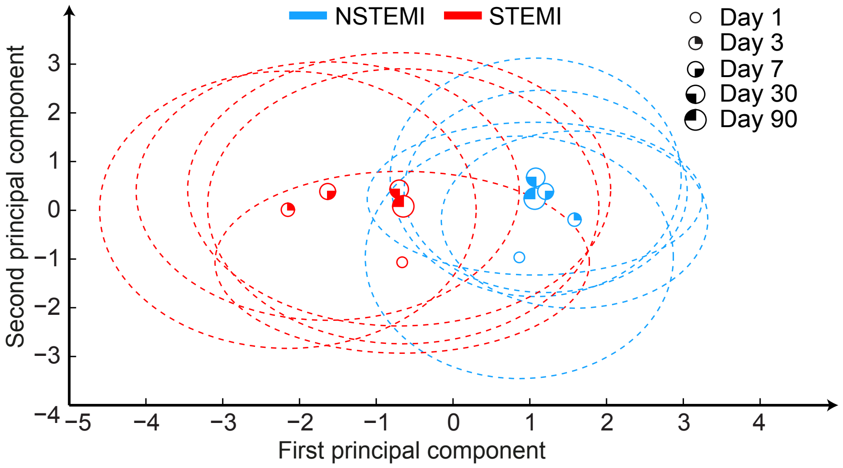

Figure 4.14: Principal component analysis of the combined expression data of $\mathcal{S}_f^*$. Dotted lines represent one standard deviation around the group-specific gene expression averages projected onto the principal components space.

as noticed in the case of the MI vs. UA study, the genes identified by l_1-StaR don't exhibit the long differential expression observed in the genes selected by the multi-stage method. This can be observed by comparing the dynamic profiles of the genes selected by l_1-StaR shown in Figure D.10 of Appendix D against the dynamic profiles of the genes selected by the multi-stage method shown in Figure D.8 and Figure D.9. It appears the multi-stage method identifies many more genes that are differentially expressed over multiple time-points that the l_1-StaR algorithm. This feature renders the multi-stage method more appropriate for feature selection studies based on time-course microarray data.

4.7 Conclusions

This chapter proposed a new multi-stage feature selection method operating in a nested cross-validation fashion. The internal loop is used to select optimal model parameters while the external loop is used to select relevant features and provide an unbiased estimate of the their predictive performance.

In Stage I, a new unsupervised filter that takes advantage of the uncertainty around the gene expression measurements summarized with multi-mgMOS was used to select biological informative genes (i.e remove nosy and housekeeping genes). If the gene expression data are summarized using other methods, the filter can be replaced by any of the unsupervised filters presented in Chapter 2.

In Stage II, differentially expressed genes were selected using either the Welch's t-test (when the population distributions were assumed to be normal) or the

Wilcoxon's rank-sum test (when no distributional assumptions were made about the data). Note that any other statistical test that is marginally independent with the filter in Stage I can be used. To correct for the multiplicity problem, pFDR control was adopted.

In Stage III, genes highly correlated with the class variable and minimally correlated with each other were selected using mRMR with MIQ formulation given the superior classification performance over the MID formulation. However the MID formulation is more stable that the MIQ formulation (Gulgezen et al., 2009). If there is little overlap between the subsets selected in the external cross-validation loop, the MID formulation could be used for more consistent results.

In Stage IV, a SVM-based wrapper approach operating in a sequential forward selection fashion was used to select the minimal subset of genes associated with the highest discriminatory power.

Excluding the new unsupervised filter, the remaining methods have been previously used in feature selection studies. However, to the author's knowledge, this is the first time the methods have been combined within a multi-stage method that not only avoids the sources of bias in performance evaluation but outputs at each step subsets satisfying stage-specific measures of relevance which can be used for other downstream analyses.

The novel multi-stage feature selection method and the l_1-StaR algorithm were compared in two differential expression studies based on time-course microarray data from patients with ACS. The l_1-StaR algorithm selected, on average, more genes than the multi-stage method (12.9 compared to 6.4 in the first study; 4.9 compared to 4.5 in the second study) providing better diagnostic performance (99.3% compared to 95.4% in the first study; 100% compared to 99% in the second study). While both methods identify genes than can efficiently distinguish between the diagnostic groups of ACS three months after hospital admission, the genes selected by the multi-stage method show longer-term differential expression than the genes selected by l_1-StaR, suggesting that the method is more appropriate for biomarker discovery in time-course microarray studies.

Given the absence of pre-admission data, extensive biological research is needed to investigate the causes underlying the observed differences in the expression level of these genes. These differences could be attributed to:

- the response of the gene regulatory network to the ACS episode
- changes in the patient's physical activity and/or dietary habits
- differences in blood cell counts
- medication

Genes responding to the ACS episode could have powerful implications in terms of diagnosis. In the context where the temporal utility for diagnosis of the known cardiac markers is at most 14 days after onset of the symptoms, genes differentially expressed up to 90 days after hospital admission could become reliable biomarkers for late diagnosis. Additionally, if some of the genes were also differentially expressed before the ACS events, new hypothesis about the genetic predisposition to heart attacks could be formulated.

Chapter 5

A novel deconvolution method for microarray gene expression data

Deconvolution of heterogeneous gene expression data represents a powerful alternative to cell separation methods for identifying cell type-specific markers and studying the dynamics of cell populations. As discussed in Chapter 3, there are three major classes of deconvolution methods estimating cell type-specific proportions and/or cell type-specific expression levels from heterogeneous microarray data. Existing methods estimating cell type-specific expression levels don't measure the contribution of each cell type to the variance of the heterogeneous gene expression measurements. This information could reveal sources of interindividual variation in gene expression patterns and provide a deeper understanding of the biological system under study.

This chapter proposes in Section 5.1 a novel approach to estimate positive cell type-specific expression levels from heterogeneous microarray data that exploits the OFR approach (Billings et al., 1988). This method naturally quantifies the contribution of each cell type to the variance of the gene expression patterns. Section 5.2 discusses a method used in econometrics for comparing coefficients between two regression models. This method is proposed for cell type-specific differential expression analysis which is concerned with identifying the cell types contributing differently to the total measured expression level of a given gene in the case and control groups. In Section 5.3, the novel deconvolution approach together with the test for cell type-specific differential expression analysis are applied on the genes distinguishing MI from UA and NSTEMI from STEMI, given associated blood cell counts, to identify the cellular sources of differential expression and measure the increments to the proportion of explained gene expression variance associated with each cell type. Concluding remarks are given in Section 5.4.

5.1 An orthogonal forward regression approach for microarray data deconvolution

Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}_+^{n \times m}$ denote the matrix of raw gene expression measurements, where n represents the number of samples and m represents the number of genes. Also, let $\mathbf{F} = [f_{ik}] \in \mathbb{R}_+^{n \times K}$ denote the matrix of cell type fractions satisfying (3.2), where K represents the number of cell types in the mixture. Deconvolution of the cell type-specific expression signatures g_{kj}, contributing to the heterogeneous measurements of gene j, consists of solving the system of linear equations (3.4). This system can be represented in the following matrix form:

$$\mathbf{x}_j = \mathbf{F}\mathbf{g}_j \tag{5.1}$$

where $\mathbf{x}_j$ represents the jth column of $\mathbf{X}$ and $\mathbf{g}_j = \left(g_{1j}, g_{2j}, \ldots, g_{Kj}\right)^\top$. The system (5.1) assumes that all the cell types in the mixture contribute to $\mathbf{x}_j$. In practice, the number and identity of these cell types are rarely known apriori. Additionally, correlation between the columns $\mathbf{f}_k$ of $\mathbf{F}$ mask how each cell type contribute to the variance of $\mathbf{x}_j$. To identify the cellular sources of gene expression and quantify their contribution to the variance of the heterogeneous gene expression patterns, an novel approach based on OFR is proposed. Before presenting the OFR framework and the particularities of the new approach, the orthogonal least squares (OLS) solution of (5.1) (which lies at the heart of the OFR method) is discussed.

Solving the deconvolution system (5.1) using OLS consists of decomposing the regression matrix into:

$$\mathbf{F} = \mathbf{W}\mathbf{A} \tag{5.2}$$

where the triangular matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$ takes the form:

$$\mathbf{A} = \begin{pmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1k} \\ 0 & 1 & a_{23} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & a_{k-1\,k} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \tag{5.3}$$

while the matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$ with orthogonal columns $\mathbf{w}_i$ satisfies:

$$\mathbf{W}^\top \mathbf{W} = \mathbf{H} \tag{5.4}$$

where the entries of the diagonal matrix **H** are defined by:

$$h_i = \mathbf{w}_i^\top \mathbf{w}_i, i = 1 \dots k \tag{5.5}$$

The factorization (5.2) can be obtained using classical Gram-Schmidt method (Chen et al., 1989) which computes one column of **A** at a time and orthogonalizes **F** as described in the following computational procedure:

- Set $\mathbf{w}_1 = \mathbf{f}_1$
- At the ith stage $(i = 2 \dots K)$, compute the ith column of **A**

$$a_{li} = \frac{\mathbf{w}_l^\top \mathbf{f}_i}{\mathbf{w}_l^\top \mathbf{w}_l}, 1 \leq l < i \tag{5.6}$$

- And make $\mathbf{w}_i$ orthogonal on the previously $i-1$ orthogonalized columns

$$\mathbf{w}_i = \mathbf{f}_i - \sum_{l=1}^{i-1} a_{li} \mathbf{w}_l \tag{5.7}$$

Using the factorization (5.2), the deconvolution system (5.1) can be written as:

$$\mathbf{x}_j = \mathbf{W}\mathbf{z}_j \tag{5.8}$$

where

$$\mathbf{z}_j = \mathbf{A}\mathbf{g}_j \tag{5.9}$$

The OLS solution $\hat{\mathbf{z}}_j$ in the space spanned by $\mathbf{w}_i$ is given by (Chen et al., 1991):

$$\hat{\mathbf{z}}_j = \mathbf{H}^{-1}\mathbf{W}^\top \mathbf{x}_j \tag{5.10}$$

or equivalently:

$$\hat{z}_{ij} = \frac{\mathbf{w}_i^\top \mathbf{x}_j}{\mathbf{w}_i^\top \mathbf{w}_i}, 1 \leq i \leq K \tag{5.11}$$

The solution $\hat{\mathbf{g}}_j$ in the space spanned by the regressors $\mathbf{f}_i$ can be recovered by solving (5.9) using backward elimination. Letting $\boldsymbol{\epsilon}_j = (\epsilon_{1j}, \dots, \epsilon_{nj})^\top$ denote the vector of residuals (fitting errors) and taking into account the orthogonality of the vectors $\mathbf{w}_i$, the variance of $\mathbf{x}_j$ can be expressed as:

$$\frac{\mathbf{x}_j^\top \mathbf{x}_j}{n} = \frac{1}{n} \sum_{i=1}^{k} z_{ij}^2 \mathbf{w}_i^\top \mathbf{w}_i + \frac{\boldsymbol{\epsilon}_j^\top \boldsymbol{\epsilon}_j}{n} \tag{5.12}$$

The first term on the right side of equation (5.12) represents the variance of $\mathbf{x}_j$

explained by the cell types while the second term represents the unexplained variance. Dividing in (5.12) by the variance of $\mathbf{x}_j$ we get:

$$1 = \frac{\sum_{i=1}^{k} z_{ij}^2 \mathbf{w}_i^\top \mathbf{w}_i}{\mathbf{x}_j^\top \mathbf{x}_j} + \frac{\boldsymbol{\epsilon}_j^\top \boldsymbol{\epsilon}_j}{\mathbf{x}_j^\top \mathbf{x}_j} \tag{5.13}$$

The quantity:

$$R_j^2 = \frac{\sum_{i=1}^{k} z_{ij}^2 \mathbf{w}_i^\top \mathbf{w}_i}{\mathbf{x}_j^\top \mathbf{x}_j} \tag{5.14}$$

is known in econometrics as the non-centered coefficient of determination (COD) (not to be mistaken with the COD discussed in Section 3.2.2 of Chapter 3) and represents the proportion of output variance explained by the regressors. The R_j^2 can be interpreted as measure of the goodness of fit of a model. A value of R_j^2 equal to one indicates that the model perfectly fits the data whereas a value of zero indicates that the model can't capture any variation in $\mathbf{x}_j$.

The increment to the proportion of explained variance associated with the ith cell type is given by:

$$\varepsilon_i = z_{ij}^2 \frac{(\mathbf{w}_i)^\top \mathbf{w}_i}{\mathbf{x}_j^\top \mathbf{x}_j} \tag{5.15}$$

These increments can be used to select a minimal subset of cell types explaining a desired proportion of explained variance, as described in the following OFR procedure:

- At the first step, for each $1 \leq r \leq k$ compute:

  1. $\mathbf{w}_1^{(r)} = \mathbf{f}_r$
  2. $z_{ij}^{(r)} = \frac{(\mathbf{w}_1^{(r)})^\top \mathbf{x}_j}{(\mathbf{w}_1^{(r)})^\top \mathbf{w}_1^{(r)}}$
  3. $\varepsilon_1^{(r)} = \left(z_{1j}^{(r)}\right)^2 \frac{(\mathbf{w}_1^{(r)})^\top \mathbf{w}_1^{(r)}}{\mathbf{x}_j^\top \mathbf{x}_j}$

  Set $\mathbf{w}_1 = \mathbf{f}_{r_1}$ where r_1 is the index of max $\left\{\varepsilon_1^{(r)} | r = 1 \dots k\right\}$

- At the ith step $(1 \leq i \leq K)$ and for $1 \leq r \leq k, r \neq r_1, \dots r_{i-1}$, compute:

  1. $a_{li}^{(r)} = \frac{\mathbf{w}_l^\top \mathbf{f}_r}{\mathbf{w}_l^\top \mathbf{w}_l}, 1 \leq l < i$
  2. $\mathbf{w}_i^{(r)} = \mathbf{f}_r - \sum_{l=1}^{i-1} a_{li}^{(r)} \mathbf{w}_l$
  3. $z_{ij}^{(r)} = \frac{(\mathbf{w}_i^{(r)})^\top \mathbf{x}_j}{(\mathbf{w}_i^{(r)})^\top \mathbf{w}_i^{(r)}}$

4. $\varepsilon_i^{(r)} = \left(z_{ij}^{(r)}\right)^2 \frac{(\mathbf{w}_1^{(r)})^\top \mathbf{w}_1^{(p)}}{\mathbf{x}_j^\top \mathbf{x}_j}$

   Set $\mathbf{w}_i = \mathbf{w}_i^{(r_i)}$ where r_i is the index of max $\left\{\varepsilon_i^{(r)} | r = 1 \ldots k, r \neq r_1, \ldots r_{i-1}\right\}$

- Terminate the procedure if $R_j^2 \geq 1 - \varrho$, where $0 \leq \varrho \leq 1$ represents the desired tolerance.

Note that for $\varrho = 0$, the OFR procedure selects all cell types and sorts them in descending order of their increment to the proportion of explained variance. Solving the system (5.9) using backward elimination yields the estimates $\hat{g}_{kj}$, which can take negative as well as positive values. The negative estimates carry no biological interpretation and confound the evaluation of the cell type-specific contributions to the variance of $\mathbf{x}_j$. To estimate positive $\hat{g}_{kj}$, in this work the system is solved using a non-negative least squares algorithm (Lawson and Hanson, 1974). In what follows, the OFR procedure using backward elimination and the procedure using non-negative least squares will be referred to as the unconstrained and constrained OFR, respectively. The two approaches produce identical results when the solution of (5.9) contains only positive entries. In this case, the association between the cell types and their increment to the explained variance is preserved.

This property doesn't hold when the unconstrained solution $\hat{\mathbf{g}}_j$ contains negative entries as during non-negative optimization, regressors with non-zero contribution to the explained variance can be associated with zero estimates of the cell type-specific expression signatures. This in turn changes the increments to the observed variance of the next regressors in the ranking. To correct these estimates, the constrained OFR deconvolution is repeated for the $\mathbf{f}_k$ associated with the non-zero parameters resulted from non-negative least squares optimization. If this step produces additional zero coefficients for some regressors, the constrained OFR algorithm is successively applied until only cell types with non-zero gene expression signature are selected. Note that since the cell counts and the gene expression measurements are positive, there is at least one regressor associated with a positive coefficient.

5.2 Testing for cell type-specific differential expression

Differential gene expression between case and control groups can be attributed to differences: (i) in cell type proportions or (ii) in the contribution of each cell type to the measured gene expression levels. Considering that the number of cell types in the mixture rarely exceeds the order of tens, testing for differences in cell type proportions can be approached using single hypothesis testing procedures.

To test for cell type-specific differential expression, Shen-Orr et al. (2010) proposed the test statistics :

$$T_{kj} = \hat{g}_{kj}^{(1)} - \hat{g}_{kj}^{(2)},\ k = 1 \ldots K \tag{5.16}$$

where $\hat{g}_{kj}^{(1)}$ and $\hat{g}_{kj}^{(2)}$ represent the cell type-specific expression levels estimated for the case and control group, respectively. The null distributions of the test statistics were estimated by permuting the class labels of the arrays and the FDR was calculated for each cell type as the ratio of the genes whose statistic exceed a given threshold in the real data to the average number of genes exceeding the same threshold in the permuted data. This approach doesn't expose which cell types contribute differently to the measured expression level of a given gene but provides an estimate of the number of gene differentially expressed in each cell type at a given significance threshold. Additionally, this approach is computationally expensive for large number of permutations. To reduce the computational burden and directly measure the significance of the difference in cell type-specific gene expression signatures between case and control deconvolution models, the following approach used in econometrics was adopted.

Let $\mathbf{x}_j^{(1)} \in \mathbb{R}_+^{n_1 \times 1}$ and $\mathbf{x}_j^{(2)} \in \mathbb{R}_+^{n_2 \times 1}$ denote the vectors of expression measurements for gene j in the case and control groups, respectively, where n_1 and n_2 represent the number of samples in each group. Also, let $\mathcal{F}^{\cup} = \mathcal{F}^{(1)} \cup \mathcal{F}^{(2)}$, $(\#(\mathcal{F}^{\cup}) = l^{\cup} \leq K)$, and $\mathcal{F}^{\cap} = \mathcal{F}^{(1)} \cap \mathcal{F}^{(2)}$, $(\#(\mathcal{F}^{\cap}) = l^{\cap} \leq K)$, denote the union and the intersection of the subsets of cell types selected by the OFR procedure for the case and control groups, respectively. Define the regression matrix for the case group $\mathbf{F}^{(1)} = [\mathbf{f}_i^{(1)}] \in \mathbb{R}_+^{n_1 \times l^{\cup}}$, with column order dictated by the order of $\mathcal{F}^{\cup}$, consisting of the counts for the cell types in $\mathcal{F}^{(1)}$ and zero entries otherwise. The regression matrix for the control group $\mathbf{F}^{(2)} = [\mathbf{f}_i^{(2)}] \in \mathbb{R}_+^{n_2 \times l^{\cup}}$ is analogously defined. The joint deconvolution model for the $n = n_1 + n_2$ expression measurements of gene j takes the form:

$$\check{\mathbf{x}}_j = \check{\mathbf{F}} \check{\mathbf{g}}_j \tag{5.17}$$

where $\check{\mathbf{x}} = [\mathbf{x}_j^{(1)}; \mathbf{x}_j^{(2)}] \in \mathbb{R}_+^{n \times 1}$, $\check{\mathbf{F}} = [\mathbf{F}^{(1)}; \mathbf{F}^{(2)}] \in \mathbb{R}_+^{n \times l^{\cup}}$ and $\check{\mathbf{g}}_j = [\mathbf{g}_j^{(1)}; \mathbf{g}_j^{(2)}] \in \mathbb{R}_+^{n \times 1}$. Note that solving (5.17) using the OLS algorithm is equivalent to simultaneously estimating the cell type-specific expression signatures for the case and control groups.

Let $\mathcal{C} = [\mathbf{1}^{n_1 \times 1}; \mathbf{0}^{n_2 \times 2}]$ denote the class vector consisting of entries of 1 for the samples in the case group and entries of 0 for the samples in the control group. To study the difference in cell type-specific expression signatures between the case and control groups, the model (5.17) is supplemented with $l^{\cap}$ interaction terms, one for each cell type in $\mathcal{F}^{\cap}$, as follows:

$$\check{\mathbf{x}}_j = \check{\mathbf{F}}\check{\mathbf{g}}_j + \beta_1(\mathcal{C} * \check{\mathbf{f}}_{k_1}) + \cdots + \beta_{l^{\cap}}(\mathcal{C} * \check{\mathbf{f}}_{k_{l^{\cap}}}) \tag{5.18}$$

where $\check{\mathbf{f}}_{k_i}$ represents the column of $\check{\mathbf{F}}$ associated with the ith cell type in $\mathcal{F}^{\cap}$ and the interaction parameters $\beta_i, i = 1 \ldots l^{\cap}$, measure the difference in contribution to the heterogeneous gene expression measurements for the common cell types. Note that solving (5.18) using OLS is equivalent to fitting the model:

$$\mathbf{x}_j^{(1)} = \mathbf{F}^{(1)}\mathbf{g}_j^{(1)} + \beta_1\mathbf{f}_{k_1}^{(1)} + \cdots + \beta_{l^{\cap}}\mathbf{f}_{k_{l^{\cap}}}^{(1)} \tag{5.19}$$

to the data of the case group, and the model:

$$\mathbf{x}_j^{(2)} = \mathbf{F}^{(2)}\mathbf{g}_j^{(2)} \tag{5.20}$$

to the data of the control group. The system (5.18) can be expressed in a more compact form as follows:

$$\check{\mathbf{x}}_j = \boldsymbol{\Phi}\boldsymbol{\theta}_j \tag{5.21}$$

where $\boldsymbol{\Phi} = [\check{\mathbf{F}} \;\; \mathcal{C} * \check{\mathbf{f}}_{k_1} \ldots \mathcal{C} * \check{\mathbf{f}}_{k_{l^{\cap}}}]$, $\boldsymbol{\theta} = [\mathbf{g}_j^{(1)}; \mathbf{g}_j^{(2)}; \boldsymbol{\beta}]$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{l^{\cap}})^{\top}$.

After estimating $\boldsymbol{\theta}_j$ in (5.21), conducting cell type-specific differential expression analysis reduces to testing that the interactions $\hat{\beta}_i$ and the parameters $\hat{g}_{kj}^{(1)}$ and $\hat{g}_{kj}^{(2)}$ associated with the cell types in $\mathcal{F}^{(1)} \setminus \mathcal{F}^{(2)}$ and $\mathcal{F}^{(2)} \setminus \mathcal{F}^{(1)}$, respectively, are significantly different from zero. This can be accomplished using the Wald test (Wald, 1943) described below.

Let $\hat{\theta}$ denote the estimated value of the parameter θ and let θ_0 denote the value proposed for comparison. The Wald test works by testing the null hypothesis $H_0 : \hat{\theta} = \theta_0$ against the alternative hypothesis $H_1 : \hat{\theta} \neq \theta_0$ using the statistic:

$$T_0 = \frac{\hat{\theta} - \theta_0}{\sigma(\hat{\theta})} \tag{5.22}$$

where $\sigma(\hat{\theta})$ represents the standard error of $\hat{\theta}$. In the context of the deconvolution model (5.21), the squared standard errors of the parameters in $\hat{\boldsymbol{\theta}}_j$ represents the diagonal entries of the covariance matrix:

$$cov(\hat{\boldsymbol{\theta}}_j) = \sigma_j(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1} \tag{5.23}$$

where

$$\sigma_j = \frac{\boldsymbol{\epsilon}_j^{\top}\boldsymbol{\epsilon}_j}{n - (l^{\cap} + l^{\cup})} \tag{5.24}$$

denotes the variance of the residuals $\boldsymbol{\epsilon}_j = \check{\mathbf{x}}_j - \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_j$. Parameters whose associated p-values are less than or equal to the significance threshold α correspond to cell types contributing differently to the total expression level of gene j. The magnitude and the directionality of the contrast is given by the absolute value and sign of the difference between the values of the same parameter in the case and control groups. Note that in the case when deconvolution is performed for multiple genes, the pFDR can be estimated for each cell type from the p-values associated with each gene, using the algorithm presented in Section 4.2.3 of Chapter 4. Thus the approach described in this section can be used to study differential gene expression for each cell type in the mixture as well as estimate the proportion of genes differentially expressed in each cell type when multiple deconvolution problems are solved.

5.3 Expression deconvolution of the genes differentiating between the ACS subtypes

The constrained and unconstrained OFR methods together with the method for cell type-specific differential expression were applied on the complete blood count (CBC) dataset presented in Section 5.3.1, to identify the cell types expressing the genes distinguishing between the diagnostic groups of ACS, to measure the contribution of each cell type to the variability of these genes and to identify the cellular sources of differential gene expression.

To this end, two studies were conducted. The first study, presented in section 5.3.2, focuses on the differentially expressed genes between MI and UA selected using the multi-stage method, while the second study, presented in Section 5.3.3, focuses on the genes distinguishing NSTEMI from STEMI. These studies indicate group-specific features of variation in gene expression patterns and identify cell type-specific differentially expressed genes.

5.3.1 The CBC dataset

A CBC measures the concentration of white blood cells (basophils, eosinophils, lymphocytes, monocytes and neutrophils), red blood cells (erythrocytes) and platelets in the blood. A total of 136 CBC tests were performed on peripheral blood samples collected from the 33 patients with ACS at day 1, day 3, day 7, day 30 and day 90 after hospital admission. Out of the 136 CBC tests, 131 correspond to patients and visits for which gene expression data was collected (see Section 4.6 of Chapter 4 for a description of the microarray dataset). Table 5.1 lists the number of arrays associated with CBC data for each visit and diagnostic group. The average count (median) for each cell type in the diagnostic groups of ACS is listed in Table 5.2. These quantities were computed using data from all the visits.

Table 5.1: Amount of arrays associated with CBC data

Visit	NSTEMI	STEMI	UA	Total
Day 1	10	8	10	28
Day 3	8	5	5	18
Day 7	12	7	8	27
Day 30	14	8	10	32
Day 90	12	7	7	26
Total	56	35	40	131

Table 5.2: Average count for each blood cell in the diagnostic groups of ACS

Cell type	NSTEMI	STEMI	UA
Basophils	0	0	0
Eosinophils	0.1	0.200	0.2
Lymphocytes	1.7	1.9	1.7
Monocytes	0.5	0.7	0.4
Neutrophils	4.45	5.6	4.3
Red blood cells	4515	4360	4465
Plateles	245.5	260	233.5

1 unit = 10^9cells/liter of blood

5.3.2 Expression deconvolution of the genes differentiating MI from UA

The results of this section are organized as follows. The first section deals with the correlation between cell type counts within each group, deconvolution of cell type expression levels and quantification of the cell type-specific contributions to the variance of the differentially expressed genes. The second section is concerned with testing for differences in cellular proportions between the MI and UA groups and identifying cell type-specific differentially expressed gene.

Basophils were removed from the study due to their low count, recorded for most of the arrays as zero. Although mature red blood cells lack a nucleus, these were included in the study following evidence of transcriptional and translational activity in human red blood cells (Kabanova et al., 2009).

This study was conducted on the raw microarray data and blood cell counts from all the visits combined due to the large sample size of both MI and UA groups. Results from visits-specific analyses were invoked to justify the aggregation of the NSTEMI and STEMI groups and to compare the distributions of the coefficients of determinations for the models fitted using data from all the visits combined with the distributions arising from models fitted using data from each visit.

The study can be repeated for each visit independently, excluding day 3 due to the number of arrays in the UA group falling below the number of cell types in the mixture. Note that in these cases, caution must be taken when interpreting the biological relevance of the findings due to the small sample size of both MI and UA groups associated with each visit (comparable at times to the number of parameters to be estimated).

Deconvolution of cell type-specific expression levels

To investigate if the NSTEMI and STEMI groups can be taken as an aggregated group for deconvolution, differences in cell type composition and in the expression level of the genes $X_1 - X_{36}$ (listed in Table B.3) between the two groups were tested at significance level $\alpha = 0.05$ using the Wilcoxon's rank-sum test for each time point independently. Significant differences in cellular compositions were identified at day 1 for monocytes and neutrophil, at day 3 for eosinophils and lymphocytes and at day 7 for platelets. Significant differences in gene expression levels are shown in Table 5.3.

Table 5.3: Gene differentially expressed between NSTEMI and STEMI

Visit	Genes
Day 1	X_9, X_{14}, X_{19}, X_{23}, X_{26}
Day 3	X_{19}
Day 7	X_9,X_{23}
Day 30	X_{19}
Day 90	X_9

Given the absence of consistent differences in cellular compositions and in gene expression levels (excluding X_9 and X_{19}) across visits, the NSTEMI and STEMI cohorts were aggregated (MI group).

As discussed previously, correlation between blood cell counts masks how each cell type contributes to the variance of the gene expression measurements. To measure the pairwise correlation between cell type counts in the MI and UA groups, Pearson correlation coefficient were computed using data from the corresponding 91 and 40 CBC tests, respectively. Significant correlations at level $\alpha = 0.05$ for the MI group and UA group are shown in Table 5.4.

Table 5.4: Correlation between blood cell counts in the MI and UA groups

Group	Correlated cell types		Correlation coefficient
MI	Lymphocytes	Eosinophils	0.39
	Lymphocytes	Monocytes	0.52
	Lymphocytes	Neutrophils	0.29
	Neutrophils	Monocytes	0.57
	Neutrophils	Platelets	0.25
UA	Red blood cells	Eosinophils	0.34
	Red blood cells	Lymphocytes	0.33
	Red Blood cells	Platelets	0.39
	Neutrophils	Platelets	0.35
	Neutrophils	Monocytes	0.49

Next, the cell type-specific expression profiles of the genes $X_1 - X_{36}$ in the MI and UA groups were deconvolved using the constrained and unconstrained OFR methods, with the tolerance ϱ set to zero for both algorithms. For each gene, the deconvolution was performed using the expression measurements and the cell counts aggregated from all the visits and for each visit independently, except at day 3 for the UA group due to insufficient data (number of measurements less than the number of cellular types). Figure 5.1 shows the distributions of the CODs for the deconvolution models fitted using the two OFR methods on the data from each visit as well as on the data from all the visits combined.

Note in Figure 5.1(a) that the distributions associated with each visit are similar to the distributions obtained using data from all the visits combined. These findings suggest that the variability of the coefficients of determination associated with models fitted on data from all the visits doesn't originate primarily from grouping sampled from different time points but from the larger variability of some genes at specific time points which can't be captured by visit-specific models.

The overall medians of the CODs summarized in Figure 5.1 (a) and 5.1 (b) are 0.89 and 0.93, respectively. Although the unconstrained OFR produces a slight increase in the overall goodness of fit at the expenses of allowing negative estimated of the cell type-specific expression signatures, the distributions of the CODs remain heavily skewed towards zero (Figure 5.1 (b)). These findings suggest that the large variability observed in the performance of the constrained OFR approach doesn't represent an artefact of the new algorithm but rather of the unaccounted sources of interindividual variability specific to blood-based gene expression measurements. These sources are of biological origin (e.g. age, gender or time of the day at which the blood samples were collected) (Whitney et al., 2003), and technical origin (e.g. RNA isolation method) (Min et al., 2010).

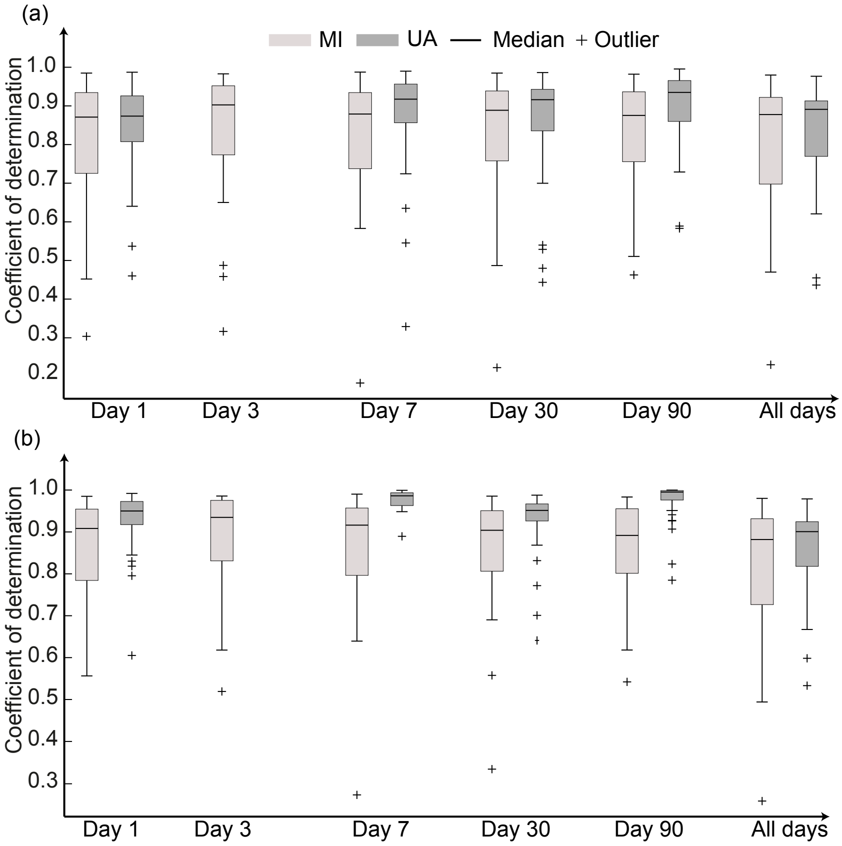

Figure 5.1: Distributions of the CODs across visits for the models fitted on data from the MI and UA groups using (a) the constrained OFR and (b) the unconstrained OFR. The lower and upper edge of the box represent the 25th and 75th percentiles, respectively, while the whiskers extend to the most extreme points not considered outliers.

Figure 5.2 shows the cell types contributing to the expression level of the genes X_1, X_3 and X_6 (exhibiting differential expression across all the visits) together with their increments to the explained variance. Note that in the case of the most relevant gene (X_1), more than half of the variance in the data is left unexplained, suggesting high variability in the gene expression measurements induced by additional sources, other than cellular composition. Moreover, the cell subsets expressing this gene are different in the MI and UA groups. This observation holds for gene X_3 whose variability is largely captured by the deconvolution model. In the case of X_6, most of the variability is attributed to platelets in both groups.

The deconvolution results for the remaining genes, presented in Table C.1, Table C.2, Table C.3, and Table C.4 of Appendix C, expose the cell type sources of gene expression and reveal features of within and between groups variation in expression measurements. Table C.7 shows the number of genes expressed in each cell type within each group. In the MI group the variability of the gene expression measurements is predominantly explained by erythrocytes and platelets. In the UA group, lymphocytes and neutrophils also capture large proportions of variance for some genes, alongside with erythrocytes and platelets.

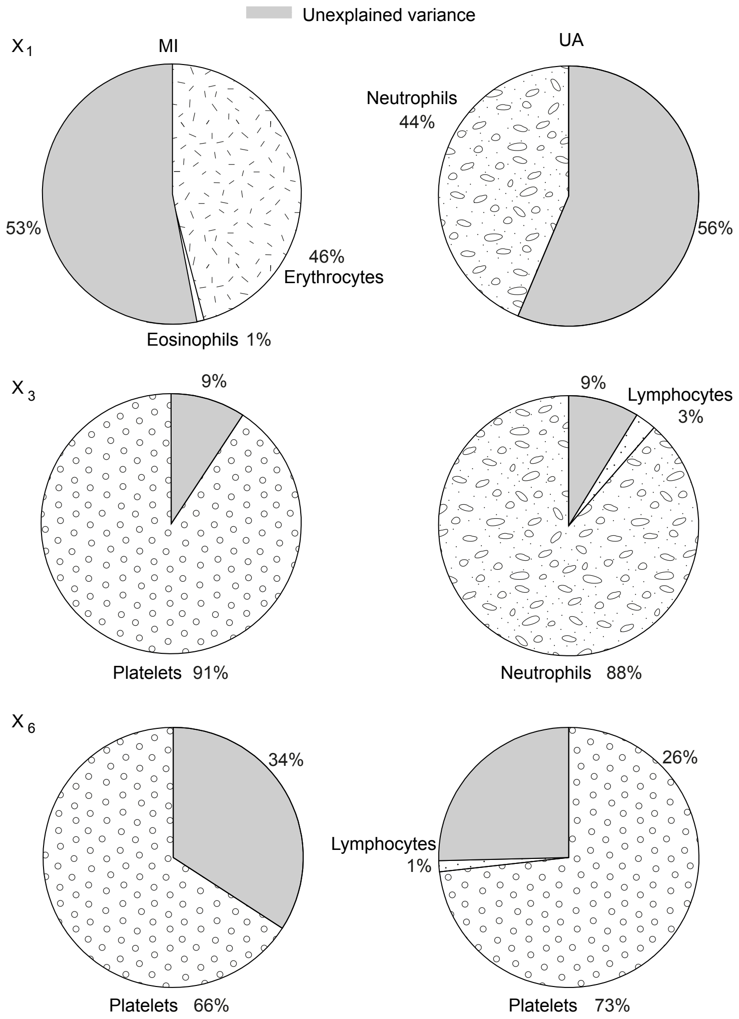

Figure 5.2: Cell type-specific contributions to the expression level of genes X_1, X_3 and X_6 in the MI and UA groups.

Cell type-specific differential expression analysis

Differences in cell type proportions between case and control groups represent a major confounding factor for the accurate interpretation of the results arising from differential gene expression studies. Testing for differences in cell type proportions between the MI and UA was performed using the two-sided Wilcoxon's rank sum test at significance level $\alpha = 0.05$ for each visit independently and for all the visits combined. Significant differences were identified at day 1 for monocytes, at day 7 for neutrophils and platelets and across all visits between monocytes and neutrophils. Removing the monocytes counts from day 1 and testing again for difference across the remaining visits showed that the contrast in monocytes counts didn't originate from the measurement taken at day one. On the other hand, the observed differences in platelets counts across all visits originated from the proportion measured at day 7. The distributions of the cell type counts for each visit are shown in Figure 5.3.

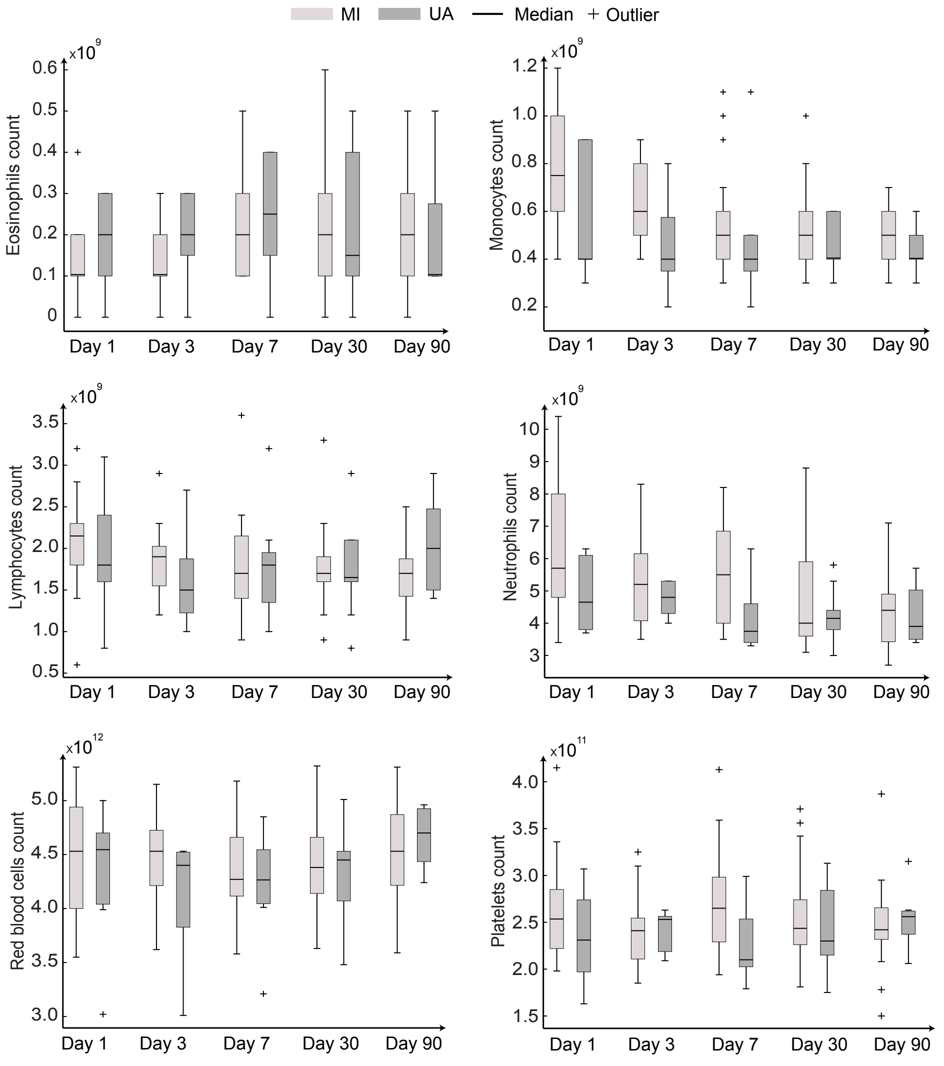

Figure 5.3: Distributions of the blood cell counts in the MI and UA groups

Given the absence of consistent differences in cell type proportions (excluding monocytes), observed differences in the expression of genes $X_1 - X_{36}$ may be attributed to cell type-specific contributions to the total transcript abundance. To identify the cellular sources of gene expression, cell type-specific differential expression analysis was carried out using data cross all visits following the approach described in section 5.2 for the significance threshold $\alpha = 0.05$. Table 5.5 list the genes differentially expressed in each cell type.

Table 5.5: Gene differentially expressed between NSTEMI and STEMI

Cell type	Genes	Total
Eosinophils	X_{28}, X_{32}, X_{35}	3
Monocytes	X_4, X_{23}, X_{24}, X_{34}	4
Neutrophils	X_1,X_3,X_{15},X_{16},X_{18}	5
Lymphocytes	X_3, X_6, X_9, X_{10}, X_{12}, X_{14}, X_{23} X_{27},X_{29}, X_{30}, X_{35}	11
Erythrocytes	X_7, X_9, X_{10}, X_{12}, X_{14}, X_{17}, X_{20} X_{22} , X_{23} , X_{24}, X_{26} , X_{34} , X_{35}	13
Platelets	X_3, X_{12}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19} X_{21}, X_{22}, X_{24}, X_{26}, X_{29}, X_{30}, X_{33}	14

Most of the genes are differentially expressed in lymphocytes, erythrocytes and platelets. The genes $X_2, X_5, X_8, X_{13}, X_{25}, X_{28}, X_{31}, X_{32}$ and X_{36} are not differentially expressed in any of the cell types. Out of these genes, X_2, X_{25} and X_{32} are expressed in monocytes in both groups. Therefore the observed difference in the expression level of these gene could be attributed to differences in the monocytes proportions between MI and UA. The source of differential expression for the remaining genes could be attributed to basophils whose counts were removed from the analysis due to the poor precision on the measurements. Basophils are known to be involved in inflammatory reactions and could play a role in coronary and myocardial diseases (Marone et al., 1989).

5.3.3 Expression deconvolution of the genes differentiating NSTEMI from STEMI

The results of this section are organized following the format and of the study presented in Section 5.3.2. This study was conducted on the raw microarray data and blood cell counts of the NSTEMI and STEMI groups taken from all the visits. Deconvolution results based on the arrays from each visit were briefly discussed when comparing the variability captured by the models fitted on the data from all the visits against the variability captured by models fitted using data from each visit independently. Cell type-specific differential expression analysis was not conducted for the models associated with each visit due to the small number of arrays in each group.

Deconvolution of cell type-specific expression levels

The first step of the analysis consisted of measuring the pairwise correlation between cell types in the NSTEMI and STEMI groups. Pearson correlation coefficient were computed using data from the corresponding 56 and 35 CBC tests, respectively. Significant correlations at level $\alpha = 0.05$ are shown in Table 5.6. These correlations mask the contribution of each cell type to the variance of the gene expression measurements. The OFR approach deals with correlation in a principled way and allows for reliable quantification of the increments to the variance explained associated with each relevant cell type.

Table 5.6: Pairwise correlations between blood cell in the NSTEMI and STEMI groups

Group	Correlated cell types		Correlation coefficient
NSTEMI	Lymphocytes	Eosinophils	0.40
	Lymphocytes	Monocytes	0.43
	Erythrocytes	Platelets	-0.26
	Neutrophils	Monocytes	0.48
	Neutrophils	Platelets	0.34
STEMI	Monocytes	Lymphocytes	0.56
	Neutrophils	Eosinophils	-0.34
	Neutrophils	Monocytes	0.58

Next, the cell type-specific expression profiles of the genes $X_1 - X_{21}$ in the NSTEMI and STEMI groups were deconvolved using the OFR method imposing positivity constraints as well as the unconstrained OFR, with the tolerance ϱ set to zero for both algorithms. The deconvolution was performed for each group using data from all the visits and for each visit independently, except at day 3 for the STEMI group due to insufficient data (number of measurements less than the number of cellular types). The distributions of the CODs for the fitted models are showed in Figure 5.4.

Figure 5.4(a) shows variability in the performance of the models fitted for each visit independently. Note that for some genes (outliers), large proportions of variance can't be explained by the deconvolution model. The performance of the models fitted on the data from all the visits combined is comparable to the performance of the visit-specific models. The overall medians of the CODs summarized in Figure 5.4 (a) and 5.4 (b) are 0.88 and 0.95, respectively. The models fitted using the unconstrained OFR explain better the observed variability in the gene expression patters at the cost of including negative estimates of the cell type-specific expression signatures. Considering that the linearity assumption of microarray convolution (and deconvolution) was shown to hold experimentally (Shen-Orr et al.,

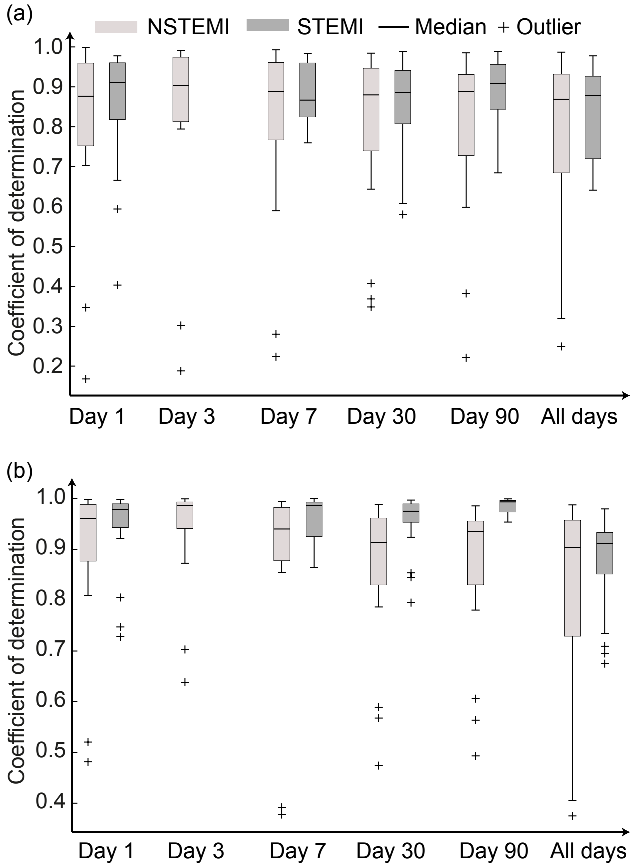

Figure 5.4: Distributions of the CODs across visits for the models fitted on data from NSTEMI and STEMI groups using (a) the constrained OFR method and (b) the unconstrained OFR method.

2010), the presence of genes whose variability couldn't be explained by both deconvolution approaches support the idea that there are other sources of variation (biological and/or technical) not included in the model and exclude the possibility that the fitting performance is an attribute of the constrained OFR method.

The deconvolution results for the first three most relevant genes are shown in Figure 5.5. The variability in the expression measurements of gene X_1 is largely explained by erythrocytes in both groups. Gene X_2 exhibits variability that can't be attributed to the considered cell counts alone. In the case of gene X_3 five cell types participate in explaining large proportions of variability in both groups. The results for the remaining genes are presented in Table C.5 and Table C.6 of Appendix C. Table C.8 shows the number of genes expressed in each cell type within each group. In the NSTEMI group the variability of the gene expression measurements is predominantly explained by erythrocytes and monocytes whereas in the STEMI group neutrophils participate alongside erythrocytes to explain large proportions of variance in the gene expression measurements.

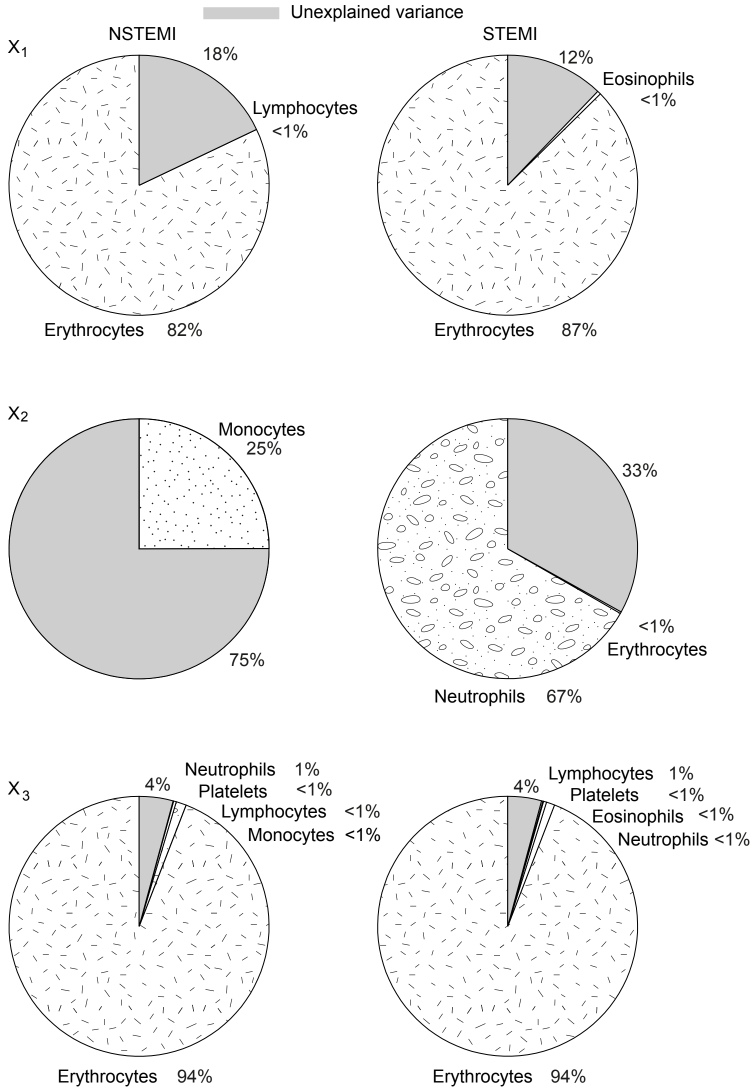

Figure 5.5: Cell type-specific contributions to the expression level of genes X_1, X_2 and X_3 in the NSTEMI and STEMI groups.

Cell type-specific differential expression analysis

Testing for differences in cell type proportions between the two groups was performed using the Wilcoxon's rank-sum test as significance level $\alpha = 0.05$. Significant differences in the data from all the visits combined were identified for eosinophils, monocytes, lymphocytes and neutrophils. These differences originate from the contrasts in eosinophils and lymphocytes at day one and in monocytes and neutrophils at day three, reported when discussing the aggregation of the

NSTEMI and STEMI groups in Section 5.3.2. The distributions of the cell type counts for each visit are shown in Figure 5.6.

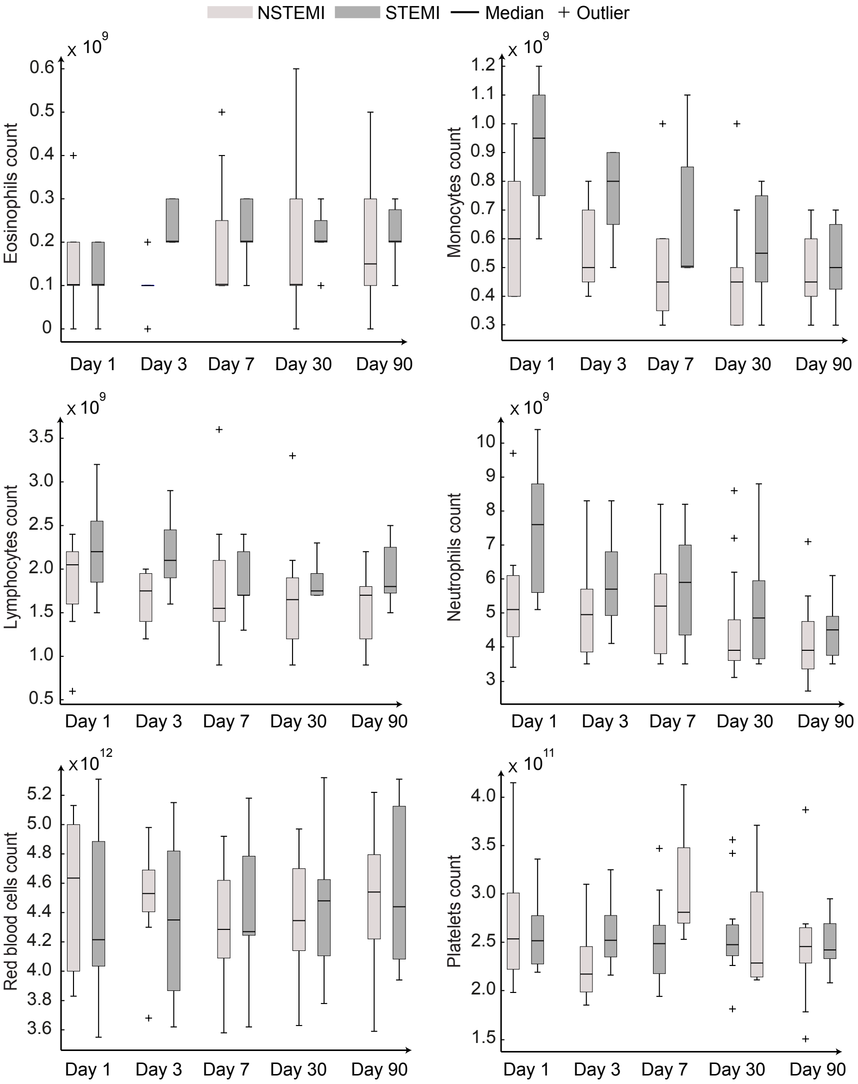

Figure 5.6: Distributions of the blood cell counts in the NSTEMI and STEMI groups

Given the absence of consistent differences in cell type proportions, observed differences in the expression of genes $X_1 - X_{21}$ may be attributed to cell type-specific contributions to the total transcript abundance. To identify the cellular sources of gene expression, cell type-specific differential expression analysis was carried out following the same approach and settings used in the first study. Table 5.7 list the genes differentially expressed in each cell type.

The genes X_3, X_9, X_{13} and X_{19} are not differentially expressed in any of the cell types. This could be attributed to the absence of basophils in the deconvolution

Table 5.7: Gene differentially expressed between NSTEMI and STEMI

Cell type	Genes	Total
Eosinophils	—	0
Lymphocytes	X_{14}	1
Monocytes	X_2, X_{17}, X_{21}	3
Neutrophils	X_5,X_8,X_{10},X_{21}	4
Platelets	X_4, X_6, X_{12}, X_{15}, X_{20}	5
Erythrocytes	X_1, X_2, X_6, X_7, X_8 X_{11} , X_{16}, X_{18}	8

model or to the small magnitude of differential expression associated with these genes (see Figure D.8 and Figure D.9 in Appendix D).

5.4 Conclusions

This chapter discussed a novel approach based on OFR for deconvolution of cell type-specific gene expression levels that naturally measured the contribution of each cellular type to the variance of the gene expression patters. This additional layer of information can reveal the sources of within and between groups variation in the heterogeneous expression measurements. To identify the cellular sources of differential gene expression, an approach measuring interaction effects was adopted. This approach is computationally more efficient than permutation based methods which require repeated runs of deconvolution with permuted data to test the significance of the contrast in the regression coefficients.

The deconvolution method was applied on the genes distinguishing between the subtypes of ACS. Results showed high deconvolution performance for the majority of genes using the constrained OFR approach. This performance was comparable to the deconvolution performance obtained in the absence of positivity constraints. The variability of genes associated to low COD could be attributed to the absence of basophils in the deconvolution model, to biological sources such as age, sex, time of the day the blood samples were collected or technical sources such as RNA isolation method.

The cell type-specific differential expression analysis identified cellular sources of (differential) expression for most of the genes distinguishing between the ACS groups. These results didn't account for the CODs of the deconvolution models. An in-depth search through the literature identified no methods that incorporate information about model fitting performance when comparing regression coefficients. Such an approach is essential for increasing the accuracy of the results or removing possible sources of bias associated with comparing coefficients of poor fitted models. Future research will investigate the relationship between the CODs and the precision of the results derived from cell type-specific differential expression analysis.

Chapter 6

A novel approach for modelling stable GRNs

A major challenge in computational biology is the reconstruction of large-scale stable GRNs. This challenge is strengthened by the shortage of data (number of temporal gene expression measurements less than the number of genes in the network). To reduce the space of candidate network topologies and avoid over-fitting the model parameters when data is scarce, strategies based on generating gene expression measurements using spline interpolation techniques and incorporation of biological constraints, were discussed in Chapter 3. As pointed out by Wu et al. (2004a), spline interpolation represents an *ad hoc* solution which may affect the biological interpretability of the model.

This chapter proposes in Section 6.1 a new approach to generate gene expression data for the unobserved time-points given unequally spaced time series microarray data. This approach accounts for the measurement noise and is congruent with the mathematical form of the response of stable dynamical system. A novel method to reconstruct large-scale stable GRNs modelled using the system of linear differential equations (3.18) is presented in Section 6.2. This method formulates parameter estimation as a nonlinear optimization problem to avoid the need for taking derivatives. Section 6.3 reviews an approach to obtain sparse topological representations for the estimated GRNs. An application to reconstruct the regulatory pathways between the differentially expressed genes selected by the multi-stage feature selection method in the studies conducted in Chapter 4 is presented in Section 6.4. Concluding remarks are given in Section 6.5.

6.1 Nonlinear approximation of gene expression dynamics by sums of exponentials

GRNs are nonlinear dynamical systems that can operate around multiple equilibrium points (Huang et al., 2005). In the neighbourhood of an equilibrium point, a nonlinear dynamical system can be approximated with a linear dynamical system (Khalil and Grizzle, 2002). Under certain assumptions, the solution of a linear dynamical system can be described using sums of exponentials. These assumptions are discussed in Section 6.1.1 where the general solution of a linear dynamical system is presented. This solution is used to derive a nonlinear model to approximate gene expression dynamics in the form of sums of exponentials. Section 6.1.2 presents a review of the methods to estimate the model parameters and discusses regularization to avoid over-fitting. Section 6.1.3 presents a strategy to select the regularization parameter that accounts for the measurement noise in the data.

6.1.1 The gene expression model

Let us consider the linear dynamical system:

$$\dot{\mathbf{x}}\left(t\right) = \mathbf{A}\left(\mathbf{x}(t) - \mathbf{x}_e\right) \tag{6.1}$$

where $\mathbf{x}\left(t\right) = \left(x_1\left(t\right), x_2\left(t\right), \dots, x_n\left(t\right)\right)^\top$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a stable matrix ($\sigma(A) \subset \mathbb{C}_H$) and $\mathbf{x}_e = \left(x_{e1}, x_{e2}, \dots, x_{en}\right)^\top$ represents the equilibrium point (steady-state). The solution of the system of linear differential equations (6.1) with initial condition $\mathbf{x}_0 = \mathbf{x}\left(t_0\right)$ is (Friedland, 2012):

$$\mathbf{x}\left(t\right) = e^{\mathbf{A}(t-t_0)}\left(\mathbf{x}_0 - \mathbf{x}_e\right) + \mathbf{x}_e,\ t \geq t_0 \tag{6.2}$$

Let $\lambda_1, \lambda_2, \dots, \lambda_k$ denote the k distinct eigenvalues of $\mathbf{A}$ and $m_i = \dim\ker\left(\mathbf{A} - \lambda_i\mathbf{I}_n\right)$ denote the geometric multiplicity of λ_i. The transition matrix $e^{\mathbf{A}(t-t_0)}$ admits the following Jordan decomposition (Teschl, 2012):

$$e^{\mathbf{A}(t-t_0)} = \mathbf{V}e^{\mathbf{J}(t-t_0)}\mathbf{V}^{-1} \tag{6.3}$$

where the columns $\mathbf{v}_i$ of $\mathbf{V}$ are the generalized eigenvectors of $\mathbf{A}$ and $e^{\mathbf{J}t}$ is the block diagonal matrix:

$$e^{\mathbf{J}t} = \begin{pmatrix} e^{\mathbf{J}_1 t} & & & \\ & e^{\mathbf{J}_2 t} & & \\ & & \ddots & \\ & & & e^{\mathbf{J}_k t} \end{pmatrix} \tag{6.4}$$

where each Jordan block $e^{\mathbf{J}_i t}$ is a square matrix of the form:

$$e^{\mathbf{J}_i t} = e^{\lambda_i t} \begin{pmatrix} 1 & \frac{t}{1!} & \frac{t^2}{2!} & \cdots & \frac{t^{m_i-1}}{(n-1)!} \\ 0 & 1 & \frac{t}{1!} & \cdots & \frac{t^{m_i-2}}{(n-2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{6.5}$$

It follows from (6.3) and (6.2) that the trajectories of the dynamical system (6.1) can be expressed using sums of complex exponentials multiplied by polynomials. Consider now the particular case when $\mathbf{A}$ has n distinct eigenvalues which implies $m_i = 1, \forall i$. The Jordan matrix decomposition (6.3) reduces to the spectral decomposition:

$$e^{\mathbf{A}(t-t_0)} = \mathbf{V} e^{\mathbf{\Delta}(t-t_0)} \mathbf{V}^{-1} \tag{6.6}$$

where $\mathbf{\Delta}$ is a diagonal matrix whose non-zero entries are the eigenvalues λ_i. Letting $\mathbf{U} = \left[\mathbf{u}_1^\top, \mathbf{u}_2^\top, \ldots, \mathbf{u}_n^\top\right]^\top$ denote the inverse of $\mathbf{V}$, equation (6.6) takes the form:

$$e^{\mathbf{A}(t-t_0)} = \sum_{i=1}^{n} \mathbf{v}_i \mathbf{u}_i e^{\lambda_i t} \tag{6.7}$$

Replacing equation (6.7) in (6.2), it follows that the solution of the dynamical system (6.1) can be expressed as:

$$\mathbf{x}(t) = \sum_{i=1}^{n} \mathbf{w}_i e^{\lambda_i t} + \mathbf{x}_e \tag{6.8}$$

where $\mathbf{w}_i = \langle \mathbf{u}_i, \mathbf{x}_0 - \mathbf{x_e} \rangle \mathbf{v}_i$ also represents an eigenvector of $\mathbf{A}$ associated with λ_i. Equation (6.8) shows that the individual trajectories $x_i(t)$ can be expressed using weighted sum of exponentials:

$$x_i(t) = w_{i0} + \sum_{j=1}^{n} w_{ij} e^{\lambda_j t}, \forall i \tag{6.9}$$

where $w_{i0} = x_{ei}$ and w_{ij} are the entries of the modal matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w_2}, \ldots, \mathbf{w}_n]$. When the gene regulatory pathways are described by (6.1), the complexity of the model (6.9) can be reduced if knowledge of the sparsity pattern of $\mathbf{A}$ is available. However, in practice we rarely known the topological structure of the network, the number of genes in the network or even if the genes we measure belong to the same network. In the absence of this information, the dynamic behaviour of the

ith gene on the array is approximated with:

$$f_i\left(\mathbf{w}_i, \boldsymbol{\lambda}_i, t\right) = w_{i0} + \sum_{j=1}^{m} w_{ij} e^{\lambda_{ij} t} \tag{6.10}$$

where $\mathbf{w_i} = (w_{i0}, w_{i1}, \ldots, w_{im})$ and $\boldsymbol{\lambda_i} = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im})$. The parameter w_{i0} models the steady state level of gene expression while w_{ij} and λ_{ij} characterize the transient behaviour. The model complexity m is chosen such that the number of model parameters $(2 * m + 1)$ doesn't exceed the length of the time series.

6.1.2 Parameter estimation using non-linear least squares

The first method for estimating the parameters in (6.10) was proposed by Prony (1795). The method consists of estimating the non-linear parameters $\boldsymbol{\lambda_i}$ as roots of a polynomial and the linear parameters $\mathbf{w}_i$ using linear least squares optimization. Kahn et al. (1992) points out that Prony's method is inconsistent even when the number of observations approaches infinity while Lanczos and Teichmann (1957) remarks that convergence to the true parameters is not guaranteed even when good initial estimates for $\boldsymbol{\lambda_i}$ are available.

Golub and Pereyra (1973), Kaufman (1975) and Ruhe and Wedin (1980) used a separable least squares approach to estimate the model parameters. Kundu and Mitra (1998) noted that this approach is sensitive to the initial value of the parameters and proposed a non-iterative algorithm that provides reliable initial parameter estimates. The algorithm modifies Prony's method by extended order modelling and singular value decomposition. Osborne and Smyth (1995) proposed a modified Prony method that is relatively insensitive to initial values of the parameters. Their method formulates the optimization problem as a non-linear eigenvalue problem which is solved iteratively.

Wiscombe and Evans (1977) proposed an iterative algorithm for fitting positive exponential sums. Their method automatically selects the number of exponential terms and has guaranteed convergence to the best fit parameters. De Groen and De Moor (1987) proposed two approaches where the nonlinear parameters $\boldsymbol{\lambda_i}$ emerge as eigenvalues of two suitably defined matrices. For both algorithms, the linear parameters $\mathbf{w}_i$ can be derived from the eigenvectors of the matrices or using linear least squares.

Non-linear least squares optimization represents the most widely used approach for exponential-sum fitting (Wiscombe and Evans, 1977) which stands out as a very versatile platform allowing incorporation of both linear and non-linear constraints. In a comprehensive review of parameter estimation methods for fitting sums of exponentials, Holmström and Petersson (2002) note that non-linear least squares optimization represents a central part of many (exponential fitting) algorithms and conclude that this optimization approach should be used when the number of exponential terms in (6.10) is known in advance.

Although the previous methods vary in their theoretical framework and performance, they don't tackle the sensitivity of the exponential-sum fitting problem to small variations in the experimental data. As Wiscombe and Evans (1977) point out, the fitting of sums of exponentials is a classically ill-posed problem in the Hadamard sense, with small perturbations in the data resulting in large changes in the model parameters (Varah, 1985). A classical approach to ill-posed problems is Tikhonov regularization (Tikhonov, 1995). This approach was used in the work of Alvarez and Lara (2011) who estimated the parameters of positive exponential sums by solving a mixed integer non-linear programming problem.

In this work, non-linear least square optimization is used with Tikhonov regularization to estimate the model parameters from time series microarray data data. Specifically, given $\{x_i\,(t_k)\}_{k=1}^{N}$ (the time course data for gene *i*), $\mathbf{w}_i$ and $\boldsymbol{\lambda}_i$ are estimated by solving the regularized nonlinear optimization problem:

$$\min_{\mathbf{w}_i, \boldsymbol{\lambda}_i} R_\eta(\mathbf{w}_i, \boldsymbol{\lambda}_i) \tag{6.11}$$

where

$$R_\eta(\mathbf{w}_i, \boldsymbol{\lambda}_i) = \sum_{k=1}^{N} \left(x_i\,(t_k) - f_i\,(\mathbf{w}_i, \boldsymbol{\lambda}_i, t_k)\right)^2 + \eta \|\mathbf{w}_i\|_2^2 + \eta \|\boldsymbol{\lambda}_i\|_2^2 \tag{6.12}$$

subject to the linear constraints $\lambda_{ij} < 0$, $\forall j$ and bound constraints $a_i \leq w_{i0} \leq b_i$. The bound constraints incorporate into the optimization framework prior biological information on the location of the steady state. The values for the lower bound a_i and upper bound b_i will be discussed in Section 6.4. In (6.12), η represents the regularization parameter while $\|\cdot\|_2$ denotes the Euclidean norm.

The objective function (6.12) (without the regularization term) has very flat valleys and very steep slopes suggesting that the minimum is badly conditioned (Varah, 1985). Wiscombe and Evans (1977) remark that the non-linear least squares approach requires good initial estimates for $\mathbf{w}_i$ and $\boldsymbol{\lambda}_i$ and suggest using less sophisticated methods for exponential-sum fitting to get these estimates. However, these estimates can still be associated with local minima. In this work, the optimization problem (6.11) was solved using the trust-region reflective algorithm implemented in the Matlab Optimization Toolbox, for random initial values of the parameters and the best-fit solution was selected.

6.1.3 Selecting the regularization parameter using the Morozov's discrepancy principle

Consider the following non-linear ill-posed problem

$$F(\mathbf{q}) = \mathbf{g} \tag{6.13}$$

where $F : \mathcal{D}(F) \subset \mathcal{Q} \to \mathcal{G}$ is a non-linear operator between the infinite dimensional Hilbert spaces $\mathcal{Q}$ and $\mathcal{G}$, with norms $\|\cdot\|_{\mathcal{Q}}$ and $\|\cdot\|_{\mathcal{G}}$, respectively. Let $\mathbf{g} = F(\mathbf{q}^{\dagger})$ represent the unperturbed data corresponding to the true solution $\mathbf{q}^{\dagger}$ and $\mathbf{g}_{\delta}$ denote the noisy data with noise level δ according to:

$$\|\mathbf{g} - \mathbf{g}_{\delta}\| \leq \delta \tag{6.14}$$

Given the optimization problem:

$$\min_{\mathbf{q} \in \mathcal{D}(F)} \left\{ \|F(\mathbf{q}) - \mathbf{g}_{\delta}\|_{\mathcal{G}}^{2} + \eta \|\mathbf{q} - \mathbf{q}^{\dagger}\|_{\mathcal{Q}}^{2} \right\} \tag{6.15}$$

the optimal regularization parameter η according to the Morozov's discrepancy principle (Scherzer, 1993) satisfies the implicit equation:

$$\|F(\mathbf{q}_{\delta}^{\eta}) - \mathbf{g}_{\delta}\|_{\mathcal{G}}^{2} = \tau^{2}\delta^{2} \tag{6.16}$$

where $\mathbf{q}_{\delta}^{\eta}$ is a solution to (6.15) and $\tau > 1$ is some constant. Since in practice $\mathbf{q}^{\dagger}$ is not known in advance, a random initial guess is taken instead. To solve the discrepancy equation (6.16), Kaltenbacher et al. (2011) proposed and inexact Newton algorithm while Tautenhahn and Jin (2003) used a secant method.

Note that Morozov's discrepancy principle requires an estimate of the measurement noise δ. In the context of microarray data summarized with multi-mgMOS, δ can be estimated from the measurement errors associated with the gene expression levels. In this work, the noise level δ_i for gene i was taken as:

$$\delta_i = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \sigma_i^2(t_k)} \tag{6.17}$$

where $\sigma_i(t_k)$ is the uncertainty around $x_i\,(t_k)$. Using this estimate, the optimal regularization parameter for the optimization problem (6.11) can be found by solving for η the discrepancy equation

$$\|\mathbf{f}_i(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta, \mathbf{t}) - \mathbf{x}_i\|_2^2 = \tau^2 \delta_i^2 \tag{6.18}$$

where $\mathbf{f}_i(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta, \mathbf{t}) = \left(f_i(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta, t_1), \ldots, f_i(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta, t_N)\right)$, $\mathbf{x}_i = (x_i(t_1), \ldots, x_i(t_N))$ denotes the vector of noisy gene expression measurements while the pair $\left(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta\right)$ represents a minimizer of (6.12) for the current η. However, for genes with small δ_i and complex dynamics that can't be fully captured by the model (6.10), the discrepancy equation (6.18) may have no solution for fixed τ. For this reason, the regularization parameter η satisfying:

$$\arg\min_{\eta} \left| \|\mathbf{f}_i(\mathbf{w}_i^\eta, \boldsymbol{\lambda}_i^\eta, \mathbf{t}) - \mathbf{x}_i\|_2^2 - \tau^2 \delta_i^2 \right| \tag{6.19}$$

is selected from a fixed range $\mathcal{R}_\eta$.

6.2 Modelling GRNs using linear dynamical systems

Although the regulatory interactions between genes are nonlinear (Heinrich and Schuster, 1996), the gene expression dynamics can be captured by linear dynamical systems modelled around the equilibrium points of the GRN (Gustafsson et al., 2005).

Let $\mathbf{x}(t) \in \mathbb{R}^n$ denote the n-dimensional vector containing the expression level of n genes at time point t. The regulatory interactions between these genes in the neighbourhood of the attractor $\mathbf{x}_e$ are modelled using the system of ordinary differential equations (6.1), where the stable matrix $\mathbf{A}$ encodes the coupling between genes. The problem of reconstructing GRNs modelled using (6.1) consists of estimating the parameters $\mathbf{w}_0 = (w_{10}, w_{20}, \ldots, w_{n0})$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{W}$ from time course gene expression data. Knowing $\mathbf{W}$ and $\boldsymbol{\lambda}$, the system matrix $\mathbf{A}$ can be recovered using the spectral decomposition $\mathbf{A} = \mathbf{W}\boldsymbol{\Delta}\mathbf{W}^{-1}$. A method for estimating the model parameters in the form of regularized nonlinear least squares optimization in presented in Section 6.2.1 while a strategy for selecting the optimal regularization parameter is discussed in Section 6.2.2.

6.2.1 Parameter estimation using non-linear least squares

Given time course gene expression measurements $\{\mathbf{x}(t_k)\}_{k=1}^N$, estimating the parameters $\mathbf{w}_0, \boldsymbol{\lambda}$ and $\mathbf{W}$ consists of solving the regularized nonlinear optimization problem:

$$\min_{\mathbf{w}_0, \boldsymbol{\lambda}, \mathbf{W}} R_\eta(\mathbf{w}_0, \boldsymbol{\lambda}, \mathbf{W}) \tag{6.20}$$

where

$$R_\eta(\mathbf{w}_0, \boldsymbol{\lambda}, \mathbf{W}) = \sum_{i=1}^{n}\sum_{k=1}^{N}\left(x_i(t_k) - w_{i0} - \sum_{j=1}^{n} w_{ij}e^{\lambda_j t_k}\right)^2 + \eta\|\boldsymbol{\lambda}\|_2^2 + \eta\sum_{i=0}^{n}\|\mathbf{w}_i\|_2^2 \quad (6.21)$$

subject to the linear constraints $\lambda_j < 0, \forall j$.

Note the similarity between the objective functions (6.12) and (6.21). The former measures the fitting error between the time series data of one gene and the sum of exponentials model (6.10) while the latter simultaneously quantifies the fitting error between the time course gene expression data of n genes and the sum of exponentials models (6.9).

The objective function (6.21) contains $n^2 + 2n$ parameters. Using time course microarray data satisfying $N < n + 2$ for parameter estimation will result in overfitting. It follows that for reliable parameter estimation, the size of the network needs to be adjusted to the length of the time series. For short-time series gene expression data this represents a limiting factor for reconstructing GRNs of arbitrarily large sizes. This problem can be avoided by modelling the gene expression dynamics as described in section 6.1 and sampling the resulting models to generate sufficient gene expression data to avoid over-fitting. Note that in this case, the number of parameters in the optimization problem (6.21) can be reduced by n if $\mathbf{x}_e$ is taken as the vector containing the steady-state values of the fitted gene expression models.

The regularization term in (6.21) can take the form of the non-linear constraint:

$$\|\boldsymbol{\lambda}\|_2^2 + \sum_{i=0}^{n}\|\mathbf{w}_i\|_2^2 \leq \zeta \quad (6.22)$$

where the regularization parameter ζ is inversely related to η. Removing the regularization term has the advantage that the objective function:

$$R_\eta(\mathbf{w}_0, \boldsymbol{\lambda}, \mathbf{W}) = \sum_{i=1}^{n}\sum_{k=1}^{N}\left(x_i(t_k) - w_{i0} - \sum_{j=1}^{n} w_{ij}e^{\lambda_j t_k}\right)^2 \quad (6.23)$$

quantifies only the fitting error. This quantity will be used in the next section to select an appropriate value for ζ.

Similar to (6.11), the optimization problem (6.20) is sensitive to the initial values of the parameters $\mathbf{w}_0$, $\boldsymbol{\lambda}$ and $\mathbf{W}$. To tackle this problem, (6.20) was solved for random initial values of the model parameters and the best fit solution was selected. The sequential quadratic programming algorithm implemented in the Matlab Optimization Toolbox was used due to its robustness to non-double results. This means that the algorithm takes smaller steps to prevent the objective function (6.21) from returning complex values, Inf (positive infinity) of NaN (not-a-number), as opposed to the active-set algorithm implemented in the same optimization toolbox.

6.2.2 Selecting the regularization parameter by cross-validation

Section 6.1.3 proposed the Morozov's discrepancy principle for selecting the regularization parameter in non-linear ill-posed problems. In the context of exponential-sum fitting to short time course gene expression data, this method is appealing as it requires only the measurement errors and therefore allows the whole time series data to dictate the model complexity and be used for parameter estimation.

The optimization problem (6.20) also consists of fitting sums of exponentials to noisy gene expression data. Thus, the Morozov's discrepancy principle could be used to select the regularization parameter, given the relative sizes of the measurement uncertainties. However, since the noise estimates are available only for the short time-course gene expression data, the Morozov's discrepancy principle can be applied only for small size GRNs.

By modelling the time course gene expression data using sums of exponentials, artificial measurements can be obtained for the unobserved time points and used to fit large scale GRNs. Taking advantage of the abundance of the artificial samples that can be generated, the regularization parameter ζ can be selected using k-fold cross-validation as described below.

Required Gene expression data $\{\mathbf{x}(t_k)\}_{k=1}^{N}$
Initial estimates $\mathbf{w}_0^0, \boldsymbol{\lambda}^0, \mathbf{W}^0$
Range $\mathcal{R}_\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_k\}$

Step 1. Split $\{\mathbf{x}(t_k)\}_{k=1}^{N}$ into k random parts of roughly equal sizes

Step 2. For every $\zeta_i \in \mathcal{R}_\zeta$

1. Solve (6.20) with objective function (6.23) subject to the linear constraints $\lambda_j < 0$ $(\forall j)$ and non-linear constraint (6.22) starting from $\mathbf{w}_0^0$, $\boldsymbol{\lambda}^0$, $\mathbf{W}^0$ using $k-1$ parts
2. Calculate the fitting error (6.23) on the hold-out part using the optimal parameters estimated in (1)
3. Repeat (1) and (2) for all possible choices of the hold-out part
4. Sum the fitting errors calculated across hold-out parts

Step 3. Select $\zeta_{opt} \in \mathcal{R}_\zeta$ with the minimum sum of fitting errors

The final model is obtained by estimating the parameters using the whole data $\{\mathbf{x}(t_k)\}_{k=1}^{N}$ and ζ_{opt}. In the context of linear inverse problems, the above procedure for the particular case when one observation is left out (leave-one-out) is widely known as the generalized cross-validation (GCV) method for selecting the regularization parameter (Golub et al., 1979). This approach admits a formulation that avoids solving the inverse problem for each left-out observation (Wahba, 1990).

Haber and Oldenburg (2000) extended the GCV method to non-linear inverse problems solved using iterative methods. The method consists of applying GCV to the linearized approximation of the non-linear ill-posed problem at each iteration. When the linearized approximation represents an adequate description of the non-linear ill-posed problem, the approach has the same performance as in the case of linear inverse problems (Farquharson and Oldenburg, 2004).

6.3 Estimating sparse GRNs

Solving the optimization problem (6.20) produces complete network topologies, given the absence of sparsity constraints. Methods for enforcing topological sparsity are based on l_1-norm regularization of the coefficients of $\mathbf{A}$.

Gustafsson et al. (2005) proposed a method to estimate sparse (but not necessarily stable) GRNs using l_1-norm regularization of the row coefficients of $\mathbf{A}$. Wu et al. (2010) used l_1-norm regularization of the entries of $\mathbf{A}$ when estimating stable GRNs and noted that this approach leads to low magnitude coefficients. Perrin et al. (2003) used a similar regularization approach and argued that for large values of the regularization parameter, the l_1-regularization term encourages real sparseness. Finally, Zavlanos et al. (2011) used a weighted l_1-norm of the entries of $\mathbf{A}$ to estimate stable and sparse GRNs from steady-state gene expression data.

In the context where the objective function (6.23) is not parametrized in terms of the entries of $\mathbf{A}$ but rather in terms of its eigenvectors and eigenvalues, none of the these approaches can be applied to enforce topological sparsity. Also note that since sparse matrices are not always associated with sparse eigenvectors, using l_1-norm regularization on the eigenvectors can induce bias in the structure of $\mathbf{A}$.

To obtain a sparse representation of the GRN modelled using (6.1), weak connections (entries of $\mathbf{A}$ close to 0) are removed using the following approach proposed by Perrin et al. (2003):

Step 1. Derive a family of connectivity matrices $\mathcal{A} = \left\{\mathbf{A}^{(k)} | k = 1 \ldots K\right\}$, by solving (6.20) from K random initial estimates for $\mathbf{w}_0^0, \boldsymbol{\lambda}^0, \mathbf{W}^0$

Step 2. Compute $\bar{\mathbf{A}}$, $\boldsymbol{\Sigma}$, where $\bar{a}_{ij}$ and σ_{ij}^2 represents the mean and variance of $a_{ij}^{(k)}$, respectively.

Step 3. Compute σ^2, the variance of the entries $\bar{a}_{ij}$

Step 4. Set to zero entries of $\bar{\mathbf{A}}$ satisfying $|\bar{a}_{ij}| < \sigma$ and $\sigma_{ij}^2 < \sigma$.

The remaining positive and negative coefficients represent probable up-regulating and down-regulating connections, respectively. Another related approach to removing weak connections (not used in this work) was proposed by Rangel et al. (2004). The method consists of setting to zero parameters for which the confidence intervals computed using bootstrap estimates contain the value 0.

6.4 Modelling GRNs for the diagnostic groups of ACS

The novel GRN reconstruction algorithm was applied on the combined data to estimate for each diagnostic group of ACS the signalling pathways between the differentially expressed genes identified in Chapter 4 using the multi-stage feature selection method. Specifically, the studies presented in Section 6.5.1 and Section 6.5.2 focus on the regulatory interactions between the genes that differentiate MI from UA and NSTEMI from STEMI, respectively. Each study consists of estimating stable GRNs and computing their sparsity pattern.

6.4.1 Modelling the regulatory interactions between the genes differentiating MI from UA

In this section the regulatory interactions between the genes $X_1 - X_{20}$ listed in Table B.3 of Appendix B were estimated for each diagnostic group. Given that the number of temporal gene expression measurements in the combined dataset ($N = 5$) was insufficient for unbiased estimation of the network parameters, the gene expression dynamics were approximated using sums of exponentials and sampled as discussed below.

Fitting sums of exponentials to gene expression data

For each diagnostic group, given the combined expression data $\{x_i(t_k)\}_{k=1}^N$ for gene X_i, where $t_k \in \{1, 3, 7, 30, 90\}$, and the associated standard errors $\{\sigma_i(t_k)\}_{k=1}^N$, the parameters of the gene expression model (6.10) with $m = 2$ were estimated by solving the nonlinear optimization problem (6.11) subject to the linear constraints $\lambda_{ij} < 0, j = 1 \dots m$, and additional boundary constraints $x_i(t_N) - 3\sigma_i(t_N) \leq w_{i0} \leq x_i(t_N) + 3\sigma_i(t_N)$. These later constraints localize the gene expression steady state

w_{i0} in the neighbourhood (99.7% confidence interval) of the final time point measurement to capture the long transient behaviour of cardiac markers (settling time up to 14 days for the known biomarkers of ACS (Ahmad and Sharma, 2012)). The regularization parameter γ satisfying (6.19) for $\tau = 1.1$ was selected from the range $\mathcal{R}_\eta = \{10^{-3}, 10^{-2.9}, \dots, 10^{-2.9}, 10^{3}\}$.

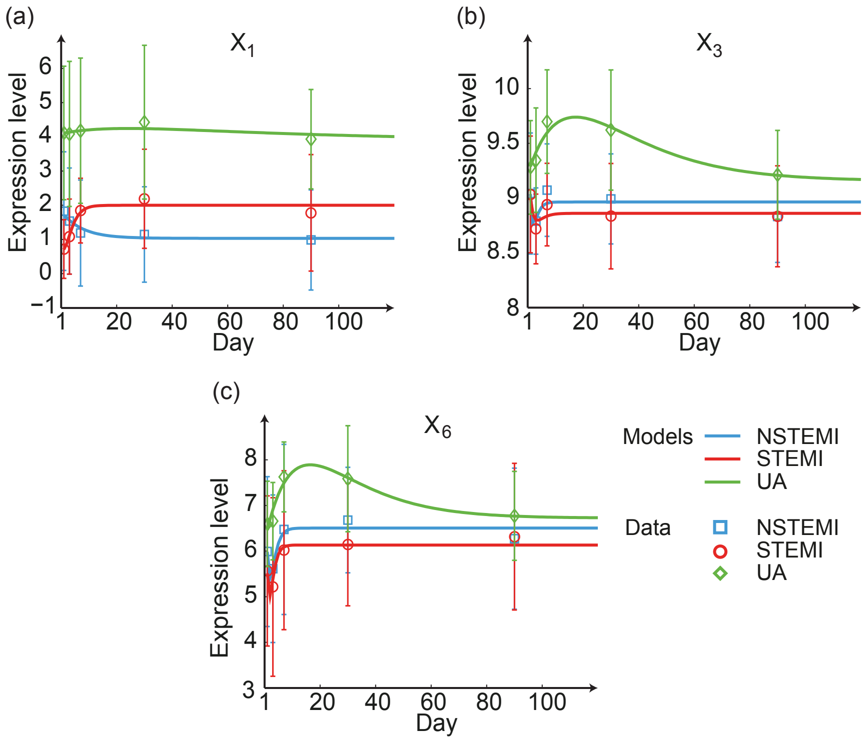

Figure 6.1: Estimated dynamic profiles for (a) X_1 (WASH1), (b) X_3 (C17orf103) and (c) X_6 (OSBP2). The bars denote one standard deviation around the group-specific gene expression averages.

For each X_i, the optimization was repeated a number of $K_{opt} = 30$ times from random initial guesses of model parameters and the best fit solution was selected. Specifically, the linear parameters w_{ij} $(j = 0 \dots m)$ were sampled from the normal distribution $\mathcal{N}(0, 10)$ while the nonlinear parameters λ_{ij} were sampled from the uniform distribution $\mathcal{U}(-5, 0)$. The dynamic profiles of the genes differentially expressed across all time point are shown in Figure 6.1. The dynamic profiles of the remaining genes, shown in Figure D.1 and Figure D.2 of Appendix D, also shade light on the long term differences between MI and UA. The residual sums of squares for each in each diagnostic group, listed in Table D.1, show that the sums of exponentials model adequately captures the gene expression dynamics. The expression trajectories for the genes $X_1 - X_{20}$ were sampled using a sampling period of $\Delta t = 1$ day over a time period of $T = 120$ days. The resulting dataset was used to estimate stable GRNs, as described in the next section.

Estimation of the network parameters

For each diagnostic group, given the sampled gene expression data $\{\mathbf{f}(t_k)\}_{k=1}^{\top}$ where $\mathbf{f}(t_k) = (f_1(\mathbf{w}_1, \boldsymbol{\lambda}_1, t_k), \ldots, f_{20}(\mathbf{w}_{20}, \boldsymbol{\lambda}_{20}, t_k))$ and $t_k = 1 \ldots 120$, the parameters of the GRN were estimated as described in Section 6.2.1. To reduce the number of parameters, $\mathbf{x}_e$ was taken as the vector containing the steady-state values of the fitted gene expression models. The regularization parameter ζ was selected from the range $\mathcal{R}_\zeta = \{1, 10^{0,1}, \ldots, 10^{2.9}, 10^3\}$ using a 10-fold cross-validation partitioning of the data $\{\mathbf{f}(t_k)\}_{k=1}^{\top}$ following the steps presented in Section 6.2.2. For each diagnostic group, the optimization was repeated a number of $K_{opt} = 20$ times from random initial values of the model parameters and the best fit solution was selected. Each time, the entries of $\mathbf{W}^0$ were sampled from the normal distribution $\mathcal{N}(0, 10)$ while the entries of $\boldsymbol{\lambda}^0$ were taken from the uniform distribution $\mathcal{U}(-10, 0)$.

Table 6.1: Eigenvalues of the GRNs associated with the ACS subtypes

Eigenvalue	NSTEMI	STEMI	UA
λ_1	-1.8222	-2.0169	-1.9416
λ_2	-0.9555	-1.2313	-0.6902
λ_3	-0.4588	-1.2305	-0.2722
λ_4	-0.4487	-0.7029	-0.2311
λ_5	-0.4362	-0.6329	-0.1388
λ_6	-0.4334	-0.6294	-0.1168
λ_7	-0.2602	-0.5961	-0.1157
λ_8	-0.1805	-0.5351	-0.1094
λ_9	-0.1555	-0.4954	-0.1052
λ_{10}	-0.1216	-0.4786	-0.0884
λ_{11}	-0.0854	-0.4463	-0.0801
λ_{12}	-0.0804	-0.4174	-0.0791
λ_{13}	-0.0729	-0.3695	-0.0741
λ_{14}	-0.0521	-0.2739	-0.0738
λ_{15}	-0.0425	-0.2446	-0.0445
λ_{16}	-0.0299	-0.2421	-0.0414
λ_{17}	-0.0285	-0.1076	-0.0379
λ_{18}	-0.0216	-0.0613	-0.0346
λ_{19}	-0.0116	-0.0137	-0.0170
λ_{20}	-0.0001	-0.0033	-0.0064

The estimated GRNs are stable dynamical systems. Table 6.1 lists the eigenvalues for the GRN of each diagnostic group. The values of ζ_{opt} for the NSTEMI, STEMI and UA networks are 15.84, 25.11 and 63.09, respectively. These values are used in the next section to estimate families of group-specific GRNs. To measure the prediction performance of the estimated models, the root mean square percentage error (RMSPE) (Holden et al., 1990) between each ith trajectory approximated using sums of exponentials and the ith trajectory resulted from simulating the networks using $\mathbf{f}(t_1)$ as initial condition, was computed for all $i = 1 \ldots 20$. The average RMSPEs (across trajectories) for the NSTEMI, STEMI and UA networks are 0.020%, 0.014% and 0.026%, respectively. These results show that the state trajectories match extremely well the trajectories approximated using sums of exponentials.

Sparse representations of the group-specific GRNs

The procedure described in Section 6.3 was applied to obtain sparse representations of the group-specific GRNs. Specifically, for each diagnostic group, a family $\mathcal{A}$ of $K = 50$ matrices was derived using $\{\mathbf{f}(t_k)\}_{k=1}^{\top}$, ζ_{opt} listed in the previous section and initial values $\mathbf{W}^0$ and λ^0 sampled from $\mathcal{N}(0, 10)$ and $\mathcal{U}(-10, 0)$, respectively. Each family consists of networks with high prediction performance (low average RMSPEs), as shows in Figure 6.2.

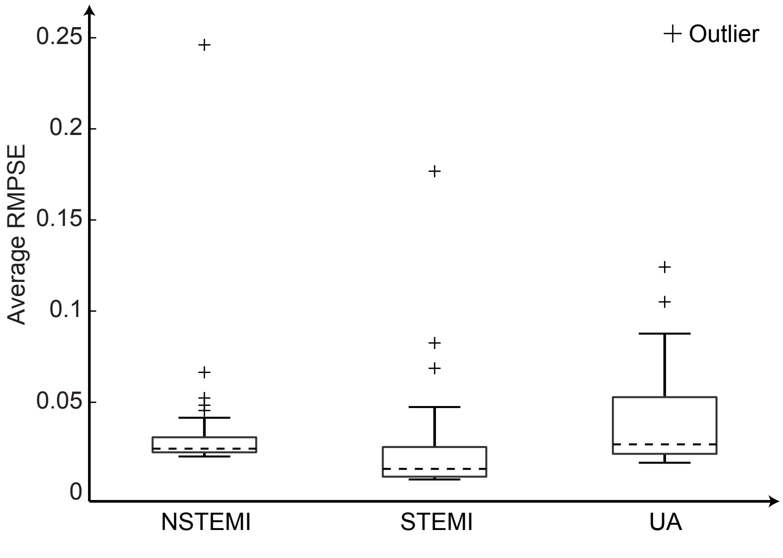

Figure 6.2: Distributions of the prediction errors for the families of GRNs.

The filtering areas for the entries of the mean connectivity matrices $\bar{\mathbf{A}}$ associated with each diagnostic group are shown in Figure 6.3. The standard deviations of the entries of $\bar{\mathbf{A}}$ for the NSTEMI, STEMI and UA groups are 2.808, 3.028 and 1.731, respectively. These values denote the boundaries of the filtering areas associated with each group. From the total number of $N = 400$ possible connections for each network, after filtering the group-specific connectivity graphs consisted of: $N_{NSTEMI} = 84$ connections (38 down-regulating and 46 up-regulating), $N_{STEMI} = 169$ connections (88 down-regulating and 81 up-regulating), and $N_{UA} = 127$ connections (58 down-regulating and 69 up-regulating). For scalability purposes, the regulators for genes $X_1 - X_5$ in each diagnostic group are shown in Figure 6.4. The complete connectivity maps for each diagnostic group are shown in Figure D.5, Figure D.6 and Figure D.7 of Appendix D.

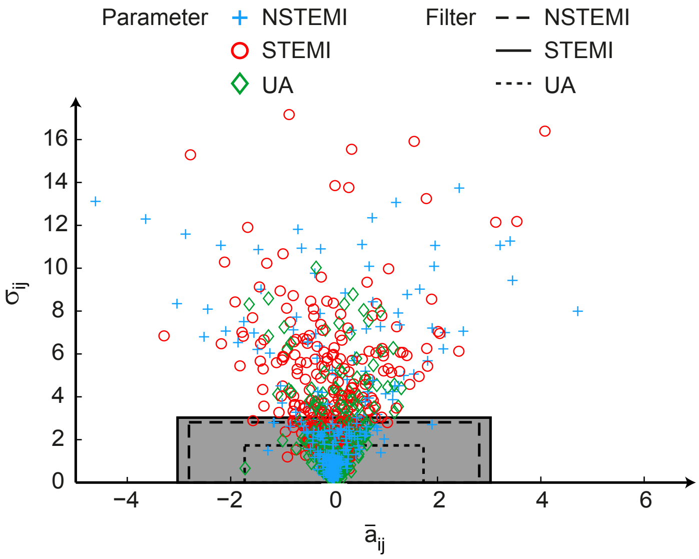

Figure 6.3: Sparsity filters for the parameters of the GRNs associated with NSTEMI, STEMI and UA

6.4.2 Modelling the regulatory interactions between the genes differentiating NSTEMI from STEMI

In this section the regulatory interactions between the genes listed in Table B.6 of Appendix B were estimated for each diagnostic group. Since the probe sets X_6 and X_7 target the same gene, only the measurements of the most frequently selected probe set (X_6) were used to represent gene NCAM1. To generate enough data for the estimation of large-scale GRNs, the gene expression dynamics were approximated using sums of exponentials and sampled as discussed below.

Fitting sums of exponentials to gene expression data

Modelling the time-course dynamics of the combined gene expression data (for each diagnostic group) was performed using the setting described in the previous study. The dynamic profiles of the genes differentially expressed across all time point are shown in Figure 6.5. The dynamic profiles of the remaining genes,

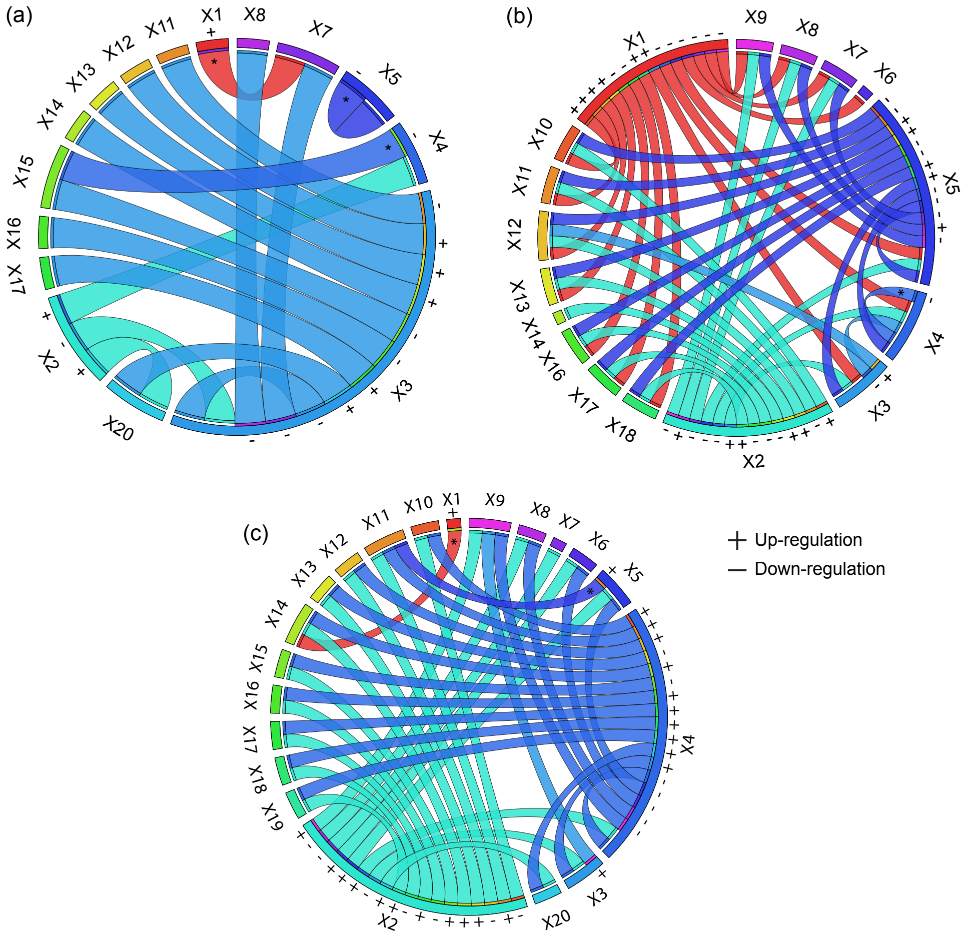

Figure 6.4: Regulatory map for genes X_1-X_5 in the: (a) NSTEMI network, (b) STEMI network and (c) UA network. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The ($*$) marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene.

shown in Figure D.8 and Figure D.9 of Appendix D, also shade light on the long term dissimilarity between MI and UA. The residual sums of squares for the fitted genes are listed in Table D.2. The expression trajectories for the unique genes were sampled using a sampling period of $\Delta t = 1$ day over a time period of $T = 120$ days. The resulting dataset was used to estimate stable GRNs, as described in the next section.

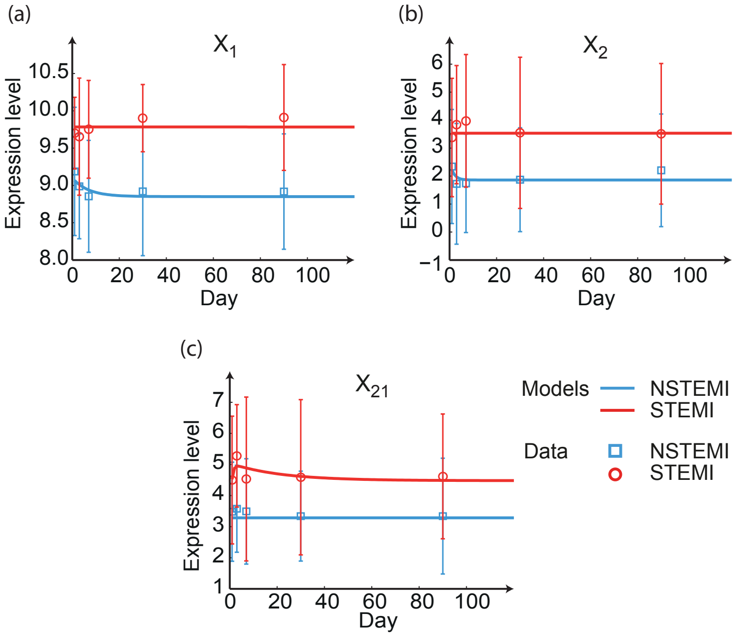

Figure 6.5: Estimated dynamic profiles for (a) X_1 (HLA-DQB1), (b) X_2 (MAPK8IP1) and (c) X_{21} (LRRC37A). The bars denote one standard deviation around the group-specific gene expression averages.

Estimation of the network parameters

Stable GRNs for the NSTEMI and STEMI groups were estimated using the parameters settings of the previous study. Table 6.2 lists the eigenvalues for the GRN of each diagnostic group. The values of ζ_{opt} for the NSTEMI and STEMI networks are 31.62 and 125.89, respectively. The average RMSPEs (across trajectories) for the NSTEMI and STEMI networks are 0.002% and 0.008%, respectively, suggesting that the state trajectories match extremely well the trajectories approximated using sums of exponentials.

Sparse representations of the group-specific GRNs

Families of $K = 50$ connectivity matrices were obtained for the GRNs of NSTEMI and STEMI groups, using the parameter settings and approach described in the previous study. Each family consists of networks with high prediction performance (low average RMSPEs), as shows in Figure 6.6. The filtering areas for the entries of the mean connectivity matrices $\bar{\mathbf{A}}$ associated with each diagnostic group are shown in Figure 6.7. The standard deviations of the entries of $\bar{\mathbf{A}}$ for the NSTEMI and STEMI are 2.64 and 2.61, respectively. From the total number of $N = 400$ possible connections for each network, after filtering the group-specific connectivity graphs consisted of: $N_{NSTEMI} = 136$ connections (74 down-regulating

Table 6.2: Eigenvalues of the GRNs associated with NSTEMI and STEMI

Eigenvalue	NSTEMI	STEMI
λ_1	-2.4359	-4.3599
λ_2	-1.8141	-2.9732
λ_3	-1.1130	-2.1789
λ_4	-1.0566	-1.1648
λ_5	-0.8944	-0.8910
λ_6	-0.8108	-0.7638
λ_7	-0.7835	-0.6522
λ_8	-0.7717	-0.5926
λ_9	-0.7544	-0.5743
λ_{10}	-0.7477	-0.2125
λ_{11}	-0.7086	-0.1782
λ_{12}	-0.6943	-0.1157
λ_{13}	-0.6624	-0.1009
λ_{14}	-0.4034	-0.0898
λ_{15}	-0.3160	-0.0848
λ_{16}	-0.2265	-0.0570
λ_{17}	-0.0667	-0.0043
λ_{18}	-0.0207	-0.0032
λ_{19}	-0.0084	-0.0021
λ_{20}	-0.0008	-0.0018

and 62 up-regulating) and $N_{STEMI} = 199$ connections (100 down-regulating and 99 up-regulating). The regulators for genes $X_1 - X_5$ in each diagnostic group are shown in Figure 6.8. The complete connectivity maps are not shown due to the large number of connections in each network.

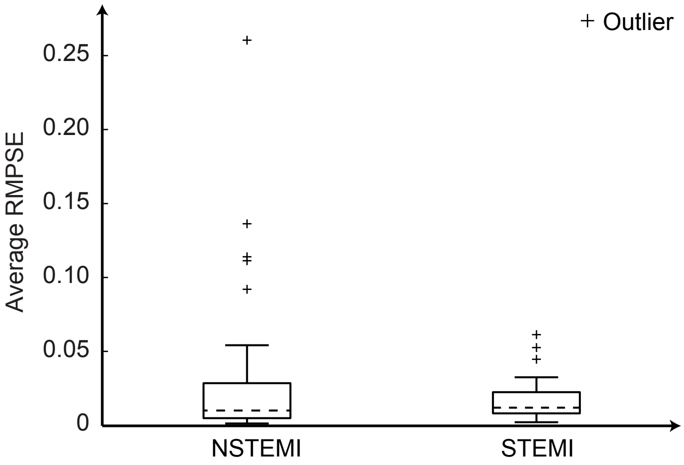

Figure 6.6: Distributions of the prediction errors for the families of GRNs

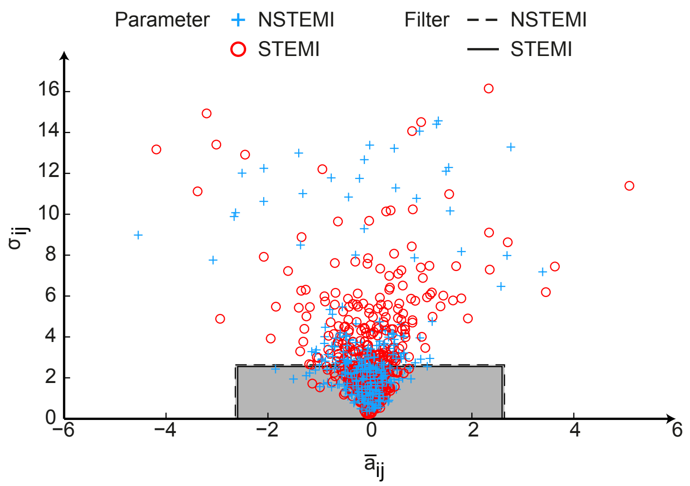

Figure 6.7: Sparsity filters for the parameters of the GRNs associated with NSTEMI and STEMI

6.5 Conclusions

This chapter proposed modelling gene expression dynamics using sums of exponentials, on the grounds that this formulation describes the response of stable dynamical systems. Estimation of the model parameters from unequally sampled microarray data was formulated as a regularized nonlinear optimization problem that incorporates information about the measurement noise in the gene expression data. This modelling approach was shown to adequately capture the time course dynamics of the genes differentiating between the subtypes of ACS.

A novel approach for modelling stable GRNs using linear dynamical systems was also derived. This approach estimates the parameters of the transition matrix from time course gene expression data by solving a regularized nonlinear optimization problem that also allows for the measurement noise to be incorporated into the estimation process. Since the method directly operates on the transition matrix, the need for computing derivatives is avoided. The novel modelling approach was used to estimate stable GRNs for each group of ACS, given uniformly sampled gene expression trajectories approximated using sums of exponentials.

Given that the estimation of the network parameters is performed in the absence of sparsity constraints, an approach that removes weak connections was discussed. The sensitivity (proportion of connections correctly identified) and specificity (proportion of non-connections correctly identified) of this approach wasn't reported in the literature. Future research will focus on estimating these quantities for the nonlinear problem at hand. Additionally, novel method for incorporating sparsity constraints will be researched.

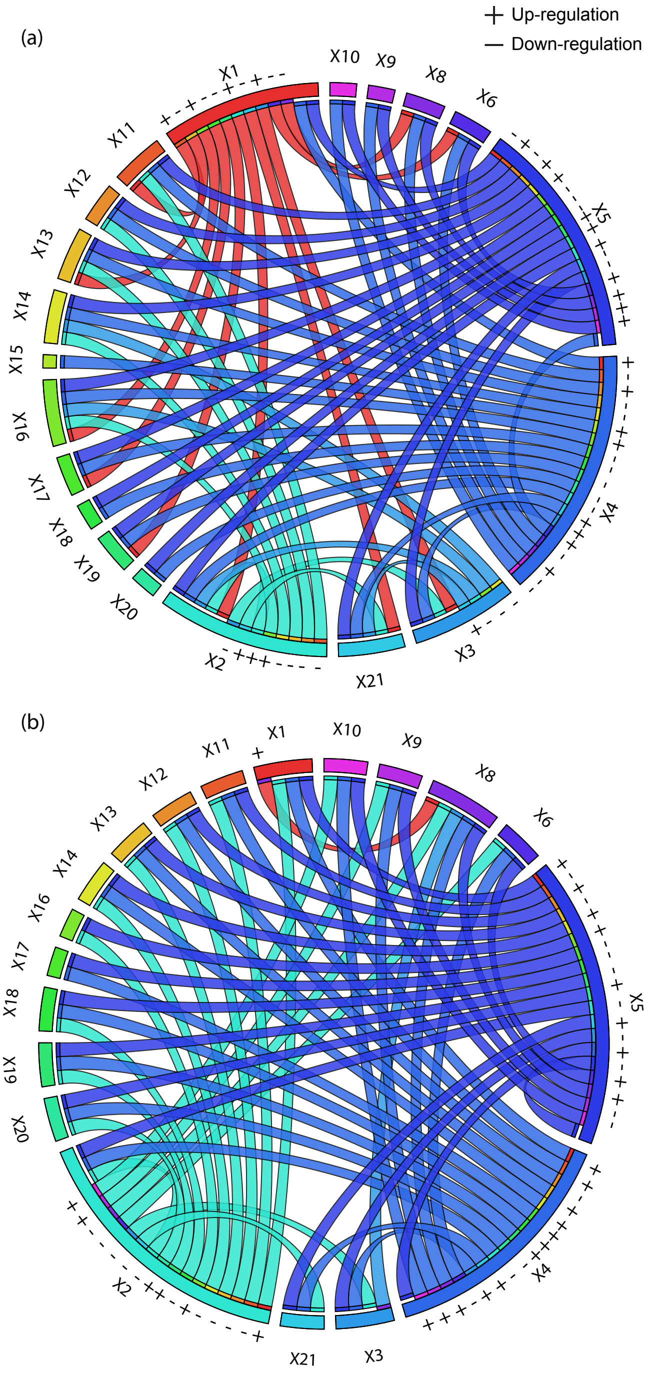

Figure 6.8: Regulatory map for genes X_1-X_5 in the: (a) NSTEMI network and (b) STEMI network. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it

Chapter 7

Conclusions

7.1 Summary and Conclusions

This thesis proposed three novel computational tools addressing major challenges related to the genetics of ACS that can provide an insight into the dynamics of the disease leading to better diagnosis solutions and improved personalized treatments. These challenges consists of:

- Selecting differentially expressed genes between the ACS subtypes
- Deconvolving heterogeneous microarray gene expression data
- Inferring gene regulatory pathways

The performance of the proposed methods was demonstrated on a real world dataset consisting of ACS time-course microarray gene expression data and associated blood count measurements.

The novel feature selection method consists of four stages imposing stage-specific levels of stringency and operating in a nested cross-validation fashion to avoid the parameter selection bias (internal loop) and the feature selection bias (external loop) and provide an unbiased estimate of the discriminatory power of selected genes. Two differential expression studies comparing the novel multi-stage method against the l_1-StaR algorithm on the task of identifying genes discriminating between the ACS subtypes showed that (i) the two approaches produced subsets of genes with comparable high diagnostic performance (the l_1-StaR performing slightly better in both studies, (ii) the multi-stage method selected, on average, less genes than l_1-StaR and (iii) that the genes selected by the multi-stage method show longer-term differential expression (up to three months) than the genes selected by l_1-StaR. These findings suggested that: (i) the multi-stage method is more appropriate for biomarker discovery in time-course microarray studies, (ii) the transcriptomic signature of the ACS subtypes is distinguishable three months after hospital admission, (iii) and that genes showing long term differential expression could become reliable biomarkers for late diagnosis or explain the genetic predisposition to ACS.

The novel deconvolution method for microarray gene expression data uses non-negative optimization within the OFR approach to identify the cellular sources of gene expression and measure their contribution to the abundance and variability of the heterogeneous gene expression patterns. An approach based on measuring interaction effects was proposed for cell type-specific differential expression analysis, which is computationally superior to permutation based methods and can be used for single gene analysis. The deconvolution method applied on the genes differentiating between the ACS subtypes demonstrated high performance in capturing the variability in the expression measurements for the majority of the genes (comparable to the performance of the unconstrained OFR approach), exposed the cell type sources of gene expression and revealed features of within and between groups variation in gene expression patterns. The approach for cell type-specific differential expression analysis identified cell types contributing differently in the case and control groups to the abundance of the genes discriminating between the ACS subtypes. Genes expressed differently in the same cell type across groups could represent cell type-specific markers for ACS.

The novel method to model GRN using stable linear dynamical systems relies on non-linear optimization techniques to recover the state transition matrix of the system. This approach (i) incorporates stability constraints, (ii) can handle unequally sampled time series data, (iii) can account for the measurement noise via regularization and (iv) bypasses the need for derivatives. When reconstruction of large scale GRN is impeded by the scarcity of time course measurements, the novel approach for modelling gene expression dynamics using sums of exponentials can be used to generate enough data. This approach formulates parameter optimization as a regularized non-linear optimization problem that presents the advantages (i)-(iii) of the GRN reconstruction method. It was shown that models consisting of exponential can capture the dynamics of the genes differentiating between the ACS subtypes, suggesting that the gene expression levels return to baseline after a transient regime initiated by the ACS events. The stable GRNs estimated for each subtype of ACS using data sampled from the gene expression profiles approximated using sums of exponentials have high prediction performance (as measured by the mean square percentage error) and vary in their sparse topological structures.

The biological findings presented in this thesis consist of or rely upon a set of genes whose association with ACS hasn't been previously reported in the literature. While the novelty of the results is the product of the new computational tools applied on a time-course microarray dataset capturing the gene expression dynamics over a period of three months after the ischemic episodes, their biological relevance strongly depends on the size of the study cohort. Validation of the findings require further medical investigations, preferably performed on a larger cohort of participants including healthy individuals on top of ACS patients.

7.2 Future work

The following points highlight current limitations of the study and suggest possible solutions and research directions:

- The novel multi-stage feature selection method addresses only binary classification problems. To extend the algorithm to multiclass problems, two major changes need to made. Firstly, the statistical test used at Stage II must be replaced by another test comparing more then two groups. Widely used parametric and non-parametric test for multiclass problems are analysis of variance (ANOVA)(Scheffe, 1999) and the Kruskal-Wallis test (Kruskal and Wallis, 1952), respectively. Secondly, the SVM classifier used at Stage IV must be replaced with a multiclass extension (Hsu and Lin, 2002). Note that the unsupervised filter used at Stage I doesn't require information about the class of the arrays while the mRMR algorithm used at Stage III can operate on multiclass data.

- The study conducted in Chapter 4 looked at differentially expressed genes between the ACS subtypes. In particular, to identify potential cardiac markers for the long term diagnosis of MI, the UA cohort was taken as the control group. While this setup aims at separating MI from non-MI episodes, it may omit genes related to ACS that are expressed at comparable levels in both cohorts. Conducting a differential expression analysis with a cohort of healthy individuals could supplement the list of genes related to ACS. Additional gene expression profiling experiments need to be performed to collect microarray data from a cohort of healthy participants presenting similar clinical variables as the ACS cohort (age, sex, type II diabetes, smoking habits, family history, race, alcohol consumption) to minimize the biases of the exploratory analysis.

- The method for conducting cell type-specific differential analysis presented in Chapter 5 disregards the goodness of fit of the regression models as do permutation based approaches (Shen-Orr et al., 2010). Studying the bias in the estimated regression coefficients as a function of the fitting performance could provide the prior information for increasing the sensitivity and specificity of the cell type-specific differential analysis approach. Additionally, investigating the sensitivity the regression coefficients through systematic degradation studies (sequential removal of regressors) could shed light into the precision of the estimated cell type-specific expression signatures when measurements of the cellular proportions are available only for some cell types in the mixture.

- The sums of exponentials proposed in Chapter 6 for modelling gene expression dynamics assume real nonlinear parameters. While this modelling approach was shown to accurately capture the dynamics of the genes discriminating between the ACS subtypes, it may not be appropriate for genes exhibiting transient oscillations or cyclic patterns of expression. Transient oscillations could be captured by allowing complex nonlinear parameters. Parameter optimization could be performed using a constrained Prony-like method (Potts and Tasche, 2013). Testing for periodic patters in equally or unequally spaced time series gene expression data can be carried out using the Fourier transform (Spellman et al., 1998) or the Lomb-Scargle periodogram (Glynn et al., 2006). The presence of cyclic regulation requires a modelling approach different from sums of exponentials fitting, which is based on the assumption that after the transient response to an external perturbation, the gene expression level returns to a constant steady state value.

- The novel approach to model GRNs using stable state-space models assumes that the eigenvalues of the dynamical system are real. Future research will investigate whether stable state-space models can be estimated using constrained Prony-like methods. Additionally, strategies to incorporate sparsity constraints will be researched and the sensitivity and specificity of the approach estimating spare topological representations (presented in Section 6.3) will be evaluated.

Appendix A

Fundamentals of genetics

Deoxyribonucleic acid (DNA) is a nucleic acid that carries the genetic information in all organisms (except some viruses), consisting of two antiparallel and complementary polynucleotide chains twisted in the form of a double helix and joined by hydrogen bonds. Each nucleotide is composed of a five-carbon sugar (deoxyribose) attached to phosphate group and a nitrogenous base, which can be either a purine base such as adenine (A) or guanine (G), or a pyrimidine base such as cytosine (C) or thymine (T). The two groups of bases complement each other and can only form hydrogen bonds with the opposing type (A with T and G with C). The structure of a DNA molecule is shown in Figure A.1 (Alberts, 2008).

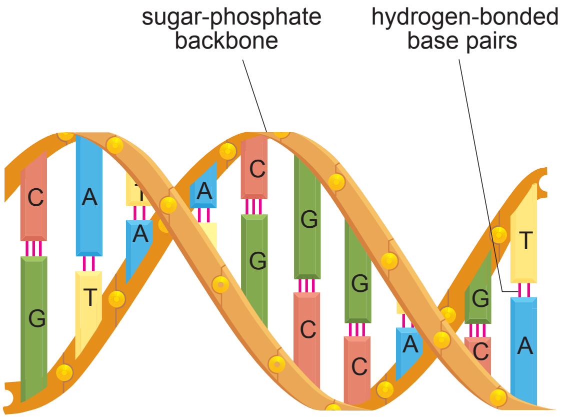

Figure A.1: DNA double helix

Ribonucleic acid (RNA) is a nucleic acid consisting of a single strand of nucleotides, often folded unto itself. The RNA nucleotide differs from the DNA nucleotide in that it has ribose instead of deoxyribose as a sugar backbone, and the pyrimidine base uracil instead of thymine. The structure of a RNA molecule is shown in Figure A.2.

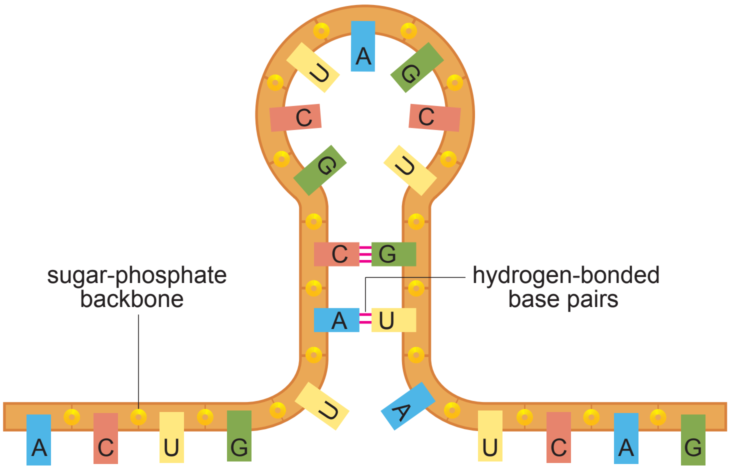

Figure A.2: RNA hairpin loop

A **gene** represents the physical and functional unit of heredity consisting of a segment of DNA that provides the coded instructions for the transcription of RNA molecules, which are translated into proteins or regulate the expression or activation of other genes. Most eukaryotic genes consists of coding regions (exons) separate by non-coding regions(introns). During transcription, introns are removed from the primary RNA trascript by splicing and the exons are covalently joined together to form a mature messenger ribonucleic acid (mRNA) molecule which is translated into proteins, as shown in Figure A.3.

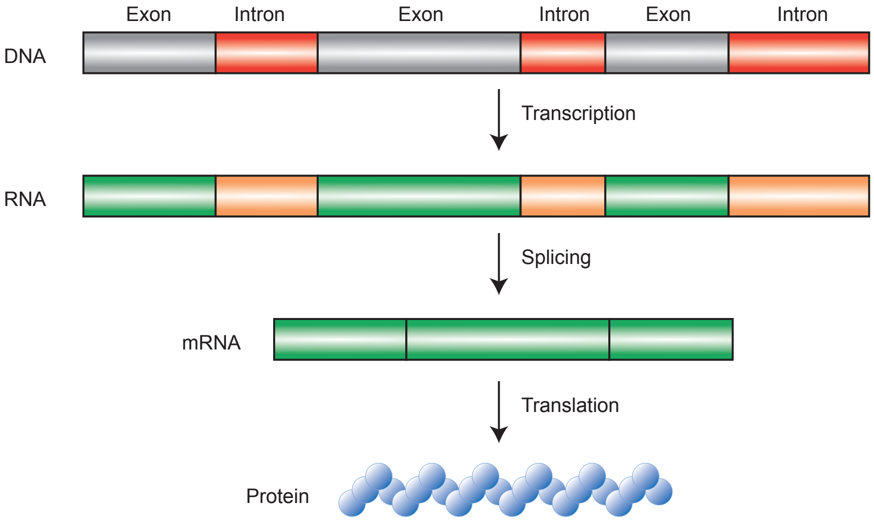

Figure A.3: The fundamental stages of protein biosynthesis

Complementary DNA (cDNA) is double-stranded DNA synthesized from a mRNA template. The synthesis reaction is catalyzed by the reverse transcriptase enzyme which produces a single-stranded complementary DNA chain on an mRNA template. The single-stranded molecule is converted into double-stranded cDNA by DNA polymerase.

Complementary RNA (cRNA) is synthetic RNA produced from a DNA molecule during an *in vitro* transcription reaction.

A **gene regulatory network (GRN)** represents a set of genes interacting with each other through transcription (RNAss) and translation (proteins) products to control a specific cell function. An example of a GRN is shown in Figure A.4.

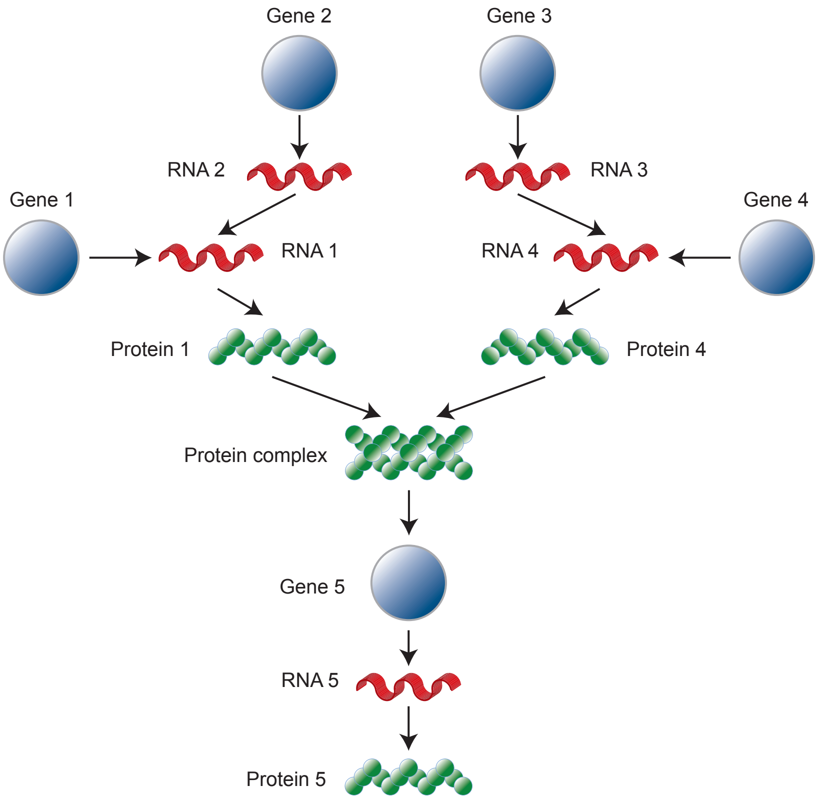

Figure A.4: Schematic representation of a GRN. Gene1 and Gene4 jointly regulate the expression level (abundance of RNA and protein) of Gene5 through the protein complex assembled for their individual translation products (Protein1 and Protein4). The amount of Protein1 is regulated by Gene2 through RNA2 that binds to molecules of RNA1, preventing further protein translation.

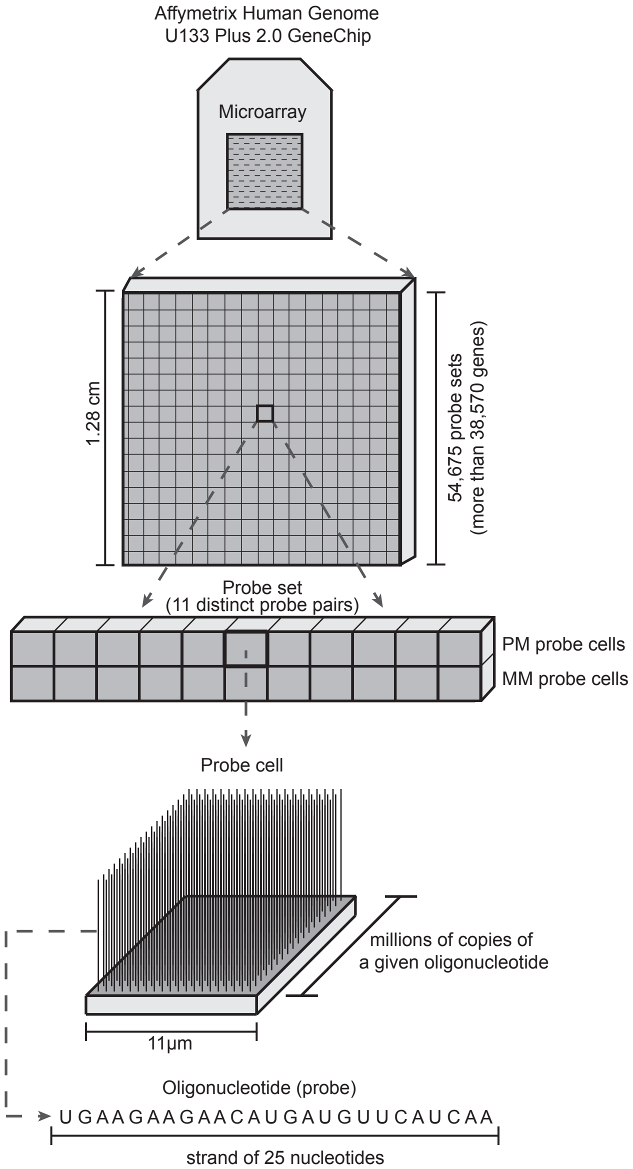

Figure A.5: Affymetrix GeneChip design

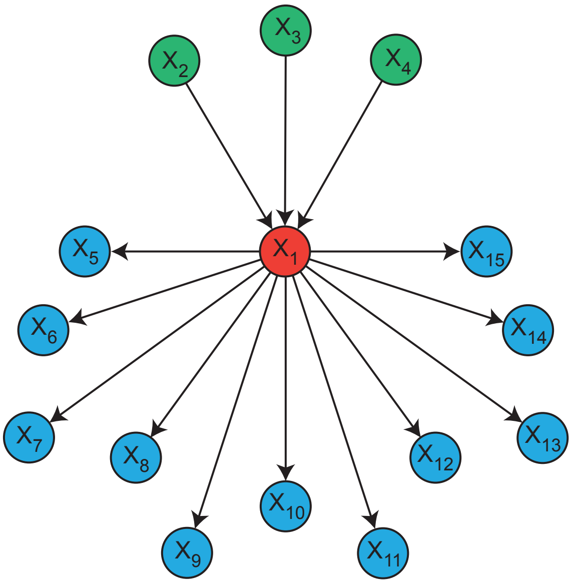

Figure A.6: Example of a gene (X_1) with three inward connections (in-degree) and eleven outward connections (out-degree). X_1 is regulated by X_2, X_3 and X_4 and regulates the expression level of genes $X_5 - X_{15}$.

Appendix B

Feature selection for ACS classification

Table B.1: Sample of estimated gene expression levels

Microarray	Probe set								
	235295_at	240274_at	1562969_at	240688_at	225316_at	1560410_at	205778_at	220615_s_at	243101_x_at
$^{1}NSTEMI_1$	6.559	4.846	2.962	-0.955	3.948	0.203	1.731	8.014	1.517
$^{7}NSTEMI_2$	6.504	4.948	0.715	-0.083	3.261	0.540	0.839	8.410	3.765
$^{90}NSTEMI_3$	6.499	3.682	2.373	-0.355	4.758	-0.039	2.233	8.216	3.790
$^{30}NSTEMI_4$	6.432	4.906	1.368	-0.755	3.120	-0.701	1.735	7.359	1.270
$^{3}NSTEMI_5$	5.651	4.642	0.476	1.880	3.580	-0.281	1.999	8.020	3.943
$^{90}NSTEMI_6$	5.705	5.666	-0.886	-1.458	3.136	-1.464	2.940	7.324	4.113
$^{1}NSTEMI_7$	6.503	4.990	2.423	0.539	3.830	-0.543	3.562	7.911	3.315
$^{30}NSTEMI_8$	5.963	4.225	2.922	-1.004	2.653	0.723	2.681	7.480	3.070
$^{3}STEMI_1$	6.408	4.593	-0.632	1.743	4.880	-0.707	2.433	7.637	2.677
$^{7}STEMI_2$	5.586	4.680	1.713	2.055	3.257	1.274	2.492	8.016	1.288
$^{30}STEMI_3$	5.848	4.640	0.761	-0.414	3.373	-0.814	2.865	7.201	2.753
$^{1}STEMI_4$	5.291	5.109	-0.533	2.038	3.710	0.154	2.046	7.462	1.270
$^{90}STEMI_5$	6.545	4.831	2.553	0.089	5.150	-0.213	1.916	7.265	0.145
$^{90}UA_1$	6.041	3.490	0.146	0.268	4.707	-0.775	0.900	8.551	-0.673
$^{3}UA_2$	5.770	4.594	1.493	2.439	1.813	-0.303	1.843	7.676	2.962
$^{7}UA_3$	5.706	4.159	1.523	3.360	3.744	0.519	2.134	8.202	0.664
$^{1}UA_4$	6.644	4.331	1.874	-1.820	4.668	-0.050	0.023	8.510	2.815
$^{30}UA_5$	5.525	3.772	1.497	-0.174	4.609	-0.135	2.175	7.862	2.056
$^{3}UA_6$	6.578	4.881	3.305	-0.491	4.760	0.383	2.481	7.911	3.491
$^{90}UA_7$	6.783	4.302	1.476	-1.529	3.607	0.040	2.988	7.334	0.088

$^{j}NSTEMI_i$ denotes the microarray obtained from the peripheral blood sample collected from the ith NSTEMI patient at day j after hospital admission. Negative values indicate absence of gene expression.

Table B.2: Standard errors of the estimated gene expression levels listed in Table B.1

Microarray	Probe set								
	235295_at	240274_at	1562969_at	240688_at	225316_at	1560410_at	205778_at	220615_s_at	243101_x_at
${}^{1}NSTEMI_1$	0.125	0.410	0.907	1.513	0.957	1.723	1.416	0.196	1.116
${}^{7}NSTEMI_2$	0.173	0.500	1.827	1.433	1.329	1.788	1.899	0.215	0.596
${}^{90}NSTEMI_3$	0.160	0.833	1.226	1.459	0.806	1.929	1.384	0.218	0.559
${}^{30}NSTEMI_4$	0.139	0.405	1.441	1.478	1.231	2.036	1.436	0.269	1.231
${}^{3}NSTEMI_5$	0.210	0.474	1.757	0.734	1.090	1.910	1.354	0.209	0.441
${}^{90}NSTEMI_6$	0.233	0.298	2.303	1.754	1.286	2.379	1.127	0.317	0.442
${}^{1}NSTEMI_7$	0.160	0.449	1.205	1.205	1.085	2.100	0.931	0.246	0.690
${}^{30}NSTEMI_8$	0.155	0.524	0.852	1.462	1.300	1.491	1.045	0.222	0.601
${}^{3}STEMI_1$	0.122	0.429	2.030	0.698	0.601	1.953	1.125	0.210	0.727
${}^{7}STEMI_2$	0.169	0.367	1.180	0.538	1.058	1.261	1.057	0.167	1.093
${}^{30}STEMI_3$	0.163	0.416	1.556	1.301	1.073	1.985	1.002	0.259	0.704
${}^{1}STEMI_4$	0.200	0.277	1.920	0.524	0.885	1.605	1.196	0.210	1.088
${}^{90}STEMI_5$	0.128	0.426	1.050	1.221	0.576	1.861	1.375	0.281	1.552
${}^{90}UA_1$)	0.201	0.870	1.967	1.304	0.795	2.167	1.794	0.174	1.920
${}^{3}UA_2$	0.222	0.526	1.479	0.614	1.726	1.968	1.451	0.260	0.758
${}^{7}UA_3$	0.301	0.772	1.636	0.5.01	1.223	1.871	1.499	0.240	1.617
${}^{1}UA_4$	0.114	0.506	1.219	1.736	0.684	1.762	1.951	0.151	0.695
${}^{30}UA_5$	0.214	0.684	1.373	1.293	0.739	1.820	1.269	0.203	0.981
${}^{3}UA_6$	0.125	0.400	0.803	1.384	0.685	1.668	1.185	0.198	0.555
${}^{90}UA_7$	0.152	0.707	1.592	1.855	1.222	1.956	1.178	0.347	1.732

${}^{j}NSTEMI_i$ denotes the microarray obtained from the peripheral blood sample collected from the ith NSTEMI patient at day j after hospital admission.

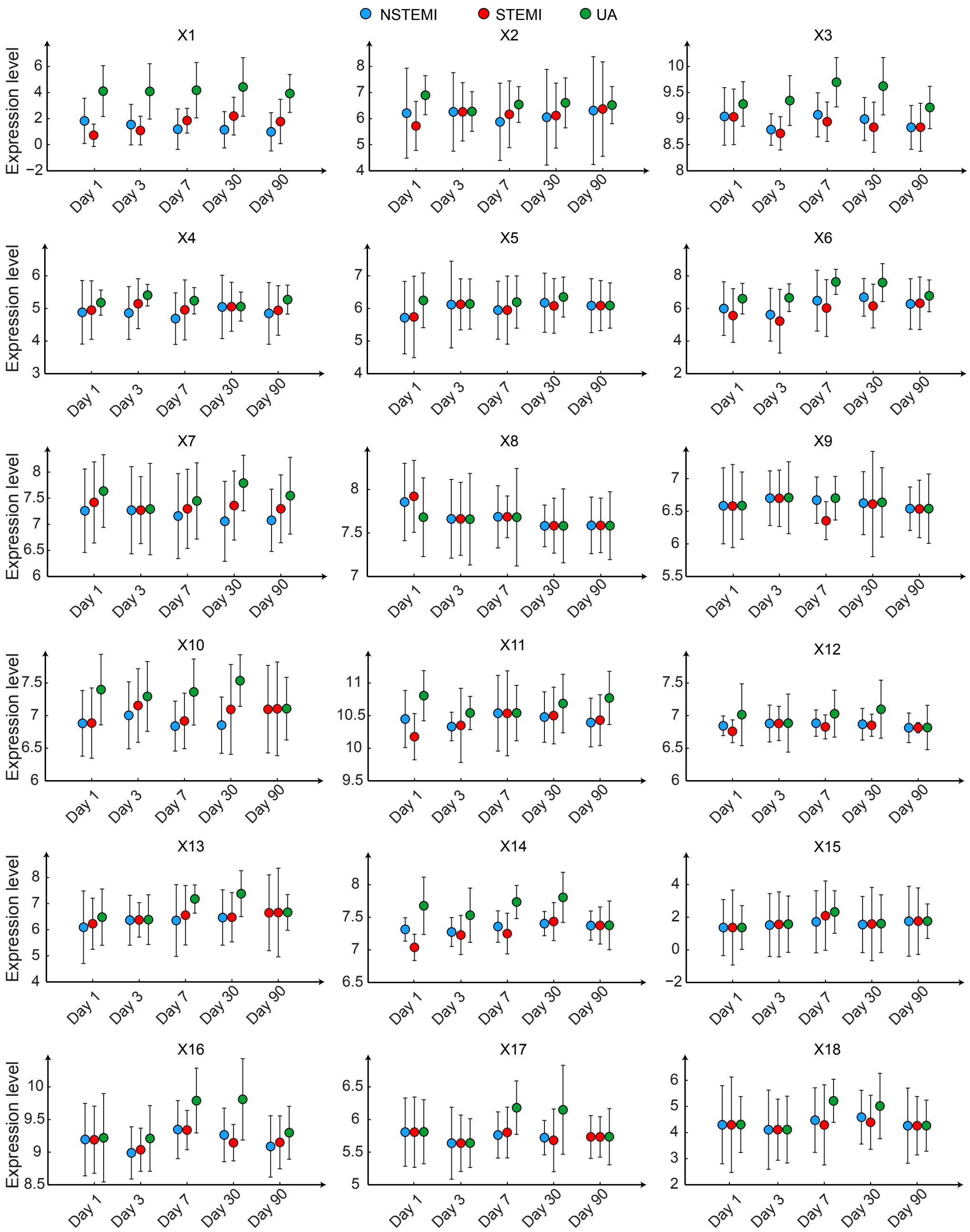

Figure B.1: Group-specific expression averages across visits for genes $X_1 - X_{18}$ in the MI vs. UA study.

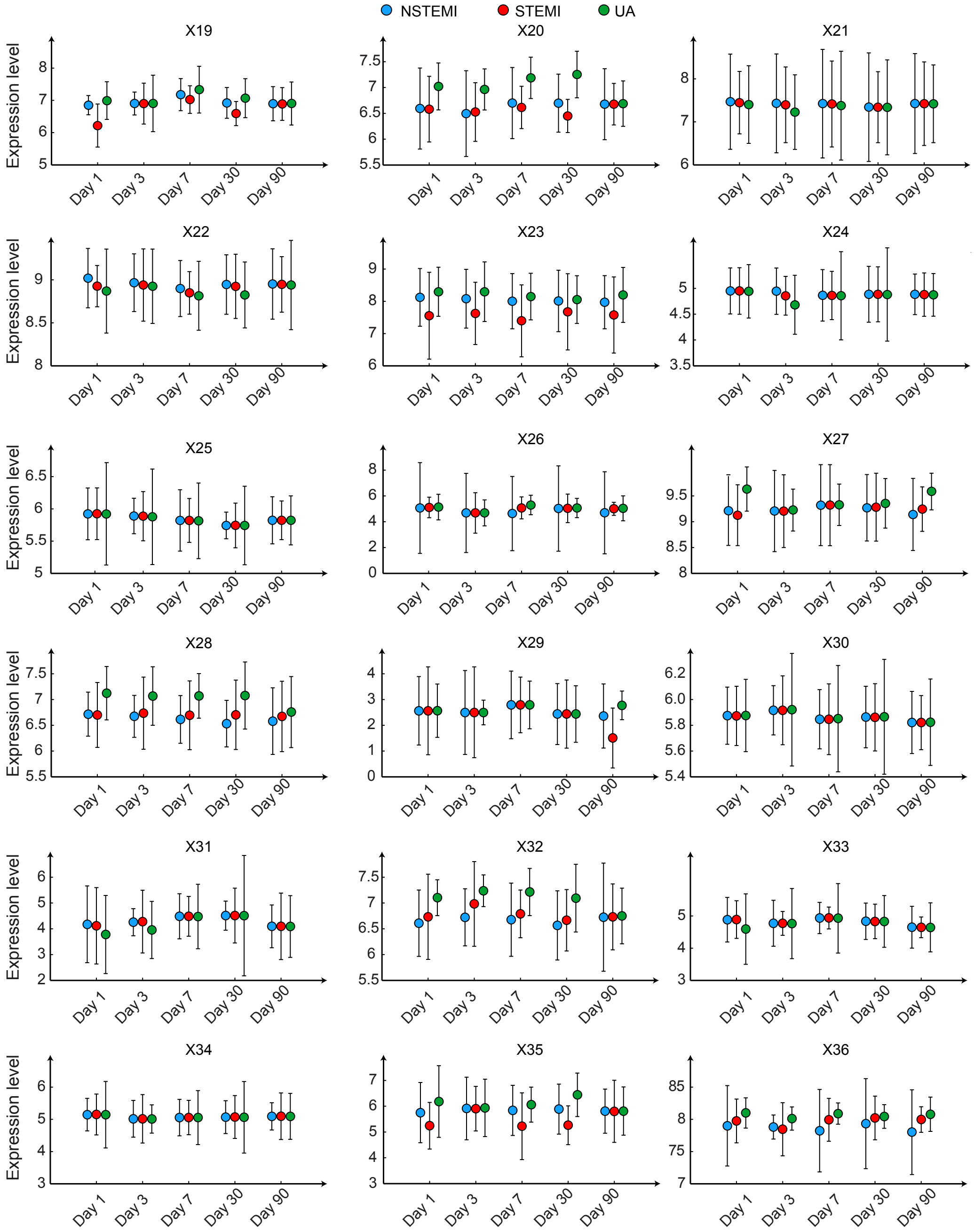

Figure B.2: Group-specific expression averages across visits for genes $X_{19} - X_{36}$ in the MI vs. UA study.

Table B.3: Differentially expressed genes between the MI and UA groups selected by the multi-stage feature selection method

Symbol	Probe set	Gene	Hedges′ g score	Frequency
X_1	1557034_s_at	WASH1	-1.8603	9
X_2	209581_at	PLA2G16	-1.1749	9
X_3	226657_at	C17orf103	-1.4369	5
X_4	232600_at	—	-0.9971	3
X_5	1569608_x_at	—	-0.9009	2
X_6	221237_s_at	OSBP2	-1.1704	2
X_7	230026_at	MRPL43	-0.9000	2
X_8	236321_at	FAM200B	0.6576	2
X_9	1557477_at	—	-0.6012	1
X_{10}	1569110_x_at	LOC728613	-1.1157	1
X_{11}	201649_at	UBE2L6	-0.9671	1
X_{12}	202405_at	TIAL1	-0.8356	1
X_{13}	203609_s_at	ALDH5A1	-0.9475	1
X_{14}	204788_s_at	PPOX	-1.2844	1
X_{15}	206136_at	FZD5	-0.8422	1
X_{16}	209019_s_at	PINK1	-1.0190	1
X_{17}	209416_s_at	FZR1	-1.2603	1
X_{18}	210151_s_at	DYRK3	-0.7842	1
X_{19}	210299_s_at	FHL1	-0.5827	1
X_{20}	219905_at	ERMAP	-1.3273	1
X_{21}	221634_at	RPL23AP7	0.6955	1
X_{22}	222791_at	RSBN1	0.7035	1
X_{23}	226558_at	LOC389834	-0.6496	1
X_{24}	226974_at	NEDD4L	0.9978	1
X_{25}	227721_at	CPAMB8	0.9759	1
X_{26}	228425_at	LOC654433	-0.7542	1
X_{27}	229390_at	FAM26F	-0.9052	1
X_{28}	229449_at	—	-1.0798	1
X_{29}	231504_at	CCDC148	-0.7238	1
X_{30}	232362_at	CCDC18	-0.5421	1
X_{31}	235758_at	LOC100287428	1.0313	1
X_{32}	235761_at	—	-1.1196	1
X_{33}	236089_at	—	0.8346	1
X_{34}	236837_x_at	LOC650794	0.6433	1
X_{35}	241233_x_at	C21orf81	-0.8792	1
X_{36}	56919_at	WDR48	-0.8420	1

Table B.4: Optimal classifier parameters for the MI vs. UA study

Fold	σ	C
1	100	10
2	1000	100
3	1000	10
4	1000	31.62
5	100	10
6	316.22	10
7	1000	10
8	31.62	10
9	316.22	100
10	100	31.62

Table B.5: Differentially expressed genes between the MI and UA groups selected by l_1-StaR

Symbol	Probe set	Gene	Hedges' g score	Frequency
X_1^*	1557034_s_at	WASH1	-1.8603	10
X_2^*	1569955_at	—	-0.5798	10
X_3^*	219300_s_at	CNTNAP2	-0.7719	10
X_4^*	224490_s_at	KIAA1267	0.5763	10
X_5^*	239853_at	KLC3	-1.0778	10
X_6^*	AFFX-LysX-M_at	—	0.5967	10
X_7^*	203638_s_at	FGFR2	-0.7399	9
X_8^*	230026_at	PLA2G16	-1.1749	7
X_9^*	1561754_at	—	-0.4367	6
X_{10}^*	213831_at	HLA-DQA1	-0.5162	6
X_{11}^*	224489_at	KIAA1267	0.4735	5
X_{12}^*	1569481_s_at	SNX22	-1.0031	4
X_{13}^*	227474_at	LOC654433	-0.6473	4
X_{14}^*	230959_at	—	-0.5009	3
X_{15}^*	207766_at	CDKL1	-1.2877	2
X_{16}^*	212768_s_at	OLFM4	-0.6197	2
X_{17}^*	216775_at	USP53	0.4529	2
X_{18}^*	224005_at	—	-0.6402	2
X_{19}^*	230053_at	—	-0.8748	2
X_{20}^*	1559477_s_at	MEIS1	-0.8264	1
X_{21}^*	203911_at	RAP1GAP	-0.8179	1
X_{22}^*	206700_s_at	KDM5D	0.5330	1
X_{23}^*	211430_s_at	IGHG1	0.4167	1
X_{24}^*	213547_at	CAND2	0.6672	1
X_{25}^*	21736_x_at	IGHA1	0.4481	1
X_{26}^*	220004_at	DDX43	-0.4885	1
X_{27}^*	228362_at	FAM26F	-0.8267	1
X_{28}^*	230336_at	—	-0.7275	1
X_{29}^*	231996_at	N4BP2	-0.7597	1
X_{30}^*	233823_at	FAM184B	0.4357	1
X_{31}^*	236962_at	—	-0.6493	1
X_{32}^*	236988_x_at	ITGB2	0.3944	1
X_{33}^*	237056_at	INSC	0.3939	1
X_{34}^*	243106_at	CLEC12A	-0.6286	1

Table B.6: Differentially expressed probe sets between the NSTEMI and STEMI groups selected by the multi-stage method

Symbol	Probe set	Gene	Hedges' g score	Frequency
X_1	209823_x_at	HLA-DQB1	-1.7879	8
X_2	213013_at	MAPK8IP1	-1.4907	5
X_3	213510_x_at	LOC220594	-0.8613	4
X_4	229778_at	C12orf39	1.4668	4
X_5	203695_s_at	DFNA5	-1.4510	3
X_6	212843_at	NCAM1	1.3154	3
X_7	227394_at	NCAM1	1.3752	2
X_8	230388_s_at	LOC64246	-1.1345	2
X_9	242874_at	—	0.9873	2
X_{10}	1552398_a_at	CLEC12A	-0.8403	1
X_{11}	1557293_at	LOC440993	1.3141	1
X_{12}	1560071_a_at	—	1.1301	1
X_{13}	203780_at	MPZL2	-0.9187	1
X_{14}	205934_at	PLCL1	1.1442	1
X_{15}	211430_s_at	IGHG1	-1.0728	1
X_{16}	212220_at	PSME4	1.1986	1
X_{17}	215761_at	DMXL2	1.0849	1
X_{18}	227421_at	C21orf57	0.9055	1
X_{19}	228518_at	IGH1/IGHM	-0.8148	1
X_{20}	231858_x_at	DKFZp761E198	0.9116	1
X_{21}	239591_at	LRRC37A	-1.1536	1

Table B.7: Optimal classifier parameters for the NSTEMI vs. STEMI study

Fold	σ	C
1	31.62	31.62
2	31.62	10
3	1000	100
4	316.22	31.62
5	100	31.62
6	10	10
7	10	31.62
8	10	10
9	3.16	10
10	10	10

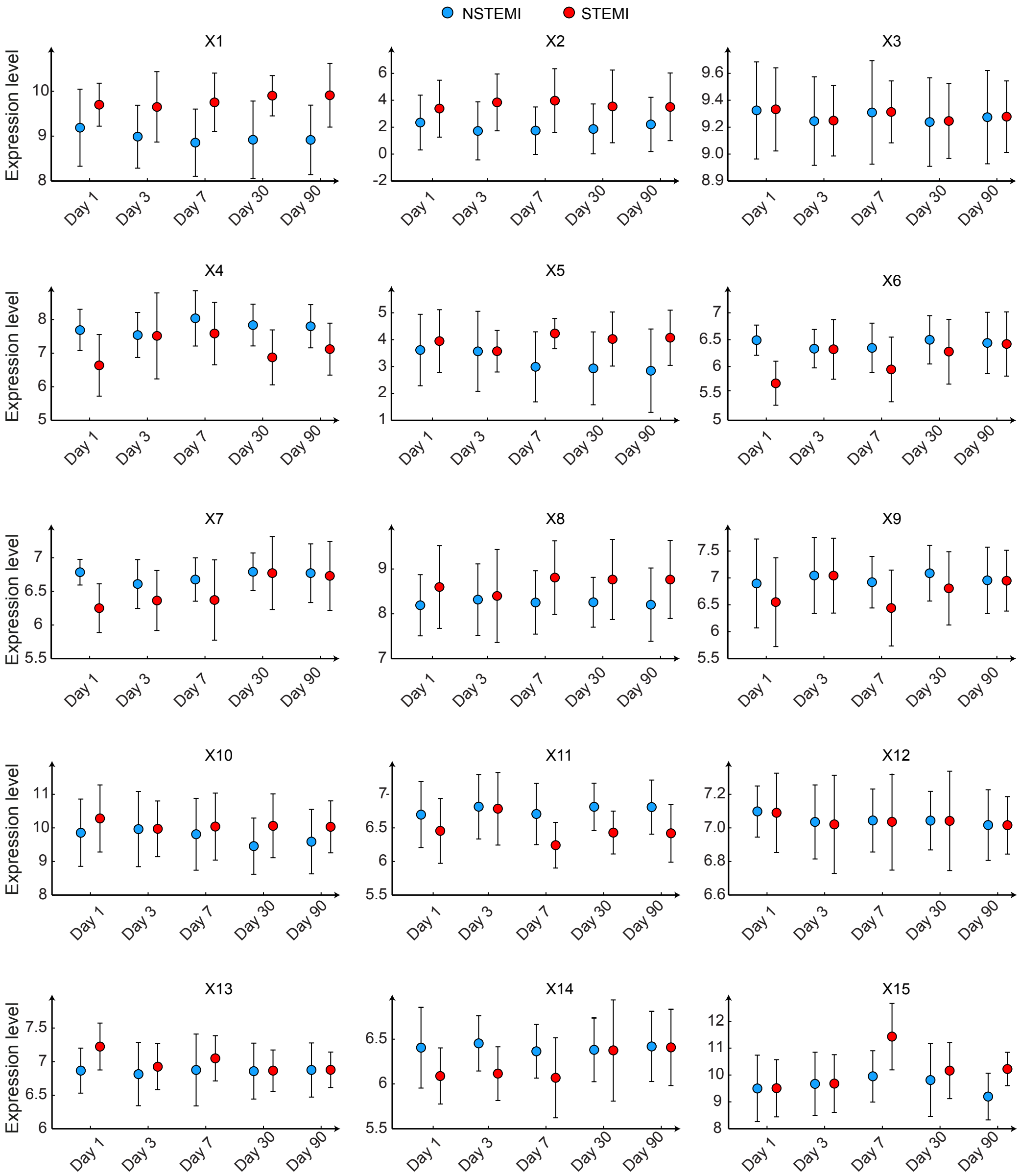

Figure B.3: Group-specific expression averages across visits for genes $X_1 - X_{15}$ in the NSTEMI vs. STEMI study.

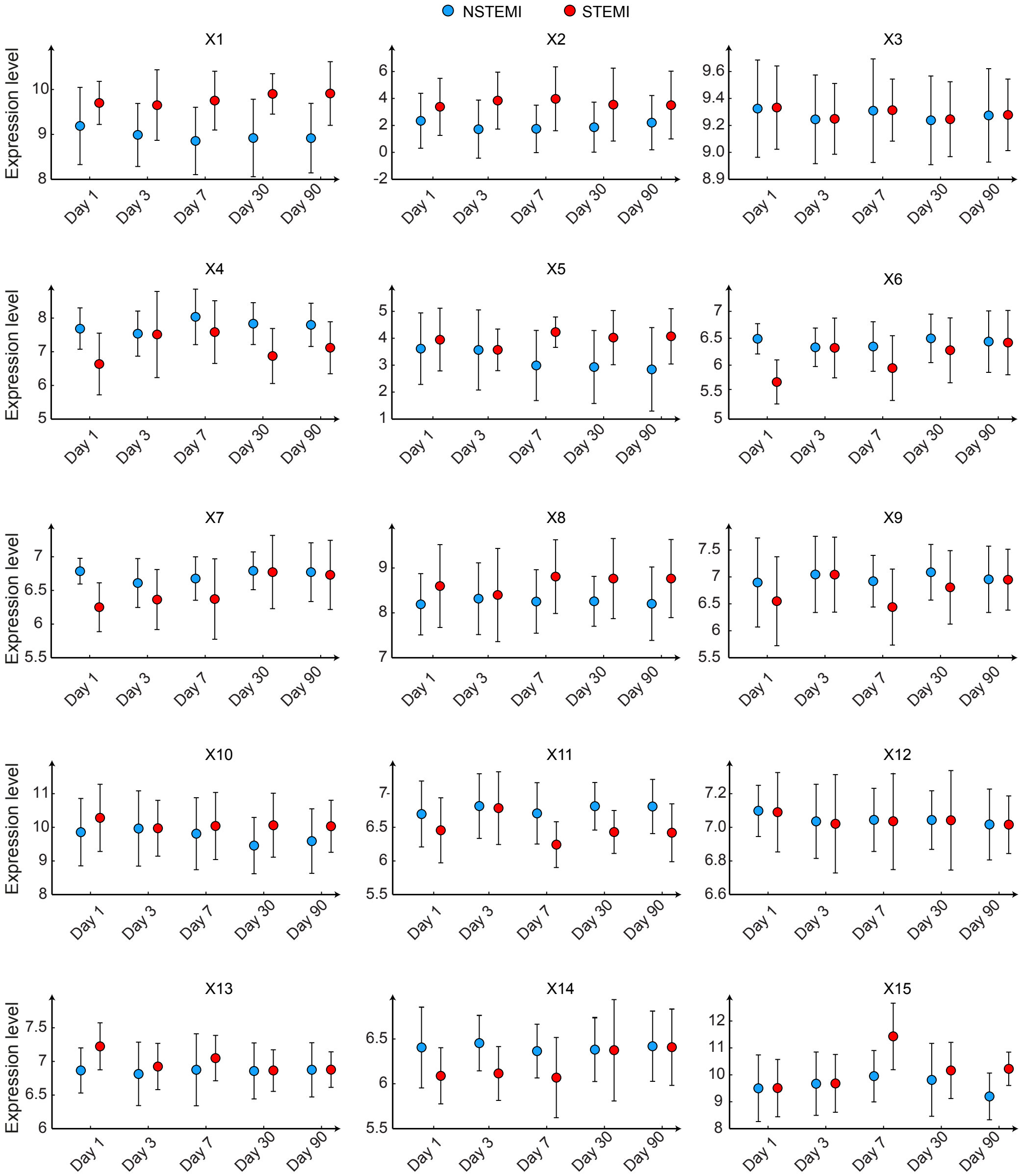

Figure B.4: Group-specific expression averages across visits for genes $X_{16} - X_{21}$ in the NSTEMI vs. STEMI study.

Table B.8: Differentially expressed genes between the NSTEMI and STEMI groups selected by l_1-StaR

Symbol	Gene	Probe set	Hedges' g score	Frequency
X_1^*	APOBEC3B	206632_s_at	-1.2503	10
X_2^*	MAPK8IP1	213013_at	-1.4907	10
X_3^*	BTNL8	220421_at	1.2567	10
X_4^*	NEBL	203961_at	0.5542	5
X_5^*	XPNPEP2	216910_at	0.5132	4
X_6^*	—	227952_at	0.8997	3
X_7^*	—	239591_at	-1.1536	3
X_8^*	NEBL	203962_s_at	0.4581	1
X_9^*	S100B	209686_at	0.4825	1
X_{10}^*	SDC2	212158_at	0.6940	1
X_{11}^*	RNF213	231959_at	0.6068	1

Appendix C

Deconvolution of microarray gene expression data

Table C.1: Cell type-specific contributions to the variance of the genes $X_1 - X_{18}$ in the MI group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_1	0.010	0	0	0	0.459	0
X_2	0	0.001	0	0	0.690	0
X_3	0	0	0	0	0	0.907
x_4	0	0	8e-4	0	0.756	0
X_5	0.003	0	0	0	0.761	0
X_6	0	0	0	0	0	0.658
X_7	1e-4	0.014	0.753	0	0.007	0
X_8	0	8e-4	0	0.020	0	0.921
X_9	0.005	0	0	0	0.911	0
X_{10}	1e-4	0	0	0	0.881	0
X_{11}	0	0	0	0.001	0.010	0.865
X_{12}	4e-4	0	0	0	0.976	0.002
X_{13}	0	0	0	0	0	0.683
X_{14}	0	0	0	0	0.958	0.004
X_{15}	0	0	0	0	0	0.230
X_{16}	0	0	0	0	0.003	0.919
X_{17}	0	0	0	0	0.947	0.005
X_{18}	0	0	0	0	0	0.669

1e-4 = 10^{-4}

Table C.2: Cell type-specific contributions to the variance of the genes $X_{19} - X_{36}$ in the MI group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_{19}	0	0	0	0	0	0.901
X_{20}	0	4e-4	0	0	0.877	0.007
X_{21}	0	0.020	0.004	0	0	0.657
X_{22}	0	0	8e-4	0	0.949	0
X_{23}	0	0	0	0	0.688	0
X_{24}	0	0.015	9e-4	0	0.900	0.005
X_{25}	0	0.009	0	0	0.933	0
X_{26}	0	0	0	0	0.547	0
X_{27}	2e-4	0	0	0	0	0.839
X_{28}	0	0	0	0.004	0.906	0
X_{29}	0	0.002	0	0	0	0.701
X_{30}	8e-4	0	9e-5	0	0.967	0.002
X_{31}	9e-4	0.002	0	0	0.768	0
X_{32}	0	4e-4	0	0	0.872	0.004
X_{33}	0	0	0	0.868	0	0.015
X_{34}	0	0.015	0	0	0.800	0
X_{35}	0.027	0	0	0	0.675	0
X_{36}	0	0.002	6e-4	0	0.932	0

1e-4 = 10^{-4}

Table C.3: Cell type-specific contributions to the variance of the genes $X_1 - X_{18}$ in the UA group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_1	0	0	0	0.436	0	0
X_2	0.008	0.002	0.005	0	0.818	0
X_3	0	0	0.026	0.886	0	0
X_4	0	0.012	0.001	0	0.945	0
X_5	0.015	0	0	0	0.823	0
X_6	0	0	0.013	0	0	0.732
X_7	0	0	0.003	0	0.889	0
X_8	0	0	0	0.905	0	0.006
X_9	0.002	0	0.028	0.002	0	0.892
X_{10}	0.006	0	0.019	0	0.917	0
X_{11}	0	2e-5	0.004	0	0.001	0.905
X_{12}	0	0	0.008	0	0	0.901
X_{13}	0	0	0.009	0	0	0.764
X_{14}	0	0	0.900	0	0	0.022
X_{15}	0	0	0	0.455	0	0
X_{16}	0	0	0	0.848	0	0
X_{17}	2e-4	0	0.002	0	0	0.887
X_{18}	0	0	0	0.6968	0	0

1e-4 = 10^{-4}

Table C.4: Cell type-specific contributions to the variance of the genes $X_{19} - X_{36}$ in the UA group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_{19}	0	0	0	0	0	0.8386
X_{20}	0	2e-6	0.001	0.003	0	0.888
X_{21}	0	0.737	0.020	0	0	0
X_{22}	0	0	0.008	0	0	0.904
X_{23}	0	0.0190	0.854	0	0	0
X_{24}	0.002	0	0	0	0.8986	0
X_{25}	7e-4	0.010	0	0	0.903	6e-4
X_{26}	0	0	0	0	0	0.822
X_{27}	0	0.007	0.019	0	0	0.880
X_{28}	0.006	0.001	0	0	0.904	0
X_{29}	6e-4	0.014	0.788	0.002	0	0
X_{30}	2e-4	0	0.007	0	0.956	0
X_{31}	0.028	0	0	0	0.685	0
X_{32}	0.010	0.001	0	1e-6	0.925	0
X_{33}	0	0	0	0.764	0	0
X_{34}	0	0	0	0.616	0	0.004
X_{35}	0	0	0.758	0	0	0
X_{36}	0	0	7e-4	0	0.975	8e-5

1e-4 = 10^{-4}

Table C.5: Cell type-specific contributions to the variance of the genes $X_1 - X_{21}$ in the NSTEMI group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_1	0	0	1e-05	0	0.820	0
X_2	0	0.249	0	0	0	0
X_3	0	3e-5	6e-4	0.012	0.941	0.003
X_4	0	8e-7	0	0.002	0	0.866
X_5	9e-5	0.002	0	0	0.686	0
X_6	0	0.002	0	0	0	0.866
X_7	0	0	0	0	0.952	0
X_8	0	0	0	0	0.818	0
X_9	0	0	0.011	0	0.887	0
X_{10}	0	0.010	0	0.658	0	0
X_{11}	0.004	0	7e-4	0	0.945	0
X_{12}	0	3e-4	0	0.002	0.977	0.006
X_{13}	0	0.013	0	0	0	0.899
X_{14}	1e-4	0	0.0064	0	0.946	0.004
X_{15}	0	0	0	0	0	0.53
X_{16}	3e-4	0	0	0	0.785	0
X_{17}	0	0.012	0	0	0.910	0.003
X_{18}	0	0	0	0	0.91	0
X_{19}	0	0.009	0.609	0	0	0
X_{20}	0	0	0	0	0	0.824
X_{21}	0	0.319	0	0	0	0

1e-4 = 10^{-4}

Table C.6: Cell type-specific contributions to the variance of the genes $X_1 - X_{21}$ in the STEMI group

Gene	Explained variance					
	Eosinophils	Monocytes	Lymphocytes	Neutrophils	Erythrocytes	Platelets
X_1	0.004	0	0	0	0.87	0
X_2	0	0	0	0.002	0.667	0
X_3	9e-4	0	0.010	0.001	0.941	0.003
X_4	0	0	0	0	0	0.828
X_5	0	0	0	0.721	0	0
X_6	0	0	0	0	0.891	0
X_7	0	0	0	0	0.915	0
X_8	0	0	0	0.823	0	0
X_9	0.013	0	6e-4	0	0.861	0
X_{10}	0	0	0	0.860	0	0
X_{11}	6e-4	0	0.010	0	0.933	0
X_{12}	0	0	0	0.001	0.976	0
X_{13}	0.004	1e-4	1e-5	0.008	0.933	6e-4
X_{14}	0.002	0	0	0.001	0.912	0.003
X_{15}	0	0	0	0	0	0.641
X_{16}	0.696	0	0	0	0.018	0
X_{17}	0	0	0.001	0.008	0.872	0.002
X_{18}	0	0	0	0	0.962	0
X_{19}	0	0.683	0	0	0	0
X_{20}	0	3e-4	0.003	0	0.900	0.005
X_{21}	0.021	0	0	0.676	0	0

1e-4 = 10^{-4}

Table C.7: Amount of genes expressed in each cell type in the MI and UA groups

Cell type	MI	UA
Eosinophils	10	8
Monocytes	12	3
Lymphocytes	7	6
Neutrophils	4	10
Erythrocytes	12	14
Platelets	9	7

Table C.8: Amount of genes expressed in each cell type in the NSTEMI and STEMI groups

Cell type	NSTEMI	STEMI
Eosinophils	4	12
Monocytes	11	11
Lymphocytes	6	21
Neutrophils	4	12
Erythrocytes	25	13
Platelets	20	16

Appendix D

Reconstruction of GRNs for the ACS subtypes

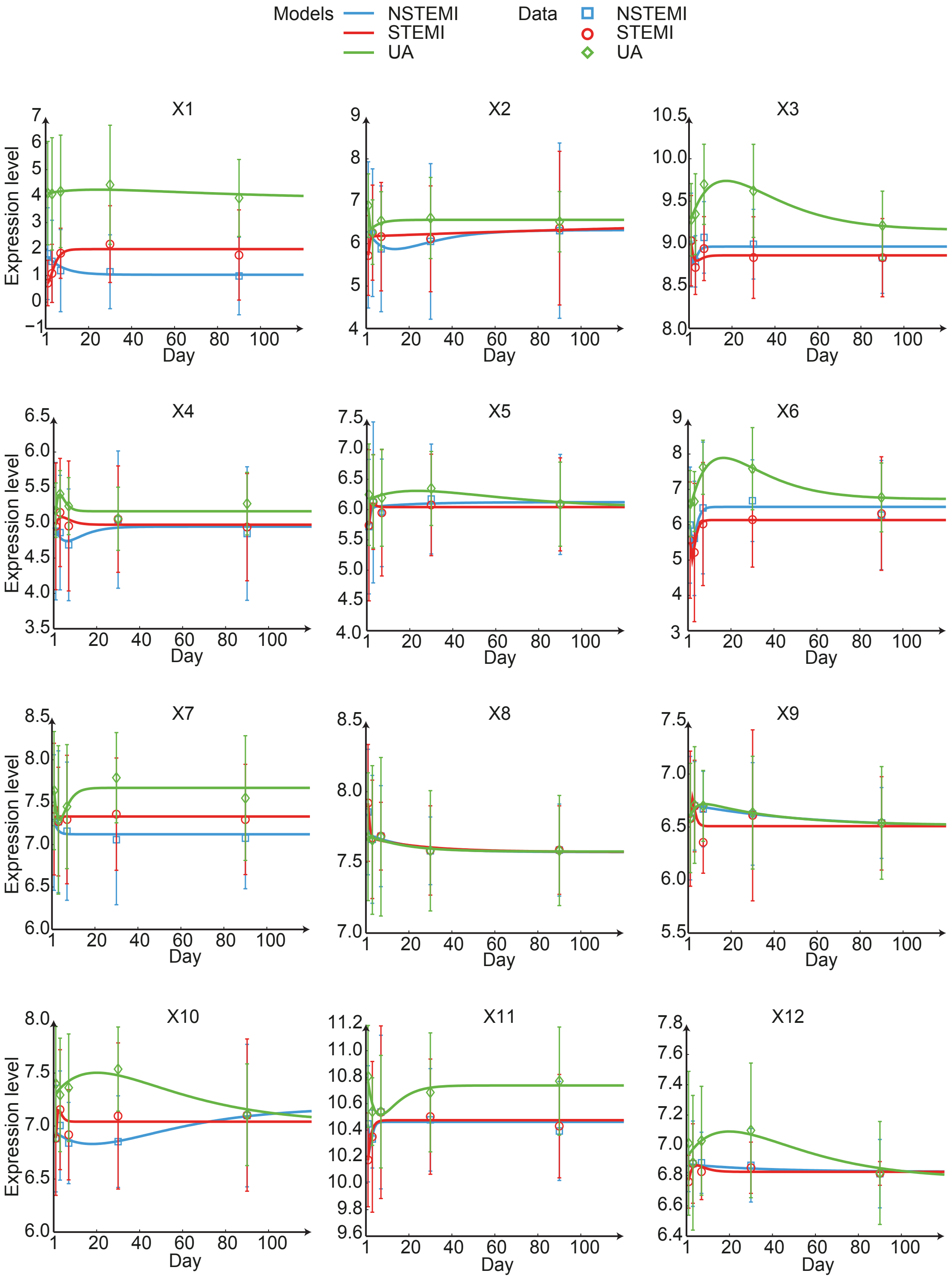

Figure D.1: Estimated dynamic profiles for genes $X_1 - X_{12}$ in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages

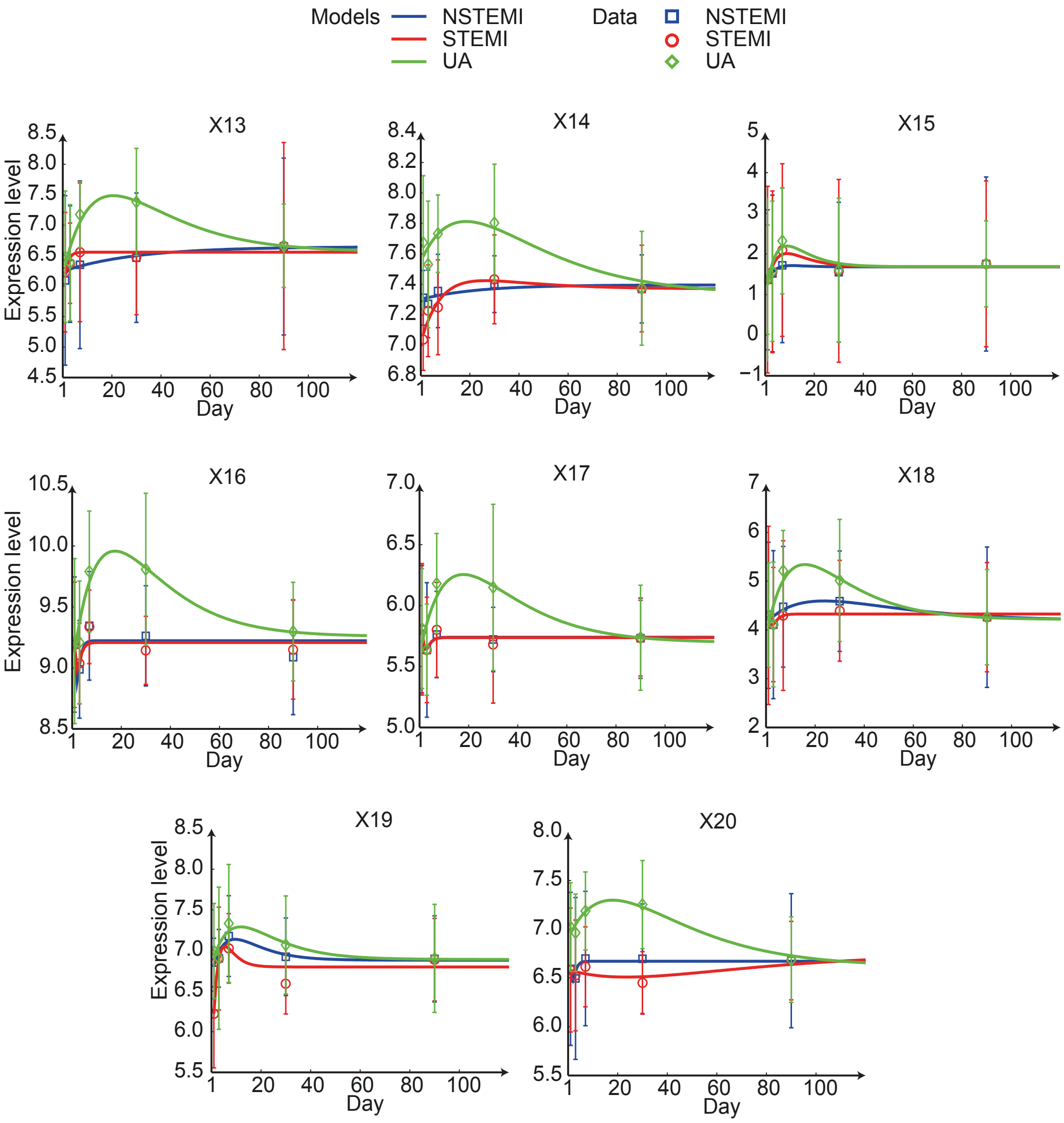

Figure D.2: Estimated dynamic profiles for genes $X_{13} - X_{20}$ in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages

Table D.1: Goodness of fit for the genes differentiating MI from UA

Gene	Residual sum of squares		
	NSTEMI	STEMI	UA
X_1	0.0502	0.0898	0.0574
X_2	0.0277	0.0240	0.0215
X_3	0.0330	0.0201	0.0162
X_4	0.0315	0.0214	0.0228
X_5	0.0277	0.0130	0.0156
X_6	0.0900	0.0708	0.1006
X_7	0.0189	0.0172	0.0289
X_8	0.0016	0.0021	0.0012
X_9	0.0008	0.0370	0.0016
X_{10}	0.0131	0.0237	0.0219
X_{11}	0.0115	0.0095	0.0077
X_{12}	0.0019	0.0027	0.0158
X_{13}	0.0186	0.0261	0.0972
X_{14}	0.0042	0.0068	0.0244
X_{15}	0.0246	0.0609	0.1104
X_{16}	0.0375	0.0267	0.0432
X_{17}	0.0015	0.0086	0.0722
X_{18}	0.0452	0.0086	0.2343
X_{19}	0.0113	0.0508	0.0432
X_{20}	0.0075	0.0122	0.0139
X_{21}	0.0066	0.0062	0.0068
X_{22}	0.0029	0.0035	0.0041
X_{23}	0.0148	0.0644	0.0429
X_{24}	0.0012	0.0016	0.0013
X_{25}	0.0001	0.0003	0.0003
X_{26}	0.0540	0.0227	0.0514
X_{27}	0.0059	0.0079	0.0058
X_{28}	0.0105	0.0114	0.0126
X_{29}	0.0255	0.0190	0.0690
X_{30}	0.0009	0.0009	0.0009
X_{31}	0.0055	0.0074	0.0114
X_{32}	0.0106	0.0160	0.0129
X_{33}	0.0121	0.0133	0.0025
X_{34}	0.0007	0.0007	0.0007
X_{35}	0.0166	0.3013	0.0902
X_{36}	0.0058	0.0046	0.0037

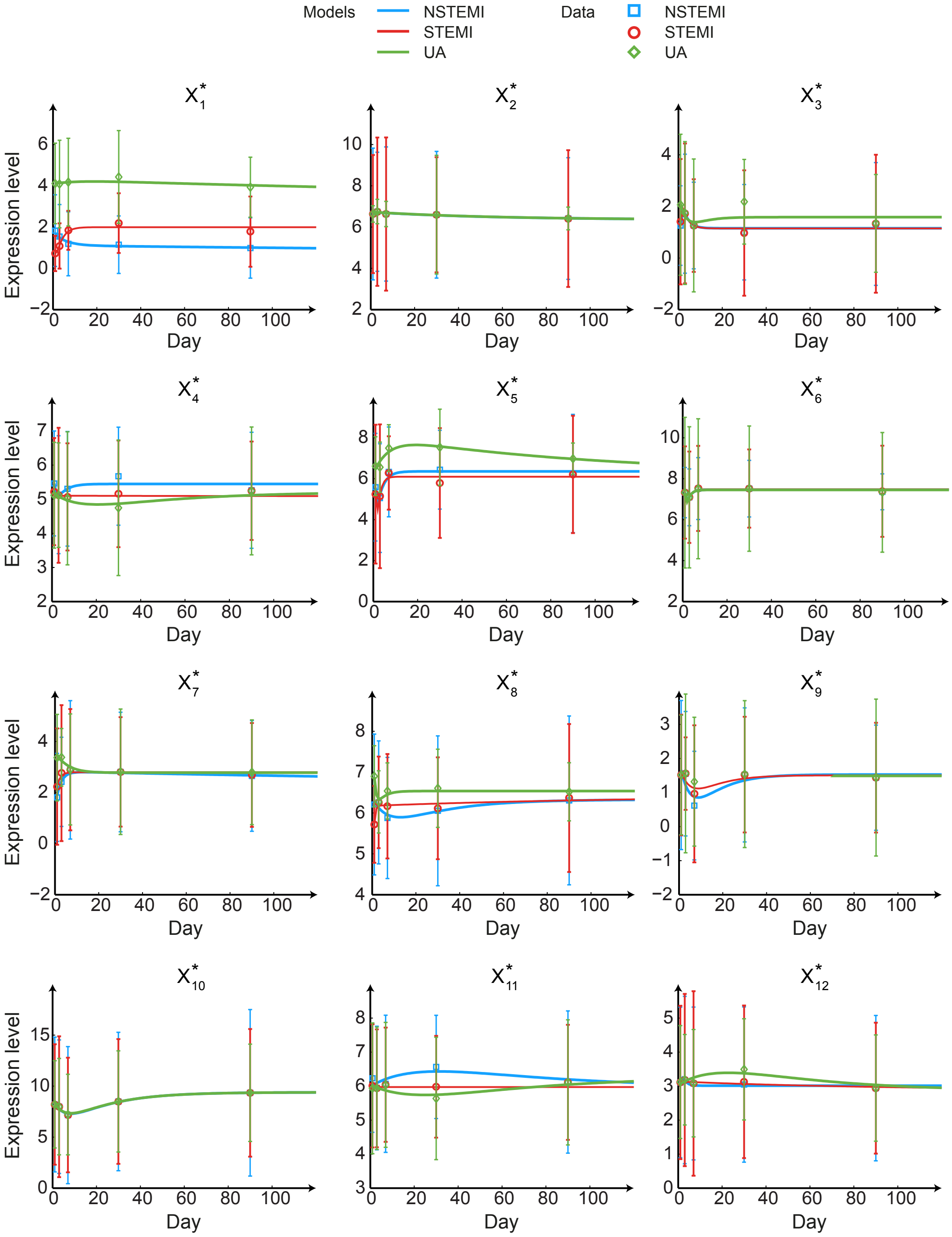

Figure D.3: Dynamic profiles of the combined expression levels for genes $X_1^* - X_{12}^*$ selected by l_1-StaR in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages

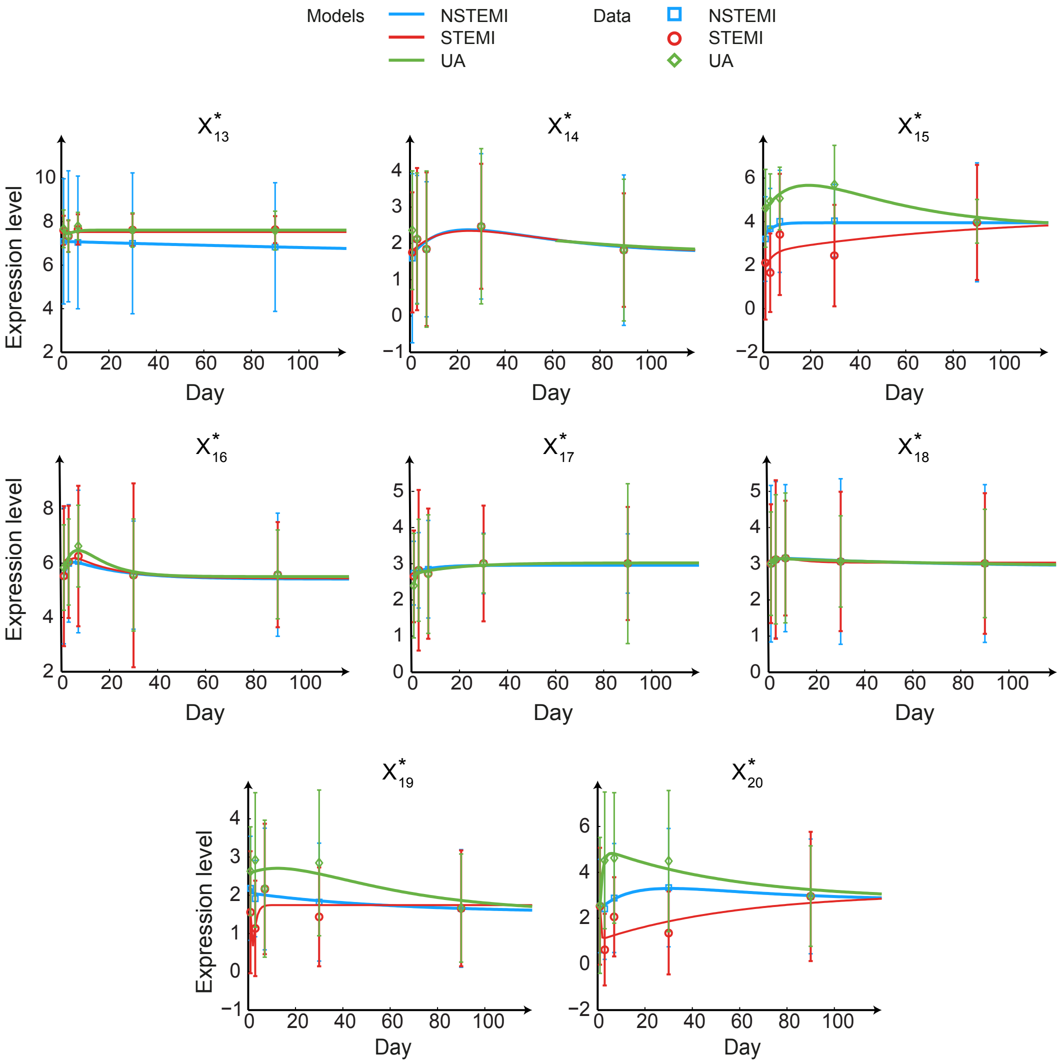

Figure D.4: Dynamic profiles of the combined expression levels for genes $X^*_{13} - X^*_{20}$ selected by l_1-StaR in the MI vs. UA study. The bars denote one standard deviation around the group-specific gene expression averages

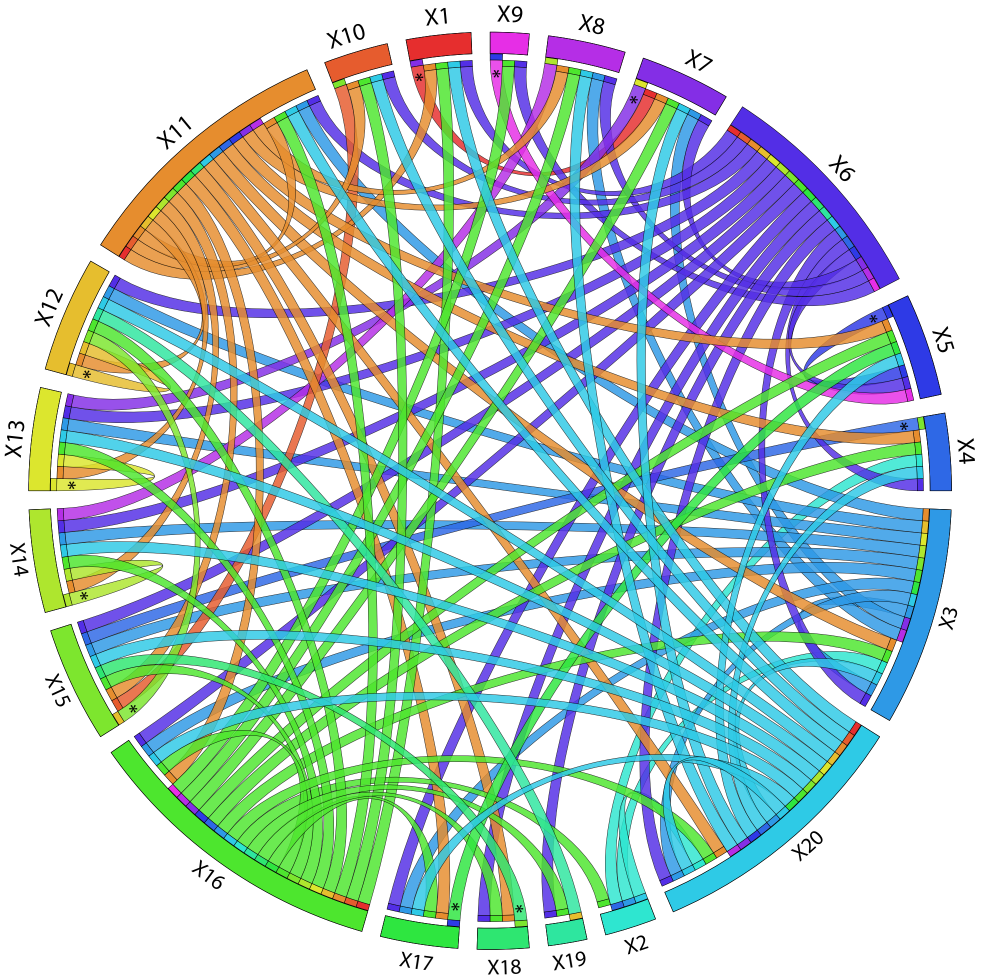

Figure D.5: Gene regulatory network for genes X_1-X_{20} in the NSTEMI group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene.

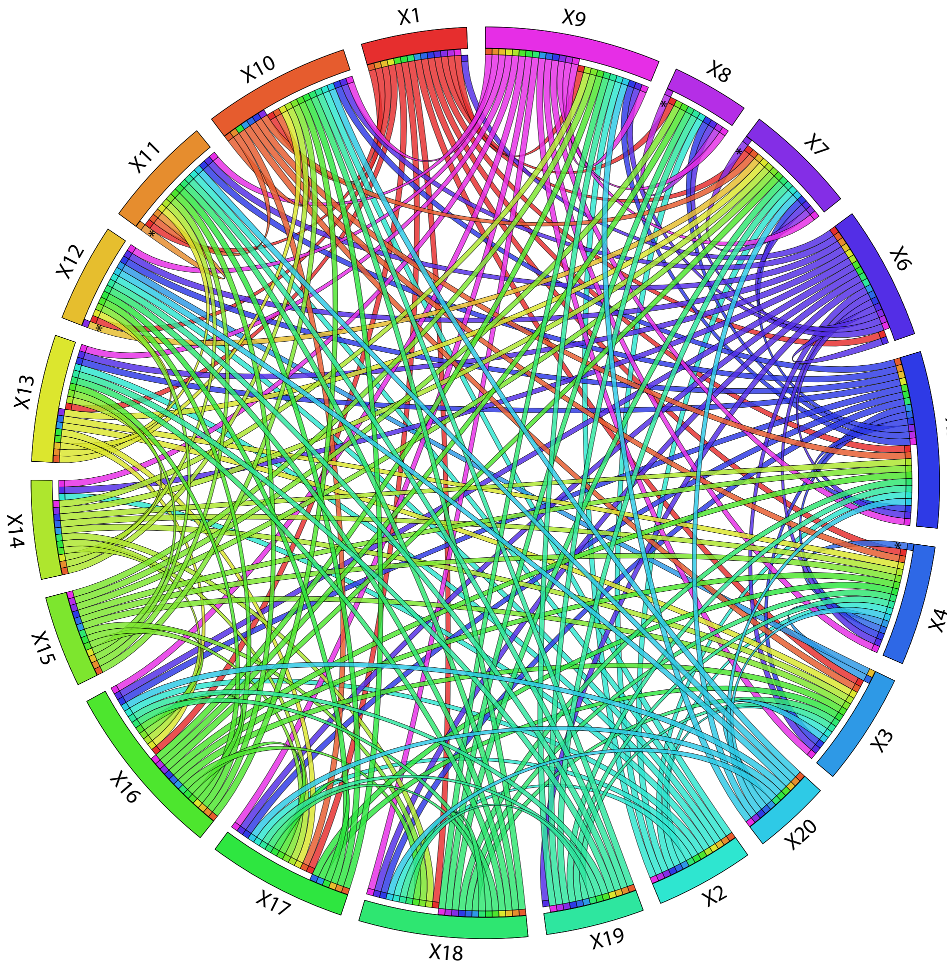

Figure D.6: Gene regulatory network for genes X_1-X_20 in the STEMI group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene.

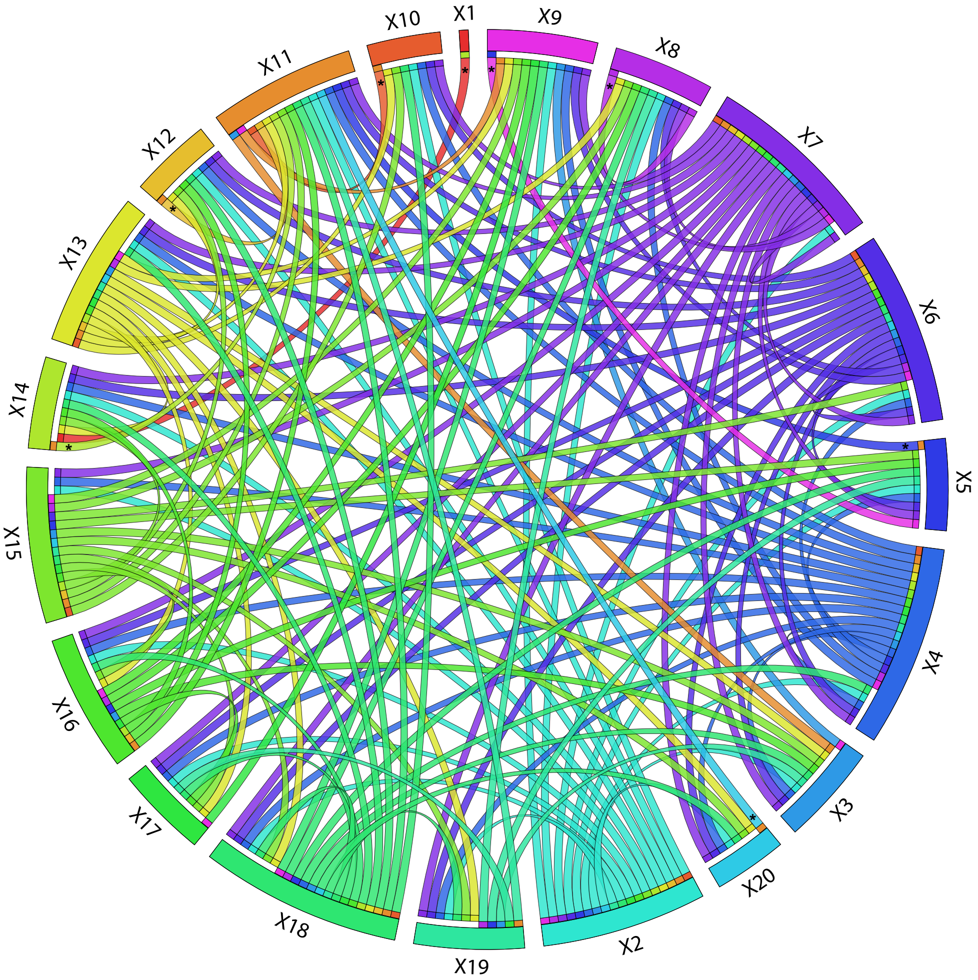

Figure D.7: Gene regulatory network for genes X_1-X_{20} in the UA group. Each gene is represented by an unique colour. Regulatory connections sharing the colour of a gene point towards the genes that regulate it. The $(*)$ marks the regulatory interactions with the largest magnitude that were removed by the sparsity filter. These connections were included to show at least one regulator per gene.

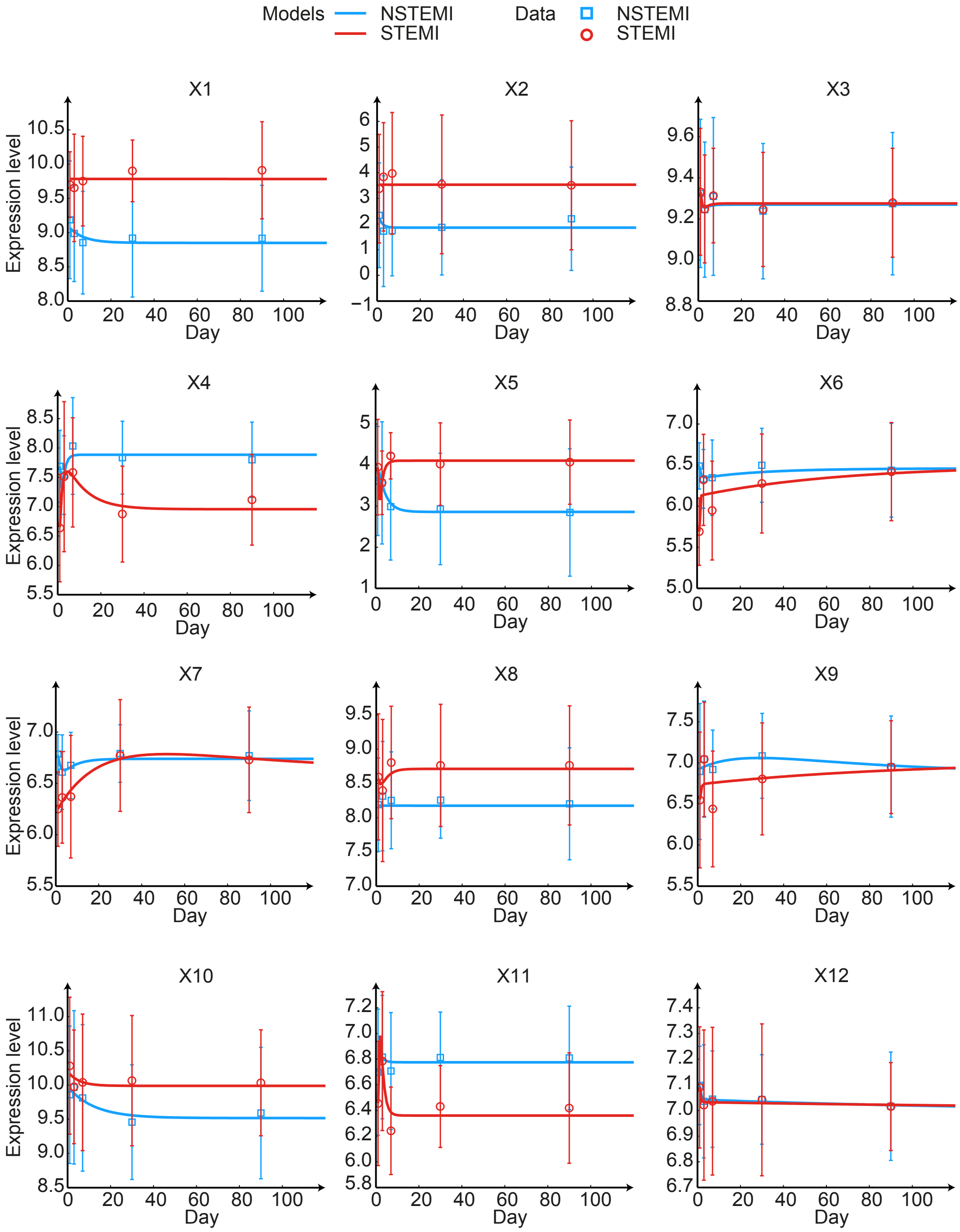

Figure D.8: Estimated dynamic profiles for genes X_1-X_{12} in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages

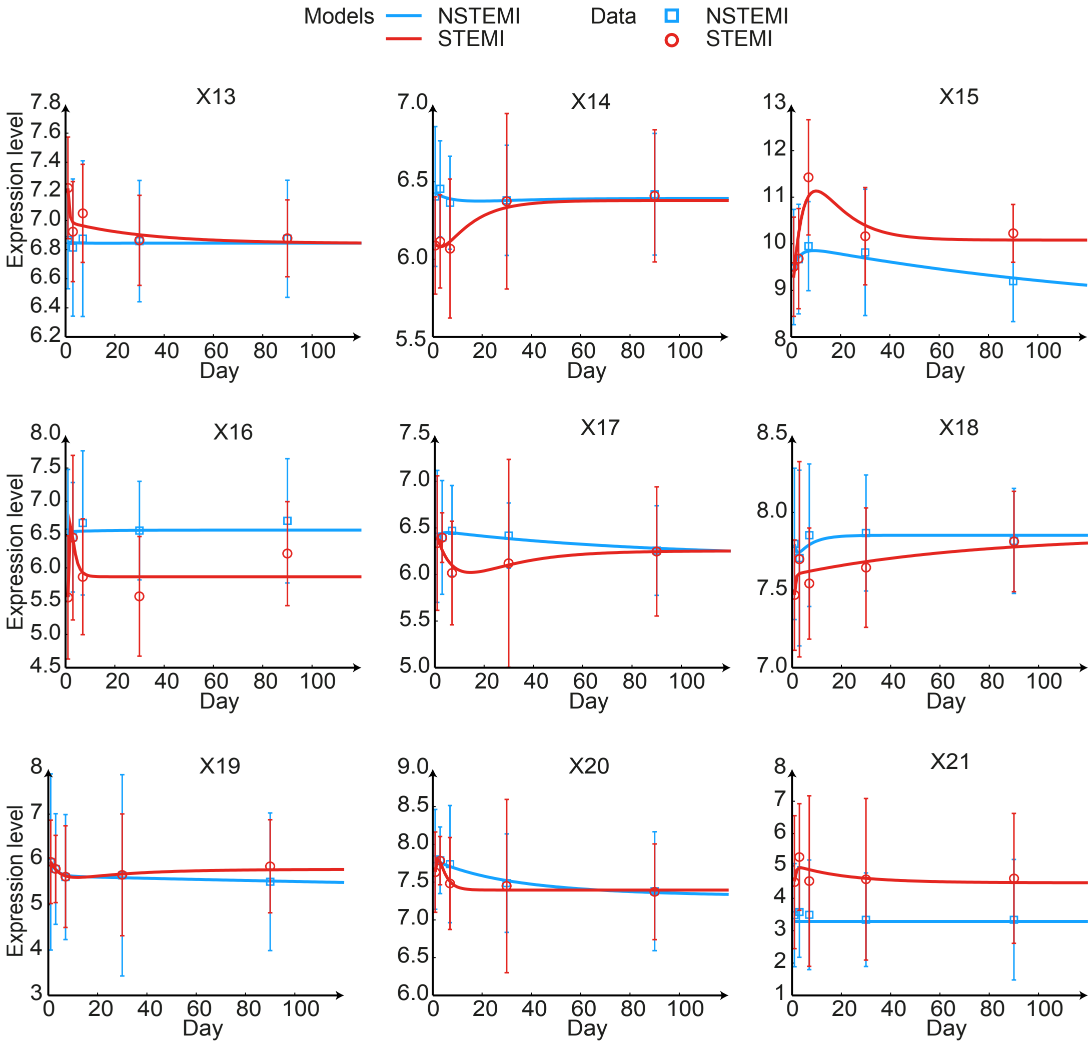

Figure D.9: Estimated dynamic profiles for genes X_{13}-X_{21} in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages

Table D.2: Goodness of fit for the genes differentiating NSTEMI from STEMI

Gene	Residual sum of squares	
	NSTEMI	STEMI
X_1	0.0313	0.0549
X_2	0.2094	0.3125
X_3	0.0030	0.0027
X_4	0.0379	0.0485
X_5	0.0137	0.0305
X_6	0.0077	0.0790
X_7	0.0050	0.0072
X_8	0.0329	0.0436
X_9	0.0158	0.1924
X_{10}	0.0313	0.0395
X_{11}	0.0076	0.0298
X_{12}	0.0002	0.0004
X_{13}	0.0034	0.0133
X_{14}	0.0038	0.0059
X_{15}	0.0331	0.5867
X_{16}	0.0425	0.2115
X_{17}	0.0043	0.0298
X_{18}	0.0067	0.0174
X_{19}	0.0053	0.0067
X_{20}	0.0038	0.0044
X_{21}	0.1717	0.2544

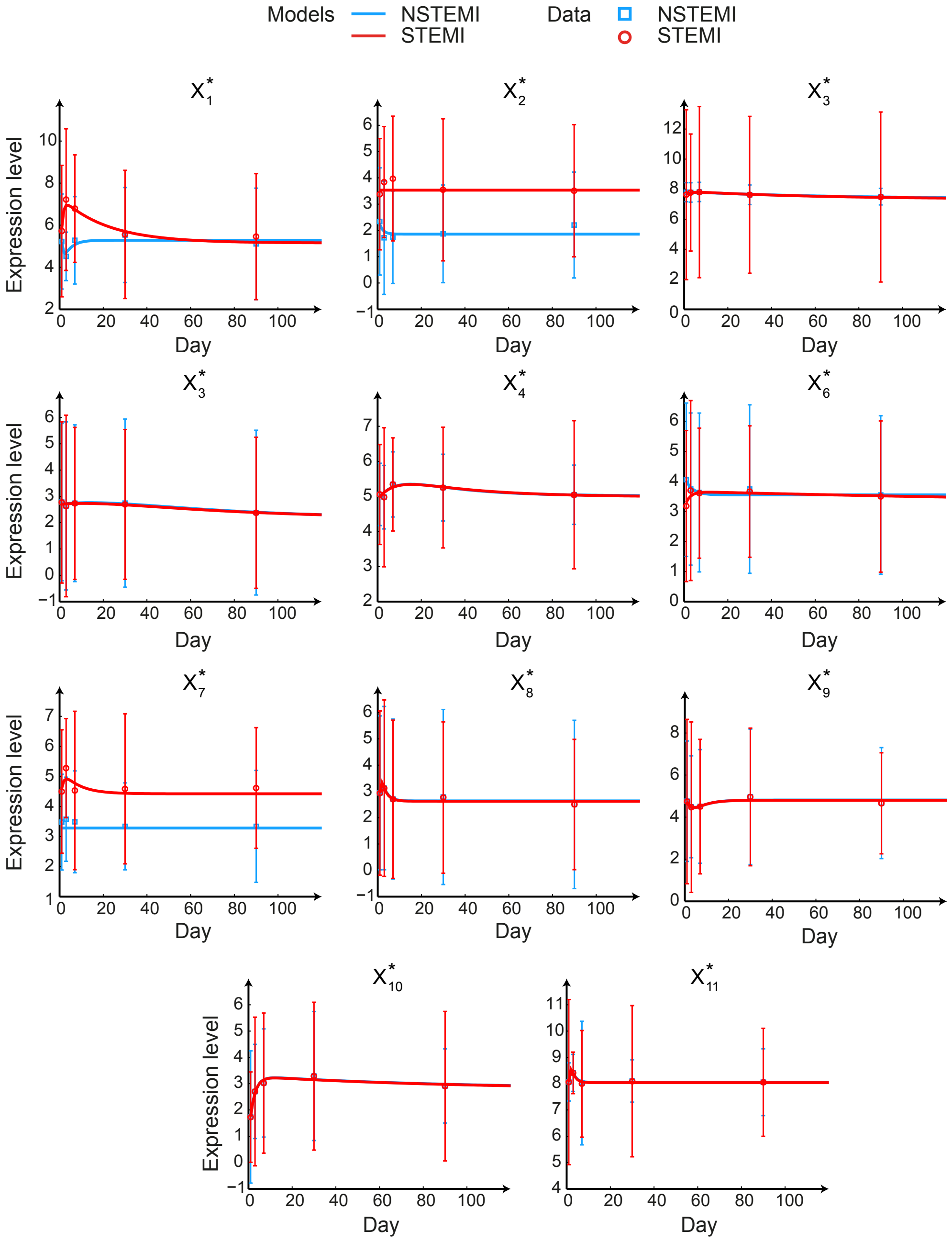

Figure D.10: Dynamic profiles of the combined expression levels for genes $X_1^* - X_{11}^*$ selected by l_1-StaR in the NSTEMI vs. STEMI study. The bars denote one standard deviation around the group-specific gene expression averages.

Bibliography

A.R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H.F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7):e6098, 2009.

Affymetrix. Genechip arrays provide optimal sensitivity and specificity for microarray expression analysis. Technical report, Stanta Clara, California, 2001.

Affymetrix. Statistical algorithms description document. Technical report, Stanta Clara, California, 2002.

M.I. Ahmad and N. Sharma. Biomarkers in acute myocardial infarction. *Journal of Clinical & Experimental Cardiology*, 2012.

M.E. Ahsen, N.K. Singh, T. Boren, M. Vidyasagar, and M.A. White. A new feature selection algorithm for two-class classification problems and application to endometrial cancer. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 2976–2982. IEEE, 2012.

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

T. Akutsu, S. Miyano, S. Kuhara, et al. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28, 1999.

R. Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

B. Alberts. *Molecular Biology of the Cell: Reference edition*. Number v. 1 in Molecular Biology of the Cell: Reference Edition. Garland Science, 2008.

S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review., 2013.

S. Aluru. *Handbook of computational molecular biology*. CRC Press, 2005.

A. Alvarez and H. Lara. A mixed integer nonlinear programming formulation for the problem of fitting positive exponential sums to empirical data. *Opuscula Mathematica*, Vol. 31, no. 4:481–499, 2011.

C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566, 2002.

M.I. Arnone and E.H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997.

C.R. Associates. The burden of acute coronary syndromes in the united kingdom. Technical report, AstraZeneca, Charles River Associates, Life Sciences Practice, 99 Bishopsgate, London, EC2M 3XD, February 2011.

M. Bansal, G. Della Gatta, and D. Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

A.L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, et al. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic acids research*, 37(suppl 1): D885–D890, 2009.

R. Battiti. Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on*, 5(4):537–550, 1994.

M.J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D.L. Wild. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2005.

J. Beirlant, E.J. Dudewicz, L. Györfi, and E.C. Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Y. Bengio and Y. Grandvalet. Bias in estimating the variance of k-fold cross-validation. In *Statistical modeling and analysis for complex data problems*, pages 75–95. Springer, 2005.

H. Bengtsson, G. Jönsson, and J. Vallon-Christersson. Calibration and assessment of channel-specific biases in microarray data with extended dynamical range. *BMC bioinformatics*, 5(1):177, 2004.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

M. Bilban, L.K. Buehler, S. Head, G. Desoye, and V. Quaranta. Normalizing dna microarray data. *Current Issues in Molecular Biology*, 4:57–64, 2002.

S. Billings, M. Korenberg, and S. Chen. Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8):1559–1568, 1988.

J. Bins and B.A. Draper. Feature selection from huge feature sets. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 159–165. IEEE, 2001.

C.M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551, 2010a.

R. Bourgon, R. Gentleman, and W. Huber. Reply to talloen et al.: Independent filtering is a generic approach that needs domain specific adaptation. *Proceedings of the National Academy of Sciences*, 107(46):E175–E175, 2010b.

P.S. Bradley and O.L. Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

U.M. Braga-Neto and E.R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

G. Brassard and P. Bratley. *Fundamentals of algorithmics*. Prentice-Hall, Inc., 1996.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, et al. Minimum information about a microarray experiment (miame)âĂŤtoward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

G. Brown. A new perspective for information theoretic feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 49–56, 2009.

G. Brown, A. Pocock, M.J. Zhao, and M. Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1):27–66, 2012.

B.L. Chen, L.Z. Liu, and F.X. Wu. Inferring gene regulatory networks from multiple time course gene expression datasets. In *Systems Biology (ISB), 2011 IEEE International Conference on*, pages 12–17. IEEE, 2011.

D. Chen, Z. Liu, X. Ma, and D. Hua. Selecting genes by test statistics. *BioMed Research International*, 2005(2):132–138, 2005.

G. Chen, P. Larsen, E. Almasri, and Y. Dai. Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC bioinformatics*, 9(1):75, 2008.

S. Chen, S.A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50 (5):1873–1896, 1989.

S. Chen, C.F. Cowan, and P.M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *Neural Networks, IEEE Transactions on*, 2(2): 302–309, 1991.

T. Chen, H.L. He, G.M. Church, et al. Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, volume 4, page 4, 1999.

J.Y. Ching, A.K. Wong, and K.C.C. Chan. Class-dependent discretization for inductive learning from continuous and mixed-mode data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(7):641–651, 1995.

S. Ciliberti, O.C. Martin, and A. Wagner. Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences*, 104(34): 13591–13596, 2007a.

S. Ciliberti, O.C. Martin, and A. Wagner. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Computational Biology*, 3(2):e15, 2007b.

G.O. Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

J.C. Cortizo and I. Giraldez. Multi criteria wrapper improvements to naive bayes learning. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 419–427. Springer, 2006.

T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

F. d'Alché Buc, P.J. Lahaye, B.E. Perrin, L. Ralaivola, T. Vujasinovic, A. Mazurie, and S. Bottani. A dynamic model of gene regulatory networks based on inertia principle. In *Bioinformatics Using Computational Intelligence Paradigms*, pages 93–117. Springer, 2005.

M. Dash and P.W. Koot. Feature selection for clustering. In *Encyclopedia of database systems*, pages 1119–1125. Springer, 2009.

M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1 (3):131–156, 1997.

E.H. Davidson. *Genomic regulatory systems: in development and evolution*. Academic Press, 2001.

M.J. Davies. The pathophysiology of acute coronary syndromes. *Heart*, 83(3): 361–366, 2000.

P. De Groen and B. De Moor. The fit of a sum of exponentials to noisy data. *Journal of Computational and Applied Mathematics*, 20:175–187, 1987.

M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In *Biocomputing 2003: Proc. Pacific Symposium*, volume 8, pages 17–28, 2002.

H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, et al. Linear modeling of mrna expression levels during cns development and injury. In *Pacific symposium on biocomputing*, volume 4, pages 41–52, 1999.

C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02): 185–205, 2005.

E.R. Dougherty, S. Kim, and Y. Chen. Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80(10):2219–2235, 2000.

S. Draghici, P. Khatri, A.C. Eklund, and Z. Szallasi. Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics*, 22(2):101–109, 2006.

W. Du, Y. Sun, Y. Wang, Z. Cao, C. Zhang, and Y. Liang. A novel multi–stage feature selection method for microarray expression data analysis. *International journal of data mining and bioinformatics*, 7(1):58–77, 2013.

K.B. Duan, J.C. Rajapakse, H. Wang, and F. Azuaje. Multiple svm-rfe for gene selection in cancer classification with expression data. *NanoBioscience, IEEE Transactions on*, 4(3):228–234, 2005.

P.E. Duda and O. Richard. Hart, pattern classification and scene analysis, 1973.

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002a.

S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, 12(1):111–140, 2002b.

S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.

D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. Expression profiling using cdna microarrays. *Nature genetics*, 21:10–14, 1999.

D. Edwards. Non-linear normalization and background correction in one-channel cdna microarray studies. *Bioinformatics*, 19(7):825–833, 2003.

B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008.

T. Erkkilä, S. Lehmusvaara, P. Ruusuvuori, T. Visakorpi, I. Shmulevich, and H. Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, 2010.

F. Falciani. *Microarray technology through applications*. Taylor & Francis, 2007.

C.G. Farquharson and D.W. Oldenburg. A comparison of automatic techniques for estimating the regularization parameter in non-linear inverse problems. *Geophysical Journal International*, 156(3):411–425, 2004.

J.S. Fentz, C. Zornig, H.H. Juhl, and K.A. David. Tissue ischemia time affects gene and protein expression patterns within minutes following surgical tumor excision. *Biotechniques*, 36(6):1030–1037, 2004.

F. Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.

W.M. Freeman, D.J. Robertson, and K.E. Vrana. Fundamentals of dna hybridization arrays for gene expression analysis. *Biotechniques*, 29(5):1042–1055, 2000.

B. Friedland. *Control system design: an introduction to state-space methods*. Courier Dover Publications, 2012.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997a.

N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 139–147. Morgan Kaufmann Publishers Inc., 1998.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

N. Friedman et al. Learning belief networks in the presence of missing values and hidden variables. In *ICML*, volume 97, pages 125–133, 1997b.

R.M. Fryer, J. Randall, T. Yoshida, L.L. Hsiao, J. Blumenstock, K.E. Jensen, T. Dimofte, R.V. Jensen, and S.R. Gullans. Global analysis of gene expression: methods, interpretation, and pitfalls. *Nephron Experimental Nephrology*, 10(2):64–74, 2002.

A. Fujita, J.R. Sato, H.M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M.C. Sogayar, and C.E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):39, 2007.

M. Gardiner-Garden and T. Littlejohn. A comparison of microarray databases. *Briefings in Bioinformatics*, 2(2):143–158, 2001.

R. Gaujoux and C. Seoighe. Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212, 2013.

E.F. Glynn, J. Chen, and A.R. Mushegian. Detecting periodic patterns in unevenly spaced gene expression time series using lomb–scargle periodograms. *Bioinformatics*, 22(3):310–316, 2006.

H. Gohlmann and W. Talloen. *Gene expression studies using Affymetrix microarrays*. CRC Press, 2010.

G.H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.

G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

T. Gong, N. Hartmann, I.S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, S. Bongiovanni, and J.D. Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One*, 6(11):e27156, 2011.

A. Goshtasby. *Image Registration: Principles, Tools and Methods*. Advances in Computer Vision and Pattern Recognition. Springer, 2012.

E.D. Grech and D.R. Ramsdale. Acute coronary syndrome: unstable angina and non-st segment elevation myocardial infarction. *BMJ (Clinical research ed.)*, 326 (7401):1259ÃćâĆňâĂİ1261, June 2003.

P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

W. Gregory Cox, M.P. Beaudet, J.Y. Agnew, and J.L. Ruth. Possible sources of dye-related signal correlation bias in two-color dna microarray assays. *Analytical biochemistry*, 331(2):243–254, 2004.

C. Guan, C. Ye, X. Yang, and J. Gao. A review of current large-scale mouse knockout efforts. *Genesis*, 48(2):73–85, 2010.

G. Gulgezen, Z. Cataltepe, and L. Yu. Stable and accurate feature selection. In *Machine Learning and Knowledge Discovery in Databases*, pages 455–468. Springer, 2009.

M. Gustafsson, M. Hornquist, and A. Lombardi. Constructing and analyzing a large-scale gene-to-gene regulatory network lasso-constrained inference and biological validation. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2(3):254–261, 2005.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

E. Haber and D. Oldenburg. A gcv based method for nonlinear ill-posed problems. *Computational Geosciences*, 4(1):41–63, 2000.

A.J. Hackstadt and A.M. Hess. Filtering for increased power for microarray data analysis. *Bmc Bioinformatics*, 10(1):11, 2009.

J.D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995): 88–93, 2004.

T. Han, C.D. Melvin, L. Shi, W.S. Branham, C.L. Moland, P.S. Pine, K.L. Thompson, and J.C. Fuscoe. Improvement in the reproducibility and accuracy of dna microarray quantification by optimizing hybridization conditions. *BMC bioinformatics*, 7(Suppl 2):S17, 2006.

L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. In *Journal of Machine Learning Research*, pages 1391–1415, 2004.

T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. Gene regulatory network inference: data integration in dynamic modelsâĂŤa review. *Biosystems*, 96(1):86–103, 2009.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

L.V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2):107–128, 1981.

P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cdna microarray analysis. *Biotechniques*, 29(3):548–563, 2000.

R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Springer, 1996.

M.J. Heller. Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

H. Hentschke and M.C. Stüttgen. Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience*, 34(12):1887–1894, 2011.

O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D.S. Charnock-Jones, S. Miyano, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 24(7):932–942, 2008.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

S. Hochreiter, D.A. Clevert, and K. Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006.

K. Holden, D.A. Peel, and J.L. Thompson. *Economic forecasting: an introduction*. Cambridge University Press, 1990.

M. Hollander, D.A. Wolfe, and E. Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

K. Holmström and J. Petersson. A review of the parameter estimation problem of fitting positive exponential sums to empirical data. *Applied Mathematics and Computation*, 126(1):31–61, 2002.

N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J.R. Banavar. Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences*, 98(4):1693–1698, 2001.

C.W. Hsu and C.J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

S. Huang. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine*, 77(6):469–480, 1999.

S. Huang, G. Eichler, Y. Bar-Yam, and D.E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 94 (12):128701, 2005.

E.B. Huerta, B. Duval, and J.K. Hao. A hybrid ga/svm approach for gene selection and classification of microarray data. In *Applications of Evolutionary Computing*, pages 34–44. Springer, 2006.

E.B. Huerta, J.C.H. Hernández, and L.A.H. Montiel. A new combined filter-wrapper framework for gene subset selection with specialized genetic operators. In *Advances in Pattern Recognition*, pages 250–259. Springer, 2010.

N.T. Ingolia. Topology and robustness in the drosophila segment polarity network. *PLoS biology*, 2(6):e123, 2004.

I. Inza, B. Sierra, R. Blanco, and P. Larrañaga. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent and Fuzzy Systems*, 12(1):25–33, 2002.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

A.K. Järvinen, S. Hautaniemi, H. Edgren, P. Auvinen, J. Saarela, O.P. Kallioniemi, and O. Monni. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168, 2004.

H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

D.S. Johnson, C.H. Papadimitriou, and M. Yannakakis. How easy is local search? *Journal of computer and system sciences*, 37(1):79–100, 1988.

I.K. Jordan, L. Mariño-Ramírez, Y.I. Wolf, and E.V. Koonin. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution*, 21(11):2058–2070, 2004.

S. Kabanova, P. Kleinbongard, J. Volkmer, B. Andrée, M. Kelm, and T.W. Jax. Gene expression analysis of human red blood cells. *Int J Med Sci*, 6(4):156–159, 2009.

M. Kahn, M. Mackisack, M. Osborne, and G. Smyth. On the consistency of prony's method and related algorithms. *Journal of Computational and Graphical Statistics*, 1(4):329–349, 1992.

A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1): 95–116, 2007.

B. Kaltenbacher, A. Kirchner, and B. Vexler. Adaptive discretizations for the choice of a tikhonov regularization parameter in nonlinear inverse problems. *Inverse Problems*, 27(12):125008, 2011.

M.D. Kane, T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, and S.J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research*, 28(22):4552–4557, 2000.

S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

L. Kaufman. A variable projection method for solving separable nonlinear least squares problems. *BIT Numerical Mathematics*, 15(1):49–57, 1975.

H.K. Khalil and J. Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, 2002.

M. Kiliszek, B. Burzynska, M. Michalak, M. Gora, A. Winkler, A. Maciejak, A. Leszczynska, E. Gajda, J. Kochanowski, and G. Opolski. Altered gene expression pattern in peripheral blood mononuclear cells in patients with acute myocardial infarction. *PloS one*, 7(11):e50054, 2012.

S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75(1):57–65, 2004.

S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, 34(5-6):975–986, 1984.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

C. Koh, F.X. Wu, G. Selvaraj, and A.J. Kusalik. Using a state-space model and location analysis to infer time-delayed regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009:3, 2009.

R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

R. Kohavi and D. Sommerfield. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *KDD*, pages 192–197, 1995.

I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

H.P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

W.H. Kruskal and W.A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

V.V. Kulkarni, R. Arastoo, A. Bhat, K. Subramanian, M.V. Kothare, and M.C. Riedel. Gene regulatory network modeling using literature curated and high throughput data. *Systems and synthetic biology*, 6(3-4):69–77, 2012.

D. Kundu and A. Mitra. Fitting a sum of exponentials to equispaced data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 448–463, 1998.

L.A. Kurgan and K.J. Cios. Caim discretization algorithm. *Knowledge and Data Engineering, IEEE Transactions on*, 16(2):145–153, 2004.

J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

H. Lähdesmäki, I. Shmulevich, and O. Yli-Harja. On learning gene regulatory networks under the boolean network model. *Machine Learning*, 52(1-2):147–167, 2003.

H. Lähdesmäki, V. Dunmire, O. Yli-Harja, W. Zhang, et al. In silico microdissection of microarray data from heterogeneous cell populations. *BMC bioinformatics*, 6 (1):54, 2005.

H. Lähdesmäki, S. Hautaniemi, I. Shmulevich, and O. Yli-Harja. Relationships between probabilistic boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal Processing*, 86(4):814–834, 2006.

T.N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. Embedded methods. In *Feature extraction*, pages 137–165. Springer, 2006.

C. Lanczos and T. Teichmann. Applied analysis. *Physics Today*, 10:44, 1957.

P. Langley and S. Sage. Induction of selective bayesian classifiers. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 399–406. Morgan Kaufmann Publishers Inc., 1994.

C.L. Lawson and R.J. Hanson. *Solving least squares problems*, volume 161. SIAM, 1974.

J.W. Lee, J.B. Lee, M. Park, and S.H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

W.P. Lee and W.S. Tzou. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics*, 10(4):408–423, 2009.

Y. Lee and C.K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9):1132–1139, 2003.

C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

Y.X. Li, Y.H. Zhu, and X.g. Ruan. Gene selection for leukemia subtype classification from gene expression profile. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 3, pages 1661–1664. IEEE, 2004.

S. Liang, S. Fuhrman, R. Somogyi, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, volume 3, pages 18–29, 1998.

P. Libby. Current concepts of the pathogenesis of the acute coronary syndromes. *Circulation*, 104(3):365–372, 2001.

D.A. Liebner, K. Huang, and J.D. Parvin. Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, page btt566, 2013.

R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nature genetics*, 21:20–24, 1999.

H. Liu and R. Setiono. A probabilistic approach to feature selection-a filter solution. In *ICML*, volume 96, pages 319–327. Citeseer, 1996.

H. Liu, F. Hussain, C.L. Tan, and M. Dash. Discretization: An enabling technique. *Data mining and knowledge discovery*, 6(4):393–423, 2002.

X. Liu, M. Milo, N.D. Lawrence, and M. Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18): 3637–3644, 2005.

X. Liu, M. Milo, N.D. Lawrence, and M. Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22 (17):2107–2113, 2006.

L. Ljung. *System identification*. Springer, 1998.

D.P. Lovell. Biological importance and statistical significance. *Journal of agricultural and food chemistry*, 61(35):8340–8348, 2013.

P. Lu, A. Nakorchevskiy, and E.M. Marcotte. Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences*, 100(18):10370–10375, 2003.

L.T. MacNeil and A.J. Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research*, 21(5): 645–657, 2011.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967.

M. Mahadevappa and J.A. Warrington. A high-density probe array sample preparation method using 10-to 100-fold fewer cells. *Nature biotechnology*, 17(11):1134–1136, 1999.

P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

S. Maldonado and R. Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009.

M. Marczyk, R. Jaksik, A. Polanski, and J. Polanska. Adaptive filtering of microarray gene expression data based on gaussian mixture decomposition. *BMC bioinformatics*, 14(1):101, 2013.

G. Marone, V. Casolaro, R. Cirillo, C. Stellato, and A. Genovese. Pathophysiology of human basophils and mast cells in allergic disorders. *Clinical immunology and immunopathology*, 50(1):S24–S40, 1989.

A. Martínez-Abraín. Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *acta oecologica*, 34(1):9–11, 2008.

M.L. Marx and R.J. Larsen. *Introduction to mathematical statistics and its applications*. Pearson/Prentice Hall, 2006.

H.H. McAdams and A. Arkin. ItâĂŹsa noisy business! genetic regulation at the nanomolar scale. *Trends in genetics*, 15(2):65–69, 1999.

J.N. McClintick and H.J. Edenberg. Effects of filtering by present call on analysis of microarray experiments. *BMC bioinformatics*, 7(1):49, 2006.

R. Mei, X. Di, T. Ryder, E. Hubbell, S. Dee, T. Webster, C. Harrington, M.h. Ho, J. Baid, S. Smeekens, et al. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12):1593–1599, 2002.

A.S. Mellick, C.J. Day, S.R. Weinstein, L.R. Griffiths, and N.A. Morrison. Differential gene expression in breast cancer cell lines and stroma–tumor differences in microdissected breast cancer biopsies revealed by display array analysis. *International journal of cancer*, 100(2):172–180, 2002.

P.E. Meyer, C. Schretter, and G. Bontempi. Information-theoretic feature selection in microarray data using variable complementarity. *Selected Topics in Signal Processing, IEEE Journal of*, 2(3):261–274, 2008.

J.L. Min, A. Barrett, T. Watts, F.H. Pettersson, H.E. Lockstone, C.M. Lindgren, J.M. Taylor, M. Allen, K.T. Zondervan, and M.I. McCarthy. Variability of gene expression profiles in human blood and lymphoblastoid cell lines. *BMC genomics*, 11(1):96, 2010.

L.C. Molina, L. Belanche, and À. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 306–313. IEEE, 2002.

K. Murphy, S. Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

F. Naef, D.A. Lim, N. Patil, and M. Magnasco. Dna hybridization to mismatched templates: a chip study. *Physical Review E*, 65(4):040902, 2002.

S. Nakagawa and I.C. Cuthill. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605, 2007.

T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5): 419–446, 2003.

B.M. Nitin K Singh. lonestar - a feature selection and classification algorithm based on l1 norm svm, 2013. URL `http://sourceforge.net/projects/lonestar/files/code/`.

M. Osborne and G.K. Smyth. A modified prony algorithm for exponential function fitting. *SIAM Journal on Scientific Computing*, 16(1):119–138, 1995.

K.J. Overbaugh. Acute coronary syndrome. *AJN The American Journal of Nursing*, 109(5):42–52, 2009.

W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.

W. Pan. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, 19(11): 1333–1340, 2003.

T. Pang-Ning, M. Steinbach, V. Kumar, et al. Introduction to data mining. In *Library of Congress*, 2006.

L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

R.D. Pearson, X. Liu, G. Sanguinetti, M. Milo, N.D. Lawrence, and M. Rattray. puma: a bioconductor package for propagating uncertainty in microarray analysis. *BMC bioinformatics*, 10(1):211, 2009.

A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S. Fodor. Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proceedings of the National Academy of Sciences*, 91(11):5022–5026, 1994.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Y. Peng, W. Li, and Y. Liu. A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. *Cancer informatics*, 2: 301, 2006.

B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. dâĂŹAlche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19 (suppl 2):ii138–ii148, 2003.

J. Piper, V.P. Pauca, R.J. Plemmons, and M. Giffin. Object characterization from spectral data using nonnegative factorization and information theory. In *Proceedings of AMOS Technical Conference*, 2004.

R. Pique-Regi, A. Ortega, and S. Asgharzadeh. Sequential diagonal linear discriminant analysis (seqdlda) for microarray classification and gene identification. In *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*, pages 112–113. IEEE, 2005.

D. Potts and M. Tasche. Parameter estimation for nonincreasing exponential sums by prony-like methods. *Linear Algebra and its Applications*, 439(4):1024–1039, 2013.

R. Prony. Essai experimental–,-. *J de lâĂŹEcole Polytechnique (Paris)*, 1(2):24–76, 1795.

L. Qian, H. Wang, and E.R. Dougherty. Inference of noisy nonlinear differential equation models for gene regulatory networks using genetic programming and kalman filtering. *Signal Processing, IEEE Transactions on*, 56(7):3327–3339, 2008.

J. Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32:496–501, 2002.

A. Rakotomamonjy. Variable selection using svm based criteria. *The Journal of Machine Learning Research*, 3:1357–1370, 2003.

C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D.L. Wild, and F. Falciani. Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

M.E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G.K. Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

A. Ruhe and P.Å. Wedin. Algorithms for separable nonlinear least squares problems. *Siam Review*, 22(3):318–337, 1980.

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

R.B. Scharpf, C.A. Iacobuzio-Donahue, J.B. Sneddon, and G. Parmigiani. When should one subtract background fluorescence in 2-color microarrays? *Biostatistics*, 8(4):695–707, 2007.

H. Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

O. Scherzer. The use of morozov's discrepancy principle for tikhonov regularization for solving nonlinear ill-posed problems. *Computing*, 51(1):45–60, 1993.

T. Schlitt and A. Brazma. Current approaches to gene regulatory network modelling. *BMC bioinformatics*, 8(Suppl 6):S9, 2007.

D. Scholtens and A. Von Heydebreck. Analysis of differential gene expression studies. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 229–248. Springer, 2005.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.

J.P. Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584, 1995.

H.M. Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

S.S. Shen-Orr, R. Tibshirani, P. Khatri, D.L. Bodian, F. Staedtler, N.M. Perry, T. Hastie, M.M. Sarwal, M.M. Davis, and A.J. Butte. Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4):287–289, 2010.

S.S. Shen-Orr, R. Tibshirani, and A.J. Butte. Gene expression deconvolution in linear space. *Nature Methods*, 9(1):9–9, 2011.

I. Shmulevich and W. Zhang. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, 18(4):555–565, 2002.

I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002a.

I. Shmulevich, E.R. Dougherty, and W. Zhang. From boolean to probabilistic boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, 90(11):1778–1792, 2002b.

I. Shmulevich, E.R. Dougherty, and W. Zhang. Control of stationary behavior in probabilistic boolean networks by means of structural intervention. *Journal of Biological Systems*, 10(04):431–445, 2002c.

I. Shmulevich, E.R. Dougherty, and W. Zhang. Gene perturbation and intervention in probabilistic boolean networks. *Bioinformatics*, 18(10):1319–1331, 2002d.

V.N. Silbiger, A.D. Luchessi, R.D. Hirata, L.G. Lima-Neto, D. Cavichioli, A. Carracedo, M. BriÃşn, J. Dopazo, F. GarcÃ■a-GarcÃ■a, E.S. dos Santos, R.F. Ramos, M.F. Sampaio, D. Armaganijan, A.G. Sousa, and M.H. Hirata. Novel genes detected by transcriptional profiling from whole-blood cells in patients with early onset of acute coronary syndrome. *Clinica Chimica Acta*, 421(0):184 – 190, 2013.

R.J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

G.K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4): 265–273, 2003.

R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, 1(6):45–63, 1996.

P. Somol, J. Novovicová, and P. Pudil. Efficient feature subset selection and subset size optimization. *Pattern Recognit Recent Adv*, 2010.

L. Song, M. Kolar, and E.P. Xing. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2009.

P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

P. Stafford. *Methods in microarray normalization*. CRC Press, 2012.

A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

A. Statnikov, L. Wang, and C.F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319, 2008.

L.J. Steggles, R. Banks, O. Shaw, and A. Wipat. Qualitatively modelling and analysing genetic regulatory networks: a petri net approach. *Bioinformatics*, 23 (3):336–343, 2007.

D. Stekel et al. *Microarray bioinformatics.* Cambridge University Press, 2003.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

J.D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035, 2003.

J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

J.D. Storey, J.E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

T. Strachan and A. Read. *Human Molecular Genetics 4*. Garland Science/Taylor & Francis Group, 2011.

R.O. Stuart, W. Wachsman, C.C. Berry, J. Wang-Rodriguez, L. Wasserman, I. Klacansky, D. Masys, K. Arden, S. Goodison, M. McClelland, et al. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (2):615–620, 2004.

A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, et al. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99(7): 4465–4470, 2002.

C.T. Su and J.H. Hsu. An extended chi2 algorithm for discretization of real value attributes. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):437–441, 2005.

Y. Sun, Y. Wang, J. Wang, W. Du, and C. Zhou. A novel svc method based on k-means. In *Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on*, volume 3, pages 55–58. IEEE, 2008.

J.A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

W. Talloen, D.A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H.W. Göhlmann. I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, 23(21):2897–2902, 2007.

E.K. Tang, P.N. Suganthan, and X. Yao. Gene selection algorithms for microarray data based on least squares support vector machine. *BMC bioinformatics*, 7(1):95, 2006.

Y. Tang, Y.Q. Zhang, and Z. Huang. Fcm-svm-rfe gene feature selection algorithm for leukemia classification from microarray gene expression data. In *Fuzzy Systems, 2005. FUZZ'05. The 14th IEEE International Conference on*, pages 97–101. IEEE, 2005.

Y. Tang, Y.Q. Zhang, and Z. Huang. Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 4(3):365–381, 2007.

Z. Tang, B.S. McGowan, S.A. Huber, C.F. McTiernan, S. Addya, S. Surrey, T. Kubota, P. Fortina, Y. Higuchi, M.A. Diamond, et al. Gene expression profiling during the transition to failure in tnf-α over-expressing mice demonstrates the development of autoimmune myocarditis. *Journal of molecular and cellular cardiology*, 36(4):515–530, 2004.

U. Tautenhahn and Q.n. Jin. Tikhonov regularization and a posteriori rules for solving nonlinear ill posed problems. *Inverse problems*, 19(1):1, 2003.

G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

B. Thompson. Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education*, 70(1):80–93, 2001.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A.N. Tikhonov. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer, 1995.

C.J. Tsai, C.I. Lee, and W.P. Yang. A discretization algorithm based on class-attribute contingency coefficient. *Information Sciences*, 178(3):714–731, 2008.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

H. Unbehauen and G. Rao. A review of identification in continuous-time systems. *Annual reviews in Control*, 22:145–171, 1998.

P. Van Overschee and B. De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

E.P. van Someren, B.L. Vaes, W.T. Steegenga, A.M. Sijbers, K.J. Dechering, and M.J. Reinders. Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics*, 22(4):477–484, 2006.

V. Vapnik. *The nature of statistical learning theory*. springer, 2000.

J.M. Varah. On fitting exponentials by nonlinear least squares. *SIAM journal on scientific and statistical computing*, 6(1):30–44, 1985.

S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl 1):S279–S287, 2001.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482, 1943.

L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589, 2006a.

M. Wang, S.R. Master, and L.A. Chodosh. Computational expression deconvolution in a complex mammalian organ. *BMC bioinformatics*, 7(1):328, 2006b.

S. Wang, D. Li, X. Song, Y. Wei, and H. Li. A feature selection method based on improved fisherâĂŹs discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38(7):8696–8702, 2011.

Y. Wang, X.Q. Xia, Z. Jia, A. Sawyers, H. Yao, J. Wang-Rodriquez, D. Mercola, and M. McClelland. In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer research*, 70(16):6448–6455, 2010.

Y. Wang, T. Joshi, X.S. Zhang, D. Xu, and L. Chen. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, 22(19):2413–2420, 2006c.

Z. Wang, F. Yang, D.W. Ho, S. Swift, A. Tucker, and X. Liu. Stochastic dynamic modeling of short gene expression time-series data. *NanoBioscience, IEEE Transactions on*, 7(1):44–55, 2008.

A.R. Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003.

B.L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, pages 28–35, 1947.

A.R. Whitney, M. Diehn, S.J. Popper, A.A. Alizadeh, J.C. Boldrick, D.A. Relman, and P.O. Brown. Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences*, 100(4):1896–1901, 2003.

J. Wildenhain and E. Crampin. Reconstructing gene regulatory networks: from random to scale-free connectivity. *IEE Proceedings-Systems Biology*, 153(4):247–256, 2006.

W. Wiscombe and J. Evans. Exponential-sum fitting of radiative transmission functions. *Journal of Computational Physics*, 24(4):416–444, 1977.

F.X. Wu, W.J. Zhang, and A.J. Kusalik. Modeling gene expression from microarray expression data with state-space equations. In *Pacific Symposium on Biocomputing*, volume 9, pages 581–592, 2004a.

F.X. Wu, L.Z. Liu, and Z.H. Xia. Identification of gene regulatory networks from time course gene expression data. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 795–798. IEEE, 2010.

Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American statistical Association*, 99(468):909–917, 2004b.

A. Wuensche. Genomic regulation modeled as a network with basins of attraction. In *Pacific Symposium on Biocomputing*, volume 3, page 44, 1998.

Y. Xiao. A tutorial on analysis and simulation of boolean gene regulatory network models. *Current genomics*, 10(7):511, 2009.

M. Xiong, X. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Research*, 11(11):1878–1887, 2001a.

M. Xiong, W. Li, J. Zhao, L. Jin, and E. Boerwinkle. Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism*, 73(3):239–247, 2001b.

R. Yamaguchi and T. Higuchi. State-space approach with the maximum likelihood principle to identify the system generating time-course gene expression data of yeast. *International Journal of Data Mining and Bioinformatics*, 1(1):77–87, 2006.

R. Yamaguchi, R. Yoshida, S. Imoto, T. Higuchi, and S. Miyano. Finding module-based gene networks with state-space models-mining high-dimensional and short time-course gene expression data. *Signal Processing Magazine, IEEE*, 24 (1):37–46, 2007.

H.H. Yang and J.E. Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *NIPS*, pages 687–702. Citeseer, 1999.

Y.H. Yang and T. Speed. Design issues for cdna microarray experiments. *Nature Reviews Genetics*, 3(8):579–588, 2002.

Y.H. Yang, M.J. Buckley, S. Dudoit, and T.P. Speed. Comparison of methods for image analysis on cdna microarray data. *Journal of computational and graphical statistics*, 11(1):108–136, 2002.

M.S. Yeung, J. Tegnér, and J.J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002.

W. Yin, T. Chen, S.X. Zhou, and A. Chakraborty. Background correction for cdna microarray images using the tv+ l1 model. *Bioinformatics*, 21(10):2410–2416, 2005.

E. Yu and S. Cho. Ga-svm wrapper approach for feature subset selection in keystroke dynamics identity verification. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 2253–2257. IEEE, 2003.

L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

F. Yue, K. Wang, and W. Zuo. Informative gene selection and tumor classification by null space lda for microarray data. In *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies*, pages 435–446. Springer, 2007.

M.M. Zavlanos, A.A. Julius, S.P. Boyd, and G.J. Pappas. Inferring stable genetic networks from steady-state data. *Automatica*, 47(6):1113–1122, 2011.

Y. Zhao and R. Simon. Gene expression deconvolution in clinical samples. *Genome Med*, 2(12):93–93, 2010.

Y. Zhong and Z. Liu. Gene expression deconvolution in linear space. *Nature methods*, 9(1):8–9, 2011.

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

L. Zhuo, J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. In *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, pages 71471J–71471J. International Society for Optics and Photonics, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

N.S. Zuckerman, Y. Noam, A.J. Goldsmith, and P.P. Lee. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS computational biology*, 9(8):e1003189, 2013.