ABSTRACT

Title of dissertation: MODEL-BASED GENOMIC/PROTEOMIC SIGNAL PROCESSING IN CANCER DIAGNOSIS AND PREDICTION

Peng Qiu
Doctor of Philosophy, 2007

Dissertation directed by: Professor K. J. Ray Liu
Department of Electrical and Computer Engineering

In recent years, high throughput measurement technologies (gene microarray, protein mass spectrum) have made it possible to simultaneously monitor the expression of thousands of genes or proteins. A topic of great interest is to study the difference of gene/protein expressions between normal and cancer subjects. In the literature, various data-driven methods have been proposed, i.e. clustering and machine learning methods. In this thesis, an alternative model-driven approach is proposed. The proposed dependence model focuses on the interactions among genes or proteins. We have shown that the dependence model is highly effective in the classification of normal and cancer data. Moreover, different from data-driven methods, the dependence model carries specific biological meanings, and it has the potential for the early prediction of cancer. The concept of dependence network is proposed based on the dependence model. The interactions and co-regulation relationships among genes or proteins are modeled by the dependence network, from which we are able to reliably identify biomarkers, important genes or proteins for

cancer prediction and drug development.

The analysis extends to cell cycle time-series, where one subject is measured at multiple time points during the cell cycle. Understanding the cell cycle will greatly improve our understanding of the mechanism of cancer development. In the cell cycle time-series, measurements are based on a population of cells which are supposed to be synchronized. However, continuous synchronization loss is observed due to the diversity of individual cell growth rates. Therefore, the time-series measurement is a distorted version of the single-cell expression. In this thesis, we propose a polynomial-model-based resynchronization scheme, which successfully removes the distortion. The time-series data is further analyzed to identify gene regulatory relationships. For the identification of regulatory relationships, existing literatures mainly study the relationship between several regulators and one regulated gene. In this thesis, we use the eigenvalue pattern of the dependence model to characterize several regulated genes, and propose a novel method that examines the relationship between several regulator and several regulated genes simultaneously.

MODEL-BASED GENOMIC/PROTEOMIC SIGNAL PROCESSING IN CANCER DIAGNOSIS AND PREDICTION

by

Peng Qiu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor K. J. Ray Liu, Chair/Advisor
Professor Carlos A. Berenstein
Professor Min Wu
Professor Robert W. Newcomb
Professor Z. Jane Wang

Dedication

To my parents.

ACKNOWLEDGEMENTS

First of all, I would like to express my deep gratitude to my advisor, Professor K. J. Ray Liu, for his invaluable guidance and support through the years of my Ph.D study. I truly thank him for all the encouragement and for helping me build up my confidence. Without his encouragement and support, I would certainly not have accomplished this thesis. I am honored to have the opportunity to conduct research under Professor Liu's supervision.

I owe my sincere gratitude to Professor Z. Jane Wang for her help and guidance during my research work. The discussions with Professor Wang have always been fruitful. Her intelligence and knowledge have enhanced the depth and width of this thesis. I am grateful to have the opportunity to collaborate with her.

I also want to thank the colleagues in our research group. Although we are working on different topics, the discussions have always been inspiring, leading to constructive suggestions.

I also would like to express my gratitude to Professor Carlos A. Berenstein, Professor Min Wu, and Professor Robert W. Newcomb for their agreement to serve in my dissertation committee and for sparing their times in reviewing the thesis manuscript.

I owe my deepest thanks to my family - my father Shubai Qiu, and my mother Xiaoxian Zhang, who have always stood by me and given me unconditional support. I also would like to thank my special someone, Enlu Zhou, for the wonderful times we spent together on both studying and having fun.

TABLE OF CONTENTS

List of Abbreviations **xii**

1 Introduction **1**
1.1 Background . . . 1
1.2 Related Prior Work and Motivation . . . 2
1.2.1 Cancer Classification and Prediction . . . 4
1.2.2 Biomarker Identification . . . 5
1.2.3 Resynchronization of Microarray Time-Series . . . 7
1.2.4 Discovering Regulatory Network from Time-Series . . . 9
1.3 Thesis Organization and Contributions . . . 11

2 Ensemble Dependence Model for Cancer Classification and Prediction **13**
2.1 Motivation . . . 13
2.2 Dependence Model . . . 15
2.3 Classification Framework . . . 18
2.3.1 Feature Selection . . . 19
2.3.2 Feature Clustering . . . 20
2.3.3 Estimating Ensemble Dependence Model . . . 22
2.3.4 Hypothesis Testing . . . 22
2.4 Classification Results of Microarray Data . . . 24
2.4.1 Microarray Datasets . . . 24
2.4.2 Classification Results for Microarray . . . 25
2.5 Classification Results of Mass Spectrum Data . . . 28
2.5.1 Protein Mass Spectrum Datasets . . . 28
2.5.2 Classification Results for Mass Spectrum . . . 29
2.6 Early Prediction of Cancer . . . 31
2.7 Chapter Summary . . . 37

3 Dependence Network for Biomarker Identification **38**
3.1 Motivation . . . 38
3.2 Dependence Network . . . 40
3.3 Biomarker Identification . . . 43
3.3.1 Dependence-Network-Based Biomarkers . . . 43
3.3.2 Classification-Performance-Based Biomarkers . . . 44

3.4 Biomarker Identification Results . . . 45
3.5 Biological Relevant of the Identified Biomarkers . . . 53
3.6 Chapter Summary . . . 57

4 Resynchronization of Microarray Time-Series 63
4.1 Motivation . . . 63
4.2 System Model for Synchronization Loss . . . 66
4.3 Polynomial-Model-Based Resynchronization . . . 67
4.3.1 Inverse Formulation of Synchronization Loss Model . . . 67
4.3.2 Estimation of Model Parameters . . . 71
4.3.3 Fitting Residue Criterion . . . 72
4.3.4 Cyclic Genes Identification Scheme . . . 73
4.4 Simulation Results . . . 74
4.4.1 Simulations based on sinusoids . . . 74
4.4.2 Simulation based on polynomials . . . 80
4.4.3 Sensitivity analysis . . . 83
4.5 Results on Real Microarry Datasets . . . 85
4.6 Chapter Summary . . . 89

5 Discovering Regulator Network from Microarray Time-Series 90
5.1 Motivation . . . 90
5.2 Analytical Form of Eigenvalue Pattern of the Dependence Model . . . 94
5.2.1 2-Dimensional Case . . . 94
5.2.2 3-Dimensional Case . . . 96
5.2.3 High Dimensional Case . . . 103
5.3 Regulatory Relationships vs. Eigenvalue Pattern . . . 109
5.4 Discovery of Regulatory Relationships . . . 113
5.4.1 Yeast Regulatory Network (Dataset and Prior Knowledge) . . . 113
5.4.2 Correlation Between Gene Expressions . . . 115
5.4.3 Correlation Between Expression and Eigenvalues . . . 119
5.5 Chapter Summary . . . 123

6 Conclusions and Future Research 125
6.1 Conclusions . . . 125
6.2 Future Research . . . 130

A Gaussian Assumption in the Dependence Model 133

B Pre-processing of Protein Mass Spectrum Data 135

C Proof – Enginvalue of Ideal-case Dependence Model 138

D GO Terms of Identified Cell-Cycle Regulated Genes 142

LIST OF TABLES

1.1 Data Format 3

2.1 Classification performance comparison on gastric cancer dataset. "EDM # " means ensemble dependence model with choice of # clusters. In each block, "#/#" means "correct classification rate for cancer samples / correct classification rate for normal samples" 26

2.2 Correct classification rate of the dependence model and SVM for cDNA datasets 27

2.3 Correct classification rate of the dependence model and SVM for Affymetrix datasets 27

2.4 Correct classification rate of the dependence model and SVM for protein mass spectrum datasets. 31

3.1 Identified biomarkers based on dependence network modeling for gastric cancer. The marker genes are mapped to the protein accession numbers in UniProt Knowledgebase (UniProtKB) [67] 54

4.1 For the simulation based on sinusoids, comparison of the normalized average fitting residues for cyclic and non-cyclic genes. 76

4.2 For the simulation based on sinusoids, we compare the proposed method and two previous studies. When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive. 79

4.3 For polynomial based simulation, we compare the normalized average fitting residues for cyclic and non-cyclic genes. 81

4.4 For the polynomial based simulation, we compare the proposed method and two previous studies. When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive. 82

4.5 The performance sensitivity to inexact prior knowledge of cell-cycle length. When the probability of correctly detecting cyclic genes (PD) is fixed, we compare the probability of false positive, under different prior knowledge of cell-cycle T. 84

LIST OF FIGURES

2.1 Ensemble dependence model. 16

2.2 Classification framework. 19

2.3 Eigenvalue pattern of gastric dataset. Fig.(a) shows the four eigenvalues of normal dependence matrices, form 200 subsets of normal data. Fig.(b) shows the eigenvalues of cancer dependence matrices, from 200 subsets of normal data. 33

2.4 The horizontal axis is variation level, which indicates how noisy the four cluster expression profiles are. As the cluster expression profiles become more noisy because of diseases, the eigenvalues of the correspondent dependence matrix will change, following the above curves. 35

2.5 Trend of eigenvalue change in the four stages of samples in the prostate dataset 36

3.1 Motivation of dependence network. 39

3.2 The analysis of binding triples based on the ovarian MS dataset. A+C is the 520 binding triples in normal samples. B+C is the 269 binding triples in cancer samples. C is the overlap, containing 80 triples. 42

3.3 Dependence networks for normal and cancer cases in ovarian cancer dataset. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features. 43

3.4 Fig (a) is the histogram of classification-performance-based biomarkers in the ovarian cancer MS dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the ovarian cancer MS dataset. In both figures, the horizontal axis is the feature index, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. . . 47

3.5 Fig (a) shows the histogram of the performance-based biomarkers for the task of normal vs early stage cancer. Fig (b) shows the histogram of the network-based biomarkers for the task of normal vs early stage cancer. Fig (c) shows the histogram of the performance-based biomarkers for the task of normal vs late stage cancer. Fig (d) shows the histogram of the network-based biomarkers for the task of normal vs late stage cancer. . . 58

3.6 Dependence networks for the prostate cancer dataset: normal, early and late cancer cases. The circles are used to indicate the core features, which are identified through visual inspection. 59

3.7 Fig (a) is the histogram of the classification-performance-based biomarkers in the liver cancer MS dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the liver cancer MS dataset. 60

3.8 Dependence networks for normal and cancer cases in liver cancer MS dataset. The circles are used to indicate the core features, which are identified through visual inspection. 60

3.9 Fig (a) is the histogram of the classification-performance-based biomarkers in the gastric cancer microarray dataset. Fig (b) is the histogram of the dependence-network-based biomarkers of the gastric cancer microarray datasets. 61

3.10 Dependence networks for normal and cancer cases in the gastric cancer microarray dataset. The circles are used to indicate the core features, which are obtained through visual inspection. 61

3.11 Fig (a) is the histogram of the classification-performance-based biomarkers in the liver cancer microarray dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the liver cancer microarray datasets. 62

3.12 Dependence networks for normal and cancer cases in the liver cancer microarray dataset. The circles are used to indicate the core features, which are identified through visual inspection. 62

4.1 For an example of a simulated gene: the simulated sinusoid underlying periodical expression, experiment observation and extracted expression. . . 77

4.2 For the simulated data based on sinusoids: Fig (a) show the histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. Fig (b) is the result of the Fourier analysis used in [16]. Fig (c) shows the upper bound of results from method in [21]. 78

4.3 For an example of a simulated gene: the simulated polynomial underlying periodical expression, experiment observation and extracted expression. . . 81

4.4 For the simulated data based on polynomials: Fig (a) show the histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. Fig (b) is the result of the Fourier analysis used in [16]. Fig (c) shows the upper bound of results from method in [21]. . . 82

4.5 The horizontal axis is the prior knowledge of cell-cycle length, though it may not be the true cell-cycle length $T = 60$. The vertical axis is the difference of fitting residues between cyclic and non-cyclic genes. 84

4.6 Histogram of fitting residues for the cdc28 dataset. Shaded part represents the histogram of fitting residues for the identified cyclic genes. 85

4.7 The experiment observed expression and the extracted periodical expression of genes identified in both the proposed scheme, the previous studies, and traditional methods. The title of each figure represents the gene's ORF name. 86

4.8 The experiment observed expression and the extracted periodical expression of genes identified in the proposed scheme, but not identified by previous studies. The title of each figure represents the gene's ORF name. . . 88

5.1 Shape of the characteristic polynomial of a dependence model with dimension being 3. 100

5.2 Partial model of the gene regulatory network of yeast cell-cycle. 114

5.3 Fig (a) shows the histogram of time-lagged correlation between gene pairs $c_{i,j}$ for all $i, j = 1, 2, ..., 30$. Fig (b) shows the histogram of time-lagged correlation between gene expression and eigenvalue of gene triple, $c_{i;(j,k,l)}$ for all $i, j, k, l = 1, 2, ..., 30$ and $j \neq k$, $j \neq l$, $k \neq l$. 116

5.4 Fig (a) shows the p-value vs absolute value of time-lagged correlation of all gene pairs i, j. Fig (b) shows the p-value vs absolute value of time-lagged correlation of all pairs of regulator and regulated triple $i; (j, k, l)$. 117

5.5 Comparison of ROC curves for two schemes: detecting regulatory relationships from time-lagged correlation between genes' expressions, detecting regulatory relationship from eigenvalue pattern. Fig (a) shows the case where 58 one-hop regulatory relationships from the partial model are considered as the ground truth. Fig (b) shows the cases where 100 regulatory relationships are regarded as the ground truth, containing both one-hop and two-hop regulatory relationships. Fig (c) considers 156 regulatory relationships as the ground truth, containing all relationships less or equal to three-hop. Fig (d) considers 214 regulatory relationships as the ground truth, containing all relationships less or equal to four-hop in the partial model. 124

A.1 Fig (a) (b) (c) show the histograms for the noise term in normal case. Fig (d) (e) (f) show the histograms for the noise term in cancer case. 134

B.1 Pre-processing of MS data. 137

List of Abbreviations

EDM	Ensemble Dependence Model
MS	Mass Spectrum
SVM	Support Vector Machine
GMM	Gaussian Mixture Model
KNN	k - Nearest Neighbors
HT	Hypothesis Testing
LS	Least Square
ML	Maximum likelihood
m/z ratio	mass-to-charge ratio
LDA	Linear Discriminant Analysis
e-value	Eigenvalue
GRN	Gene Regulatory Network
PBN	Probabilistic Boolean Network
DBN	Dynamic Bayesian Network

Chapter 1

Introduction

1.1 Background

As reported by the Center of Disease Control, cancer is the fourth most common disease and the second leading cause of death in the United States. More than 500,000 people die from various forms of cancer each year in the US. Cancer causes a significant financial burden to the health care system, in addition to the tremendous toll on patients and their families. Therefore, understanding the mechanism of cancer development, accurate detection, classification and early prediction of cancer is a research topic of significant importance.

Life science-based research has evolved rapidly during the past decade, driven largely by the sequencing of the complete genome of many organisms and high-throughput technological advances, such as microarray technology and mass spectrum (MS) technology, with a shift from a reductionist approach towards an integrated approach. The new integrated approach investigating "complex" systems

instead of individual components leads to the emerging field of systems biology, which aims at a system-level understanding of biology systems.

The microarray and MS technologies provide us with high throughput measurements at gene and protein level. The gene microarray technology measures the abundance of mRNAs of thousands of genes, and thereby infers how much each gene is expressed [1]. On the other hand, the MS technology measures the proteins. For protein samples, MS converts proteins or peptides to charged pieces that can be separated on the basis of the mass-to-charge ratio (m/z). By measuring the intensity for different m/z ratio, the abundances of different proteins and peptides can be assessed [2].

These high throughput technologies make it possible to systematically study the genes and proteins related to cancer, and would eventually lead to breakthroughs in cancer research. Recently, gene microarray techniques are shown to provide insight into cancer research [3, 4]. In this thesis, we place our emphasis on signal processing and modeling of genomic and proteomic data from microarray and MS technologies, as they are clearly among the leading frontiers that will rapidly reshape cancer study.

1.2 Related Prior Work and Motivation

The expression data from microarray and MS technologies share a common format, as shown in Table 1.1. For each sample, the expression data is a long vector, with each element being the expression level (relative abundance) of a particular gene

or protein. In current available datasets for cancer study, there are usually about one hundred samples, containing both normal and cancer cases. While on the other axis, there are more than ten thousand gene/protein features. The expression data post challenges: the dimension of the data is very high, and all the gene/protein features are potentially inter-correlated. Thus, it is hard to interpret the data. This is where statistics and signal processing can provide help, to systematically interpret the expression data.

	Sample 1	Sample 2	...	Sample 100
gene/protein 1	1.4	5.46	...	1.41
gene/protein 2	3.8	3.08	...	0.67
gene/protein 3	2.34	9.5	...	0.32
gene/protein 4	1.85	1.05	...	1.2
gene/protein 5	3.8	0.19	...	0.2
gene/protein 6	2.23	1.12	...	6.47
gene/protein 7	1.75	2.79	...	0.22
gene/protein 8	6.2	3.96	...	1.27
...	...	...	...	...
gene/protein 12000	1.85	0.19	...	3.01
	Normal samples			Cancer samples

Table 1.1: Format of microarray and MS expression datasets for cancer research.

In the literature, microarray and MS expression data have been examined for cancer classification, biomarker identification, cell-cycle analysis, regulatory network discovery, etc. The rationale behind these applications is based on the belief that the overall behavior of cancer is determined by the expression at the gene/protein level. We now give some specific examples of microarray and MS's applications in cancer research and present the motivation of our study.

1.2.1 Cancer Classification and Prediction

For the purpose of cancer classification, various methods have been proposed for classifying normal samples vs cancer samples or classifying different subtypes of cancer samples. Current methods for cancer classification can be divided into two categories. One is based on the clustering of samples. Some example schemes include Hierarchical Clustering [5], Local Maximum Clustering [6], Self-Organizing Map [7], and K-means Clustering and its variations [8]. These clustering methods usually do not require many prior assumptions, i.e., the underlying model. However, determining the number of clusters is a challenging problem itself, and there is lack of widely-accepted measures to evaluate the clustering performance. The other category is mainly based on machine-learning. Motivated by the success of machine learning algorithms in image and speech processing, many researchers have applied them to the analysis of gene and protein expression data. For example, K-Nearest Neighbors (KNN) [9], Support Vector Machine [10] and Neural Network [11]. Machine learning methods generally yield better results than those of the clustering methods. The clustering and machine learning methods are mostly **data-driven**, which are quite powerful in exploring the data's numerical domain. However, in these methods, gene/protein features are usually treated in a quite separated fashion. The features group behavior and interactions are not considered. Also, in these data-driven methods, without a model to describe the system, it is hard to draw biology insights.

In our work, we propose an alternative **model-driven** approach, called the

dependence model, which focuses on the gene/protein's group behavior and interactions. Because of the limited size of current available datasets and the noisy nature of the expression data, it is not feasible to reliably examine the relationships among all features. Therefore, we propose to group features into clusters, so that the noise level will be reduced and we will be able to reveal the big picture, the ensemble dependence dynamics of the clusters. By doing that, our hypothesis is: the health status information can be reflected by the ensemble dependence relationships among gene clusters. And the hypothesis is validated by the excellent classification performance of the dependence model. In Chapter 2, we will present the dependence model in detail. In addition to the classification performance, we will show the uniqueness of the dependence model, in terms of its biology meaning and its potential in early prediction of cancer.

1.2.2 Biomarker Identification

Biomarker identification is another interesting topic in cancer research. During cancer development, the cancerous cells may release unique genes, proteins and other molecules, which may be regarded as biomarkers. Here biomarkers are defined as the alternations of patterns at the cellular, molecular or genetic level. Caused by the presence of specific diseases, these biomarkers normally serve as the indicators of diseases. Correctly identifying biomarkers for cancer holds enormous potential for the early detection of cancer and drug development [12]. Recently, microarray and MS data have been applied for cancer biomarker identification. For instance, in

[13], a panel of three biomarkers that best separate cancer and normal samples are selected using the linear combination based Unified Maximum Separability Analysis (UMSA). In [14], a particle swarm optimization technique is combined with support vector machine to identify protein biomarkers. In the literature, classification-based biomarker identification criteria are quite popular. The basic idea is to identify features that have the highest discrimination power (classification performance) between normal and cancer cases. However, the performance-based identification results are not quite consistent and reproducible.

The lack of reproducibility is a serious concern. For a particular method, if the conclusion based on one dataset cannot be generalized to other datasets (for the same type of disease), the validity of the conclusion is questionable and the validity of the method is also questionable. In our study, we address this issue by proposing an alternative network-based criterion for reliable biomarker identification. We build dependence networks for both normal and cancer cases, and identify biomarkers by comparing these networks. The basic idea is to identify features with most topology change as biomarkers. In Chapter 3, the results show that the network-based identification criterion yields much more consistent and reproducible results than the performance-based criterion. The biological relevance of the network-based biomarkers is validated by the analysis of their sequence annotations and functionalities.

1.2.3 Resynchronization of Microarray Time-Series

Besides the direct comparison between normal and cancer cases, the understanding of the fundamental cell-cycle system could also contribute a lot to cancer research. One interesting topic is to study the regulatory network in yeast cell-cycle based on the time-series data [15]. The time-series data is obtained by measuring one sample at multiple time points during a certain biology process, such as cell-cycle. Therefore, different from Table 1.1, in time-series data, the horizontal axis of the table represents different time points (rather than different samples). From time-series data, we are able to see how the sample evolves along time. The first critical task in understanding the cell-cycle system is to identify the genes which are periodically expressed during the cell-cycle.

In the current technologies, most expression data are measured based on a population of cells which are synchronized to exhibit similar behaviors [16]. However, even with the most advanced synchronization method, maintaining a tightly synchronized population even over a couple of cycles is a challenging research issue, since continuous synchronization loss is gradually observed due to the diversity of individual cell growth rates [17]. Because of the synchronization loss, the gene expression data observed from a population of cells is different from the gene expression data of a single cell. Therefore, in addition to the noise effect on the measurements, a significant difficulty in identifying cell-cycle regulated genes from time-series data arises from synchronization loss. Direct periodicity test could be misleading or fail due to the fact that the time-series measurements are contributed by a mixture of

cell population growing at different rates, rather than a synchronized population.

Several approaches for identifying cell-cycle regulated genes, when taking into consideration the issue of synchronization loss, have been proposed in the literature. They can be divided into two major categories, differentiated by the absence or presence of other complementary information besides time-series gene expression data. Most studies in the literature belong to the former category, which relies solely on the expression data. Fourier analysis is employed for periodicity test in [17, 18, 19]. The authors present an exact statistical test to identify periodically expressed genes by distinguishing periodicity from random processes in [20]. In [21], a periodic-normal mixture (PNM) model is proposed to fit the transcription profiles of periodically expressed genes. In the second category, an algorithm combining budding index and gene expression data is recently proposed to deconvolve expression profiles in [22]. Regardless these developments, efforts are still needed to accurately identify cyclic genes and recover a more accurate single cell time-series expression compared with the current expression measurements.

In our study, we developed an efficient scheme for identifying periodically expressed genes and reconstructing the underlying single cell gene expression profiles. Our main contributions are two fold. (1) We propose a synchronization loss model by representing the gene expression measurements as a superposition of different cell populations growing at different rates, and we develop a model-based estimation algorithm to reconstruct the underlying single cell gene expression. In previous studies, the single cell expression is often assumed to be sinusoids. However, the proposed algorithm does not make such assumption. It is able to handle a much

larger variety of single cell expression. (2) Using the fitting residue error as criterion, we develop a supervised learning scheme for identifying the cell-cycle regulated genes. The performance of the proposed scheme are examined via both simulations and real microarray gene expression data of *Saccharomyces Cerevisiae*.

1.2.4 Discovering Regulatory Network from Time-Series

The resynchronization analysis of time-series serves as a good pre-processing step to improve the data quality by removing the effect of synchronization loss. After this pre-processing step, a more significant topic is to identify the regulatory network from the time-series data, where the regulatory network describes the complex relationship about how a cell system evolve along time. Discovering and identifying such regulatory network will greatly improve our understanding of cell systems at the gene level. The knowledge of regulatory network will lead to the discovery of the signaling pathways of different biological processes and different diseases, which will greatly facilitate the development of effective drugs.

In the literature, many methods have been proposed to model the gene regulatory network (GRN). In [23, 24], the boolean network is introduced to model the gene regulatory network as boolean relationship in combinatorial logic circuits. In [25], the boolean network is extended to a probabilistic boolean network (PBN), which is a probabilistic mixture of several boolean networks. [26] apply PBN to iteratively grow a regulatory network from microarray time-series data. The Bayesian network models the relationship among genes in terms of conditional probability

distributions and joint probability distributions. Recently, Bayesian network has been used to analyze gene microarray data [27], where the gene regulatory network is modeled as a directed acyclic graph. Further, the Dynamic Bayesian Network (DBN) is proposed in [28], followed by a number of studies [29, 30, 31, 32]. Differential equations are also used to model gene regulatory networks in the literature. In [33], the relationship among genes, mRNAs and proteins are modeled as differential equations. In [34, 35], differential equations are used to model the regulatory relationship among genes, and the parameters are determined through evolutionary programming. In [36, 37], maximum likelihood criterion is applied to determine the parameters of the differential equations. [38] propose to model gene regulatory network using stochastic differential equations. In [39], regulatory relationships with different time lags are examined. In [40] pairwise mutual information and minimum description length (MDL) is applied to infer existence of regulatory relationships. In [41], fuzzy logic is applied to model gene regulatory network.

There is a common property among existing methods, boolean network, Bayesian network, differential equations, etc. The relationship between one or several regulators and one regulated gene is examined. To our knowledge, there is no method that examines several regulated genes simultaneously. In our study, we will address this issue by providing a tool to examine the relationship between one or several regulators and several regulated genes. Since the proposed dependence model and its eigenvalue pattern are able to describe the group behavior of several genes, we will infer the regulatory relationship between the regulators and a group of regulated genes from the relationship between the regulators' expressions and the regulated

genes' group behavior (eigenvalue pattern). Therefore, we are able to examine the regulatory relationships in a novel way, compared with the existing literature.

1.3 Thesis Organization and Contributions

This thesis focuses on the concept of model-driven approach in genomic and proteomic signal processing. Models are developed to address several challenges in bioinformatics and cancer research.

We begin, in Chapter 2, with the cancer classification problem. We first introduce the dependence model and apply it to examine the big picture (the ensemble dependence among clusters of genes), yielding excellent classification performance. Then, we present the biological meaning of the dependence model, which is the uniqueness that distinguishes our work with the dominating data-driven approaches. We show that the eigenvalue pattern of the dependence model is a consistent indicator of dependency and subjects health status. Further, we show that the dependence model has the potential for the early prediction of cancer, and our arguments are supported and validated by several gene and protein datasets.

In Chapter 3, we zoom in to study the details, the dependence relationships among individual gene and protein features. The dependence network is constructed, with each connection representing the dependence relationship between connected features. By comparing the dependence networks constructed for normal and cancer cases, we are able to identify biomarkers. Compared with existing biomarker identification methods, the dependence-network-based biomarkers are much more

consistent and reproducible. The biological relevance of the dependence-network-based biomarkers is validated.

In Chapter 4 and Chapter 5, we shift our attention to time-series analysis. In time-series experiments, there is an inherent problem of synchronization loss, which degrades the data quality. In Chapter 4, we present a polynomial approach that successfully removes the effect of synchronization loss. This is an effective pre-processing step that greatly improves the quality of the data. Comparisons with existing literature show that we are able to better discover cell-cycle regulated genes based on the resynchronized data.

In Chapter 5, the resynchronized data is examined to identify gene regulatory relationships. Existing methods in the literature can only examine one regulated gene at one time. In our study, we propose a method that examines several regulated genes at one time. In order to test the regulatory relationship between a pair of genes (a regulator and a regulated gene), existing methods examine the time-lagged correlation between their expressions. However, for a pair of regulator and regulated genes, such correlation could be weak because of the noisy nature of the expression data. On the other hand, the proposed method uses the eigenvalue pattern of the dependence model to identify regulatory relationships. Analysis on yeast cell-cycle time-series shows that the proposed method performs better than the time-lagged correlation from the Neyman-Pearson point of view. Therefore, we are able to better discover regulatory relationships based on the eigenvalue pattern.

Finally, in Chapter 6, we draw conclusions and discuss some possible future directions.

Chapter 2

Ensemble Dependence Model for Cancer Classification and Prediction

2.1 Motivation

With the rapid development of microarray technology [1] and protein mass spectrum technology [2], it is possible to monitor the expression level of thousands of genes and proteins simultaneously. The large amount of data generated by these high throughput technologies have stimulated the development of many computational methods to study different biological processes at the gene and protein level. Among them, understanding the difference between cancer and normal cells is of particular interest. This includes the difficult task of distinguishing cancerous subtypes, such as benign, invasive, neoplastic or metastatic.

Current methods for the classification of gene and protein expression data can be divided into two categories. One is based on the clustering of samples, which can be used to distinguish cancer and normal samples and to distinguish subtypes of cancers. Some example schemes include Hierarchical Clustering [5], Local Maximum Clustering [6], Self-Organizing Map [7], and K-means Clustering and its variations [8]. These clustering methods do not require many prior assumptions, i.e., the underlying model. However, determining the number of clusters is a challenging problem itself, and there is lack of widely-accepted measures to evaluate the clustering performance. The other category is mainly based on machine-learning. Motivated by the success of machine learning algorithms in image and speech processing, many researchers have applied them to analyze gene and protein data. For example, K-Nearest Neighbors (KNN) [9], Support Vector Machine [10] and Neural Network analysis [11]. Machine learning methods generally yield better results than those of the traditional clustering methods. However, in many machine-learning methods, although gene and protein features form a feature vector and are processed jointly, they are still treated in quite a separate fashion. Gene and protein features' group behaviors and interactions are not considered. Moreover, the existing methods are mostly data-driven methods. Without a model to describe the system, it is hard to draw biology insights.

In our study, we propose to take gene and protein features' group behaviors and interactions into account. We propose a dependence model to study the big picture, the ensemble dependence relationship among clusters of gene and protein features. Because of the limited size of current available datasets and the noisy

nature of the expression data, it is not feasible to reliably examine the relationships among all features. However, if features are clustered properly, the noise level in the resulting cluster expression will be reduced, and we will be able to reveal the ensemble dependence dynamics of gene and protein clusters. We will show that, different from the data-driven methods, the dependence model not only has excellent classification performance, but also carries certain biology meanings. Moreover, the dependence model has the potential for cancer prediction.

2.2 Dependence Model

Because of the limited size of current available datasets, it is not feasible to examine the relationships among all genes. In the proposed dependence model, genes are clustered into several clusters and the clusters' ensemble dependence relationship is studied. We predict, given appropriate and well-sorted clustering results, that genes' group behavior and ensemble dynamics can be revealed. In the following section, several clustering methods are compared, and we will discuss what is appropriate way to cluster genes. In this section, we assume we can cluster genes appropriately and focus on the proposed ensemble dependence model.

After clustering, each cluster contains specific genes that have a well-defined mathematical relationship to one another. To average out experiment noise and enhance genes' common expression within each cluster, the average gene expression profile is used to represent each cluster. Without any prior knowledge, we assume that each cluster is, to some extent, dependent on all the other clusters. Linear

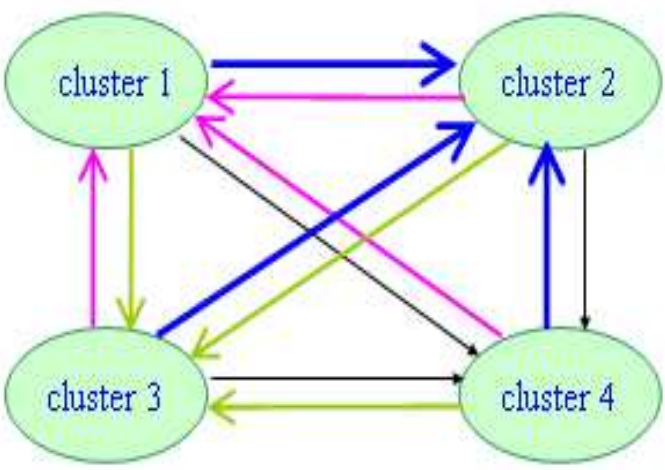


Figure 2.1: Ensemble dependence model.

dependence relationship is studied here, as shown in Figure 2.1, where each arrow represents an inter-cluster dependence relationship. There is a weight a_{ij} associated with each arrow, which indicates to what extent cluster i depends on cluster j. The so-called self-regulation is assumed to be zero, i.e. $a_{ii} = 0$, $i = 1, 2, 3, 4$. Because the cluster average is used to represent each cluster, the intra-cluster dependence relationship within each cluster is averaged out. Later, it is clear that, from a mathematical point of view, allowing non-zero a_{ii} terms will make the model-learning process trivial and un-reasonable, since the results will simply be $a_{ii} = 1$ for any i, and $a_{ij} = 0$ for any $i \neq j$.

The dependence relationship shown in Figure 2.1 can be expressed as the

following linear equation:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ a_{21} & 0 & a_{23} & a_{24} \\ a_{31} & a_{32} & 0 & a_{34} \\ a_{41} & a_{42} & a_{43} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix}, \tag{2.1}$$

or equivalently defined as

$$\mathbf{X} = \mathbf{A}\mathbf{X} + \mathbf{N}, \tag{2.2}$$

where, $\mathbf{A}$ matrix is what we call the dependence matrix; x_i, $i = 1, 2, 3, 4$, are the expression profiles for each gene cluster, which are considered to be random variables. There is a noise-like term $\mathbf{N}$, which could be contributed by model mismatch (i.e. those clusters' expression profiles may not be linearly related) and measurement uncertainty from microarray experiments. For simplicity, the noise-like term is modeled as a Gaussian random vector. Later, we will show that the dependence matrix and statistics of the noise term could be used to distinguish cancer and normal samples.

Equation (2.1) may appear similar to the space-time model of a discrete linear time invariant system in control theory. However, they are quite different. In the state-space model of a discrete linear time invariant system, matrix $\mathbf{A}$ describes how the system state will evolve from the current time step to the next time step. In our case, there is no time concept in the dependence model. The $\mathbf{X}$ vectors on both sides are actually the same. Therefore, each element of the dependence matrix $\mathbf{A}$ does not imply any time evolvement, while it only indicates to what extent one gene cluster is dependent on another cluster.

In this section, the dependence model is introduced in the context of examining the ensemble dependence among gene clusters. However, the dependence model is also applicable to examine protein clusters. Also, the dependence model is not only applicable in studying feature clusters.In Chapter 3, we will see that the dependence model is also applicable to examine the dependence relationship among individual gene and protein features.

2.3 Classification Framework

Since not all genes' expression profiles are informative in understanding the difference between normal and cancer cases, feature selection is needed to exclude irrelevant genes. As required in the ensemble dependence model, gene clustering is performed to group together genes with similar expressions. After feature selection and clustering, selected genes are divided into several groups, each of which is represented by the group average expression. Then, the proposed ensemble dependence model is used to describe the dynamics of gene clusters, one model for the normal case, and another for the cancer case. With these two dependence models, a hypothesis-testing based method is applied to classify normal and cancer data. The main flow of the proposed classification method is shown in Figure 2.2. It includes four main components: feature selection, gene clustering, ensemble dependence model and hypothesis-testing. We will discuss these components as follows.

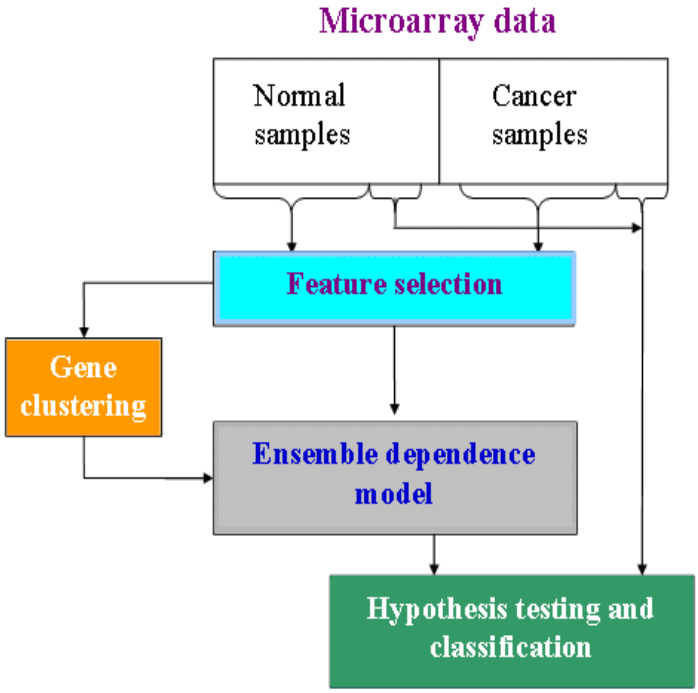


Figure 2.2: Classification framework.

2.3.1 Feature Selection

In this study, we employ two feature selection methods. T-test feature selection criterion is quite popular in microarray analysis. In T-test, each gene is given a score, which evaluates the similarity between its expression profiles in normal and cancer samples. All genes are ranked according to their T-test scores. A p-value is chosen, and genes with scores lower than the chosen p-value are believed to behave most differently between normal and cancer samples.

We also apply another feature selection criterion used in [42, 43]. Equation (2.3) is used to calculate a score for each gene,

$$F(j) = |\frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-}|, \tag{2.3}$$

where, μ_j^+ and σ_j^+ are the mean and standard deviation of gene j 's expression level

in cancer samples, μ_j^- and σ_j^- are the mean and standard deviation of gene j 's expression level in normal samples. Similarly, genes are ranked according to $F(x_j)$ scores. Compared with T-test approach, in this criterion, genes with highest scores are believed to behave most differently between normal and cancer samples.

2.3.2 Feature Clustering

As mentioned above, a proper way of gene clustering is required by the ensemble dependence model. Three standard clustering algorithms are compared: Self-Organizing Map (SOM) [7], K-means [8], and Gaussian Mixture Model (GMM) [44]. In these clustering algorithms, the number of clusters can be pre-defined, as we do in the proposed dependence model. However, K-means clustering is an unstructured method, and it depends more on algorithm initials. SOM is a soft clustering method, but it blurs the difference between adjacent clusters, which is what we want to examine. Therefore, GMM is chosen to cluster genes, since it is a soft clustering method, it can capture cluster difference, and it is much more stable than K-means clustering.

No matter which clustering method is chosen, a similarity measure should be defined. In this study, Euclidean distance of genes' expression profiles is chosen to measure the similarity, because genes with similar expression profiles are likely to share similar functionality [5]. One may argue that Euclidean distance may not cluster genes correctly in terms of their functionalities, and genes in different clusters may share similar functions or are functionally closely related. For example, suppose

that two genes, gene a and gene b, are directly down-regulated by each other. When expression of gene a increases, the expression of gene b decreases, and vice versa. In terms of the Euclidean distance of their expression profiles, gene a and gene b could be far away from each other, thus it is likely that they will fall into different clusters. In this case, mutual information or Euclidean distance of expressions' derivatives would be more appropriate similarity criteria. However, in the proposed method, the average gene expression profile over all genes within one cluster is used to represent each cluster. Even if gene a and gene b are in the same cluster, the example above will be averaged out. That's why we choose the Euclidean distance of genes' expressions as the similarity criterion. Although functionally related genes may fall into different clusters, at least, genes with similar behaviors will be grouped together, thus would represent ensemble mean behaviors more clearly.

Before clustering, the number of clusters needs to be decided. The optimal number of clusters is difficult to determine, because it may depend on different diseases, and different sets of genes under investigation. To determine this parameter, we examined different choices, apply the proposed classification method and suggest the best one by comparing the overall classification performance. In this study, the number of clusters is chosen to be four, according to section 2.4.2. In two of the investigated datasets, the number of normal samples is only around 6, which means we cannot afford to analyze many clusters with the limited size of the available datasets. Although the appropriate number of clusters is hard to determine, in general, the more clusters, the more the dependence relationship is examined, and the more the difference between normal and cancer samples could be revealed.

2.3.3 Estimating Ensemble Dependence Model

Given the gene clustering result, cluster expression profiles can be easily obtained by taking the cluster average. Then, the dependence matrix **A** can be estimated row by row, based on the least squares (LS) criterion. For example, for the first row of **A** matrix,

$$x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + n_1, \tag{2.4}$$

by using the LS criteria, coefficients $a_{1i}, i = 2, 3, 4$ that minimize the noise-like term n_1 are estimated. The statistics of the noise-like term n_1 is estimated at the same time.

For each dataset, after feature selection and gene clustering, we can estimate two dependence models. One for the normal case, and the other for the cancer case. Part of the cancer samples are used to estimate a model for the cancer case, represented by the dependence matrix ($\mathbf{A_c}$) and the distribution of the noise term ($\mathbf{N_c}$); part of the normal samples are used to estimate a model for the normal case, the dependence matrix ($\mathbf{A_n}$) and the distribution of the noise term ($\mathbf{N_n}$). With these two models, the classification problem becomes a hypothesis testing problem.

2.3.4 Hypothesis Testing

In binary hypothesis-testing problems [45], there are two possible hypotheses, H_0 and H_1, associated with two probability distribution functions, f_0 and f_1, on the observation space. In this study, H_0 and H_1 represent the normal case and the cancer case, respectively. Under each hypothesis, the observation Y—-gene

expression, follows a certain probability distribution, written as

$$\begin{aligned} H_0 : Y \sim f_0. \\ H_1 : Y \sim f_1. \end{aligned} \tag{2.5}$$

where f_0 and f_1 are the distribution of the gene expression in normal and cancer samples, respectively. A decision rule δ is a partition of the observation space Γ into Γ_1 and $\Gamma_0 = \Gamma_1^c$, where Γ_1^c is the complement set of Γ_1. In this study, the Maximum Likelihood (ML) approach is used to form the decision rule, that is to compare the conditional probability of observation Y, given underlying hypothesis H_0 or H_1,

$$\Gamma_1 = \{Y \in \Gamma | f_1(Y) > f_0(Y)\}. \tag{2.6}$$

Therefore, the two dependence model can be written in the following hypothesis-testing formulation:

$$\begin{aligned} H_1 : \mathbf{X} = \mathbf{A_c X} + \mathbf{N_c}. \\ H_0 : \mathbf{X} = \mathbf{A_n X} + \mathbf{N_n}. \end{aligned} \tag{2.7}$$

For each incoming unknown sample X (samples not used in model learning), the ML decision rule is applied to predict whether it is normal or cancer. That is, we check whether the incoming sample fits the normal dependence model better, or fits the cancer dependence model better, by comparing the following two log-likelihoods

$$\begin{aligned} Pr(\mathbf{X}|H_1) = -0.5 \log((2\pi)^k |\mathbf{V_c}|) - 0.5(\mathbf{X} - \mathbf{A_c X} - \mathbf{M_c})^T {\mathbf{V_c}}^{-1} (\mathbf{X} - \mathbf{A_c X} - \mathbf{M_c}) \\ Pr(\mathbf{X}|H_0) = -0.5 \log((2\pi)^k |\mathbf{V_n}|) - 0.5(\mathbf{X} - \mathbf{A_n X} - \mathbf{M_n})^T {\mathbf{V_n}}^{-1} (\mathbf{X} - \mathbf{A_n X} - \mathbf{M_n}) \end{aligned} \tag{2.8}$$

where, k is the number of clusters, $\mathbf{V_c}$, $\mathbf{M_c}$, and $\mathbf{V_n}$, $\mathbf{M_n}$ are the first- and second-order statistics of the Gaussian noise-like terms in cancer and normal cases, respectively.

From equation (2.8), it can be seen that the noise-like term N is assumed to be multivariate gaussian. In order to validate this assumption, we analyzed some real microarray and mass spectrum data and examined the kurtosis of the noise-like term. Some results are shown in Appendix A. The kurtosis of the elements of the noise-like vector are close to 3, which validates the gaussian assumption.

2.4 Classification Results of Microarray Data

2.4.1 Microarray Datasets

Since in general the cDNA microarray gene expression data follows standard format and pre-processing operations (e.g. normalization), five public-available cDNA datasets are investigated in detail first. Each of them contains both normal samples and cancer samples. They are, a gastric cancer dataset [46], containing 90 cancer samples and 22 normal samples; a liver cancer dataset [47], containing 82 cancer samples and 74 normal samples; a prostate cancer dataset [48], containing four stages of samples: normal adjacent prostate (NAP), benign prostatic hyperlasia (BPH), localized prostate cancer (PCA) and metastatic cancer (MET), which can be roughly regarded as 15 normal samples (7 NAP and 8 BPH) and 25 cancer samples (14 PCA and 11 MET); a cervical cancer dataset [49], containing 25 cancer samples and 8 normal samples; and a lung cancer dataset [50], containing 37 cancer samples and 6 normal samples.

To be complete, we also investigate three Affymetrix datasets:, a colon cancer

dataset [51], containing 40 cancer samples, 22 normal samples; a prostate cancer dataset [52], containing 77 cancer samples and 59 normal samples; and a lung cancer dataset [53], containing 150 cancer samples and 31 normal samples.

2.4.2 Classification Results for Microarray

As mentioned in the subsection "Feature Clustering", the number of clusters has to be pre-determined for the dependence model. The optimal number of clusters is difficult to determine. In this study, we heuristically choose the parameter as follows: we examine different choices, apply the proposed classification method and suggest the best one by comparing the overall classification performance. In Table 2.1, for the dependence model, different choices of feature selection and number of clusters are examined on the gastric cancer dataset. The performance is shown under leave-one-out cross-validation [54]. From this table, we can see that the choice of feature selection does not affect the classification performance significantly. We believe that using a purely mathematical criterion to select genes is not enough, and that a more meaningful gene selection method which can incorporate biology knowledge is desirable. In the dependence model, different choices of the number of clusters yield slightly different results. Although it is hard to conclude which choice is the best, in general, with sufficient samples, the more clusters, the more the dependence relationship is examined, thus, the better the classification performance could be achieved. Since the number of samples is limited, we can not afford to analyze many clusters. As illustrated in Table 2.1, the performance of the 5-cluster

case is worse than that of the 4-cluster case. The number of clusters is heuristically chosen to be four. We also investigated four other datasets, and observed similar results.

	Golub's approach 100 genes	Golub's approach 500 genes	T-test 3319 genes	All features 6688 genes
EDM 2	98.8% / 95.4%	98.8% / 95.4%	98.8% / 100%	98.8% / 100%
EDM 3	98.8% / 100%	98.8% / 95.4%	100% / 100%	98.8% / 100%
EDM 4	98.8% / 100%	98.8% / 100%	100% / 100%	98.8% / 100%
EDM 5	98.8% / 90.9%	98.8% / 100%	100% / 100%	98.8% / 100%

Table 2.1: Classification performance comparison on gastric cancer dataset. "EDM # " means ensemble dependence model with choice of # clusters. In each block, "#/#" means "correct classification rate for cancer samples / correct classification rate for normal samples"

For each dataset, we use Golub's approach for feature selection, employ the gaussian mixture model to group selected genes into four clusters, and apply the proposed classification scheme to do leave-one-out cross-validation. The results are shown in Table 2.2, where we can see that the proposed scheme yields high classification accuracy. In the reference papers mentioned in the "Microarray Datasets" section, Hierarchical Clustering method is applied to group samples. Since Hierarchical Clustering does not give precise classification results, it is hard to compare the proposed method with it. To examine the proposed scheme, we compare it with the widely-applied linear support vector machine (SVM) approach. The SVM algorithm is a supervised machine learning algorithm. It is a powerful tool in classification and pattern recognition, commonly used in the areas of face detection [55], speaker/speech recognition [56], and handwritten recognition [57]. It also has been applied in the problem of microarray data classification [10, 58], where it is illustrated that SVM provides excellent classification performance. In Table 2.2,

we compare the dependence model and SVM based on several cDNA microarray datasets, and we notice that the linear SVM and the proposed algorithm perform comparably, both providing very high classification accuracy. An interesting observation during the result-checking procedure is that, the classification errors in nearly all leave-one-out validation experiments happen with the same two samples, which may be because of sample mis-labelling. We also compare the dependence model with SVM based on several Affymetrix datasets, with results shown in Table 2.3. We notice that the overall classification performance ranges from 85% to 98% for different types of cancer. Also, we notice that the performance of the proposed dependence model is comparable to that of SVM.

cDNA datasets	Dependence model	SVM
gastric cancer	100%	99.1%
liver cancer	98.72%	98.72%
prostate cancer	97.5%	100%
cervical cancer	93.9%	93.9%
lung cancer	95.35%	97.67%

Table 2.2: Correct classification rate of the dependence model and SVM for cDNA datasets

Affymetrix datasets	Dependence model	SVM
colon cancer	88.71%	85.48%
prostate cancer	85.29%	91.18%
lung cancer	97.79%	99.45%

Table 2.3: Correct classification rate of the dependence model and SVM for Affymetrix datasets

Although SVM and the dependence model provide comparable classification performance, it is worth mentioning that the proposed approach has its advantages. The linear SVM is a hard test approach since a hyper-plane in the feature space is generated to classify test samples. In the proposed ensemble dependence model,

two likelihoods are evaluated to determine the class index. The proposed scheme is a soft test approach, where not only the class index is determined, but also the confidence level of each classification operation can be obtained.

2.5 Classification Results of Mass Spectrum Data

2.5.1 Protein Mass Spectrum Datasets

The two investigated protein mass spectrum (MS) datasets are an ovarian cancer dataset, with 91 normal samples and 161 cancer samples, and a prostate cancer dataset, with 81 normal samples, 84 early stage cancer samples and 84 late stage cancer samples. Datasets are gathered from the National Cancer Institute and Eastern Virginia Medical School. In mass spectrum data, the features are not actual proteins. The features are mass-to-charge (m/z) ratios. In the data matrix for protein data, the horizontal axis represents different samples, and the vertical axis is the m/z ratio. Each m/z ratio corresponds to a protein or a segment of a protein. Because of the noisy nature of mass spectrum datasets, proper pre-processing is needed before any analysis. The details of pre-processing is available in Appendix B. After pre-processing, peaks in the mass spectra are identified as features for further analysis.

2.5.2 Classification Results for Mass Spectrum

Similar with microarray datasets, for protein mass spectrum (MS) datasets, we also need feature selection and feature clustering. The model estimation component and the hypothesis-testing component are also the same. The only difference is how to find a representative to effectively represent each cluster. A most straightforward way would be using the average of all features within one cluster as the cluster representative. However, due to the specific properties of the protein MS data, we propose a concept of *virtual protein*. Here virtual protein is defined as a linear weighted combination of different MS features within a cluster. In order to represent each cluster, a virtual protein is generated as the cluster representative.

We argue that a virtual protein representation makes more sense than a straightforward averaging for two main reasons. First, in MS data, some features correspond to high intensity peaks, while some features correspond to low intensity peaks. In order to avoid high intensity features dominating its cluster, the virtual protein is generated by the weighted average of the cluster members, which can provide more information than the straightforward averaging. Secondly, MS measures the abundance of different peptides with different mass-to-charge (m/z) ratios. Due to the measurement process of MS, one particular cancer-related protein can be represented by several peptides, each of which corresponds to a certain m/z feature. Thus, a linear combination of m/z features may lead to a virtual protein which represents the underlying cancer-related protein. For the purpose of constructing a virtual protein, i.e. determining the weights, we employ the linear

discriminant analysis (LDA) [60]. Since we are interested in the virtual proteins which are cancer-related and thus best represent the difference between a cancer and non-cancer sample, we believe LDA provides an efficient way to construct such a virtual protein.

In one cluster, each feature contains some information of the difference between normal and cancer samples. Through linear discriminant analysis, a set of weights is determined in order to extract one virtual feature that best distinguishes the two cases. For one cluster, denote the training dataset as a matrix $[P, Q]$, where each row corresponds to one feature in this cluster; each column of P corresponds to one cancer sample; and each column of Q corresponds to one normal sample. Linear discriminant analysis finds a linear combination w that maximizes the ratio of between class variance and within class variance, which is wS_Bw^T/wS_Ww^T. S_B and S_W can be calculated as follows,

$$\begin{aligned} S_B &= (\mu_P - \mu)(\mu_P - \mu)^T + (\mu_Q - \mu)(\mu_Q - \mu)^T \\ S_W &= (P - \mu_P)(P - \mu_P)^T + (Q - \mu_Q)(Q - \mu_Q)^T \end{aligned} \tag{2.9}$$

where μ_P is the average of cancer sample; μ_Q is the average of normal sample; and μ is the average of all samples. Through the Lagrange method, the w that maximizes wS_Bw^T/wS_Ww^T is the eigenvector of $S_W^{-1}S_B$ that corresponds to the largest eigenvalue. Then, weights w are used to generate the virtual protein of this cluster.

With the virtual protein as cluster representative, we follow the classification framework in Figure 2.2 to assess the classification performance of the dependence model. The performance of the dependence model and SVM is compared under

leave-one-out cross-validation. From Table 2.4, we can see that in the ovarian cancer dataset, the proposed model and SVM have comparable performance. In the prostate cancer dataset, when we classify normal samples against late stage cancer samples, the two schemes also have comparable performance. However, in the prostate cancer dataset, when we classify normal samples against early stage cancer samples, where the classification task appears to be more difficult, the proposed ensemble dependence model out performs SVM. Since SVM produces a linear boundary that best separates the training data, in the case where normal and early cancer samples are not well separated, SVM does not perform well. However, the proposed model fits the data well. It can reduce the noise, and yield satisfactory separation between normal and cancer data.

Protein Mass Spectrum	Dependence model	SVM
ovarian cancer	96.60%	96.83%
prostate: normal vs early cancer stage	98.79%	78.79%
prostate: normal vs late cancer stage	99.39%	98.79%

Table 2.4: Correct classification rate of the dependence model and SVM for protein mass spectrum datasets.

2.6 Early Prediction of Cancer

In the previous section, we have shown that the dependence model and SVM achieve comparable performance, while the dependence model is sometimes better. In this section, we will present the uniqueness of the dependence model as a model-driven approach, in terms of its biology meaning and its potential for early prediction of cancer.

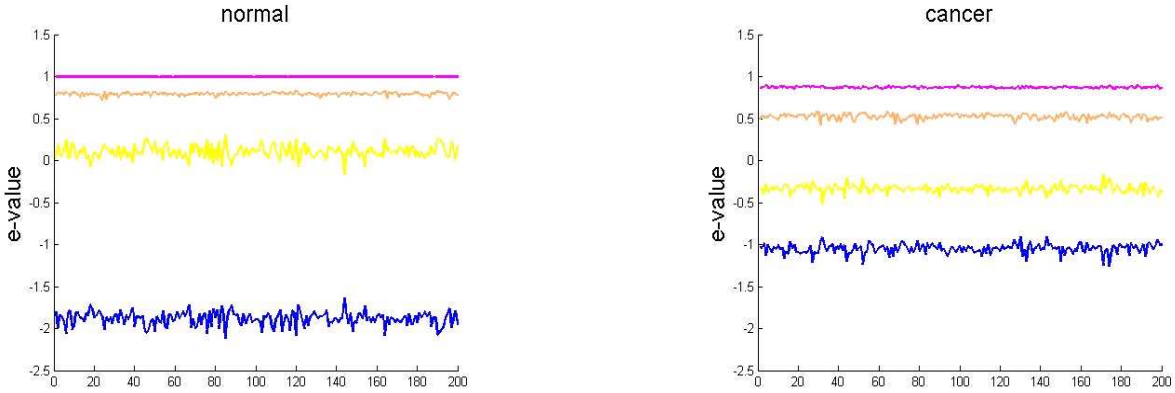
Below are typical examples of the estimated cancer dependence matrix $\mathbf{A_c}$ and the normal dependence matrix $\mathbf{A_n}$:

$$\mathbf{A_c} = \begin{bmatrix} 0 & 0.3676 & 0.1098 & -0.0398 \\ 1.6274 & 0 & -0.5400 & 0.0067 \\ 0.2103 & -0.2336 & 0 & 0.3922 \\ -0.1537 & 0.0058 & 0.7912 & 0 \end{bmatrix}. \tag{2.10}$$

$$\mathbf{A_n} = \begin{bmatrix} 0 & 0.4502 & 0.5154 & -0.4208 \\ 1.8188 & 0 & -1.0142 & 0.5021 \\ 0.6592 & -0.3210 & 0 & 0.7028 \\ -0.7767 & 0.2294 & 1.0145 & 0 \end{bmatrix}. \tag{2.11}$$

Comparing these two matrices entry-wisely does not reveal a clear difference. However, when exploring the eigenvalue domain, we observe that, there are clearly two different patterns, in Figure 2.3. Figure 2.3 is derived from the gastric cancer microarray dataset. In order to obtain Figure 2.3(a), we randomly pick 80% of the normal samples, estimate the normal dependence model, calculate the eigenvalues, and repeat this 200 times. Therefore, we have 200 sets of eigenvalues derived from different subsets of the normal samples, and we plot the eigenvalues in Figure 2.3(a). From this figure, we can see that the eigenvalue pattern derived from different subset of the normal samples is quite consistent. We did exactly the same thing on the cancer samples to obtain Figure 2.3(b). It can be observed that, in general, the eigenvalues for the normal dependence matrix have larger absolute values than those of the cancer case. The difference is most distinguishing at the smallest eigenvalue.

We believe that the different patterns in eigenvalue domain could play an important role in predicting whether an unknown sample is normal or cancer.



(a) (b)

Figure 2.3: Eigenvalue pattern of gastric dataset. Fig.(a) shows the four eigenvalues of normal dependence matrices, form 200 subsets of normal data. Fig.(b) shows the eigenvalues of cancer dependence matrices, from 200 subsets of normal data.

Recall that, after gene clustering, the dependence matrix is obtained from the cluster representative expression profiles. What is the relationship between cluster expression profiles and the eigenvalue pattern of the dependence matrix? What kind of cluster expression profiles will result in the two different patterns observed in Figure 2.3? To answer these questions, an ideal case is defined, where there is no noise-like term in equation (2.1), meaning the four cluster expression profiles are completely linearly dependent. In the other words, each cluster expression profile could be exactly written as a linear combination of the other clusters' expression profiles. Thus, the noise-like term is zero. More specifically, if the four clusters' expression profiles satisfy

$$x_1 = \alpha_1 x_2 + \alpha_2 x_3 + \alpha_3 x_4, \tag{2.12}$$

then the noise-like term is zero. In this case, the dependence matrix will have a

special structure as follows,

$$\mathbf{A_{ideal}} = \begin{bmatrix} 0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \frac{1}{\alpha_1} & 0 & -\frac{\alpha_2}{\alpha_1} & -\frac{\alpha_3}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & 0 & -\frac{\alpha_3}{\alpha_2} \\ \frac{1}{\alpha_3} & -\frac{\alpha_1}{\alpha_3} & -\frac{\alpha_2}{\alpha_3} & 0 \end{bmatrix}. \tag{2.13}$$

We can show that the eigenvalues of the above matrix are 1,1,1,-3, no matter what are the values of $\alpha_i, i = 1, 2, 3$. We define the above case in (2.13) as the *ideal case*. This property can be generalized into cases with higher dimensions. For example, if we have M clusters, the eigenvalues of the M-by-M matrix $\mathbf{A_{ideal}}$ are $\{1, 1, ..., 1, -(M-1)\}$, no matter what are the values of $\alpha_i, i = 1, 2, ..., M-1$. (Proof in Appendix C).

We simulate the ideal case. Based on the ideal case, we gradually introduce larger and larger random variation to make the four cluster expression profiles more and more noisy, thus more and more independent. At each variation level, a dependence matrix is estimated, and the corresponding eigenvalues are calculated. Compared with the ideal case, as the cluster expression profiles suffer more and more noisy variations, the eigenvalues of their dependence matrix will change and follow the trends shown in Figure 2.4. Compared with Figure 2.3, it can be suggested that the cluster expression profiles in cancer samples correspond to a much larger variation level than that of the normal samples, which means the gene clusters' behavior in cancer samples is much more noisy than that of the normal samples. Here we propose to explain this intuitively. In the normal samples, gene clusters' dependence relationship is clearer, and gene clusters work more cooperatively. Therefore, we ob-

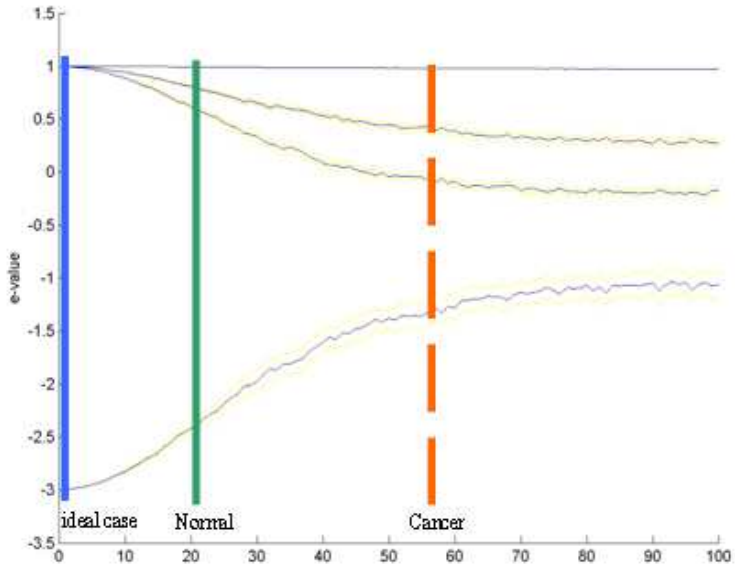


Figure 2.4: The horizontal axis is variation level, which indicates how noisy the four cluster expression profiles are. As the cluster expression profiles become more noisy because of diseases, the eigenvalues of the correspondent dependence matrix will change, following the above curves.

serve that the clusters behave more dependently. On the other hand, in the cancer case, the dependence relationship among gene clusters is overwhelmed by a large variation caused by diseases, which thus make gene clusters work less cooperatively, and make the cell system become worse and worse.

This is the biology meaning behind the dependence model, where normal and cancer cases are distinguished by the strength of dependence among gene clusters. On the other hand, in SVM, a data-driven method, we can find a hyperplane that best separates normal and cancer cases. But SVM can not explain why this side corresponds to normal and the other side corresponds to cancer, why not the other way around. This is one big advantage of the dependence model, compared with existing data-driven methods.

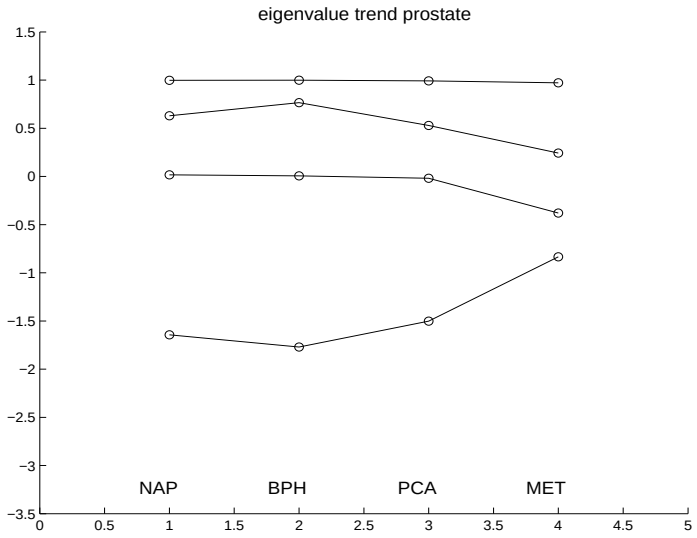


Figure 2.5: Trend of eigenvalue change in the four stages of samples in the prostate dataset

Moreover, the transition stage between normal and cancer eigenvalue patterns suggests that the eigenvalue pattern from the dependence model can be used as features to predict early stage of cancer development, whether a sample is in transition from healthy to cancer. For example, if the eigenvalue pattern of a patient falls between the normal and cancer eigenvalue patterns, we will tell the patient to take treatments to prevent the possible development of cancer. To support the above argument, we use the prostate cancer dataset as an example. As mentioned earlier, it contains four stages of data, NAP, BPH, PCA, MET, which can be simply regarded as from normal (NAP and BPH), to cancer stage (PCA), to cancer in stage (MET). The dependence matrix and eigenvalues of each stage are calculated. As shown in Figure 2.5, the overall trend of eigenvalues from normal to cancer follows the trend in Figure 2.4, which verifies the above argument. In the future study, we will obtain more datasets to further verify the argument. This argument of early prediction of cancer is the biggest implication and significance of the dependence model.

2.7 Chapter Summary

In this chapter, we develop a model-driven approach, called the dependence model. The dependence model is highly efficient in classification of normal and cancer samples, using gene microarray data and protein mass spectrum data. We compare the proposed approach with the widely-applied support vector machine algorithm. Although these two algorithms show comparable performance, our algorithm presents a fundamental departure from the existing SVM approach because it develops a more plausible ensemble dependence model by taking genes group behaviors and interactions into account, and thus may have potential to classify intransigent data on which other classifiers balk.

An interesting observation is noted in the eigenvalue domain: two distinguishing eigenvalues patterns of the dependence models are noted for the normal and cancer cases. From the eigenvalue pattern, we derived the biology meaning behind the dependence model, showing that the gene clusters are working more cooperatively in the normal case and less cooperatively in the cancer case. By examining one prostate cancer dataset, we also illustrated the dependence model's potential in early prediction of cancer. The biology meaning and the ability for cancer prediction represent the key difference between our study and the existing literature. More details can be found in [61, 62, 63]. In the future study, we will obtain more datasets to further verify the dependence model.

Chapter 3

Dependence Network for Biomarker Identification

3.1 Motivation

The functionality of a gene or a protein is not solely characterized by its own structure. Its surroundings and interacting genes and proteins also play important roles in determining the function. In short, the gene and protein interaction network can provide detailed functional insights [64]. In Chapter 2, we develop the dependence model and apply it to examine the big picture, the ensemble dependence among groups of genes. Although the big picture reflects some biology insights, the detail relationships among individual gene and protein features are missing. Another motivation is that the ensemble dependence model in Chapter 2 requires a clustering method, which is heuristic. Therefore, in this chapter, we will zoom in to examine the missing detail dependence relationships among individual features, and avoid

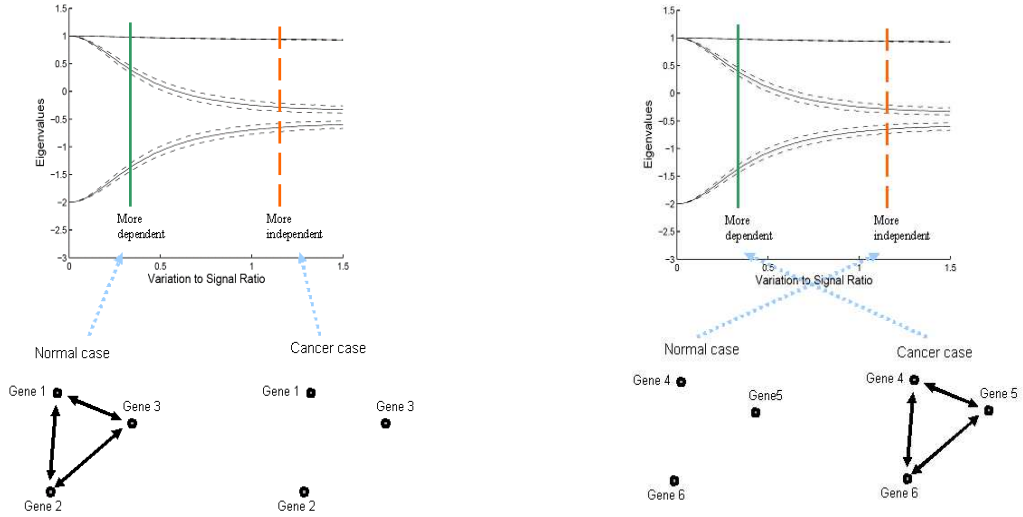


Figure 3.1: Motivation of dependence network.

the heuristic clustering component. The dependence relationships among individual features construct the dependence network, from which we are able to identify important genes and proteins for cancer development and effective treatment.

Figure 3.1 shows the eigenvalue vs dependence relationship among 3 features. This is almost the same with the Figure 2.4 in the previous chapter. Small variation levels, meaning the eigenvalue pattern close to the left side, indicate more dependent; while large variation levels, meaning the eigenvalue pattern close to the right side, indicate more independent. In Figure 3.1(a), we have an example of a gene triple consisting of gene 1, 2, 3. They work together in the normal case. But their relationships are disturbed in the cancer case. For this triple, normal implies more dependent, while cancer implies more independent. On the other hand, in Figure 3.1(b), there is another example, a triple consisting of gene 4, 5, 6. They are not related in the normal case. But their dependence relationships are activated in the

cancer case. Then, for the second example, normal implies more independent, while cancer implies more dependent.

From these examples, we can imagine that, if we use the dependence relationship to define a network and draw such networks for normal and cancer cases, we will see some or maybe a lot of difference. And such difference may help us in identifying biomarkers, the important genes for cancer prediction and effective treatments. In this chapter, we will present how to define and construct a dependence network, and apply the dependence network for biomarker identification.

3.2 Dependence Network

A dependence network is a set of nodes (such as gene or protein features) and linear dependence interactions among them. All the nodes and connections collectively carry out specific functions. Each connection represents an inter-component dependence relationship with an associated weight a_{ij} indicating to what extent component i depends on component j. In the following, we describe how a dependence network is constructed.

In Chapter 2, it is shown that the eigenvalue pattern is closely related to the dependence relationship of a group of features, especially the smallest eigenvalue. Take a three-feature case for example. From the noise-free ideal case, as the three features become more and more independent, the eigenvalues of their dependence matrix will change and follow the trends shown in Figure 3.1. In the three-feature example, when the three features are perfectly linearly dependent, the smallest

eigenvalue is -2. When the dependence relationship become weaker and weaker, the smallest eigenvalue increases, and eventually saturate to around -0.7. Thus, for any feature triple, by examining the eigenvalue pattern of their dependence matrix, we are able to tell how dependent they are, how closely related they are.

Since, the eigenvalue pattern can serve as an indicator of strength of dependency, if we examine three individual features at one time, we can find all closely related feature triples through an exhaustive search. The elements in each triple share strong dependence relationships, which indicates that they have a strong influence on each other in the dependence network. Take the ovarian cancer protein MS dataset as an example. For the normal case, we exhaustively examine the eigenvalue pattern for all possible feature triples. A threshold -1.5 is applied. If the smallest eigenvalue of a feature triple is lower than the threshold, there exists a strong dependence relationship within the triple, which is called the "binding triple". Similar analysis is applied to the cancer samples. The results are shown in Figure 3.2. In the normal case, 520 triples pass the threshold; while in the cancer case, 269 triples pass the threshold. Moreover, there are only 80 triples in the overlap between normal and cancer cases. The results suggest that, from healthy to cancerous, some dependence relationships are disabled; while some other dependence relationships are activated. The small overlap indicates that, from healthy to cancerous, the overall dependence relationships go through a major change.

The dependence network is constructed from binding triples. As in graph theory, the topology of an n-node network can be represented by an $n \times n$ adjacency matrix D. If feature i and feature j both appear in a binding triple, it is

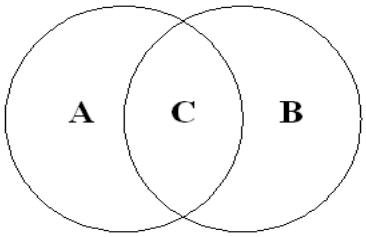
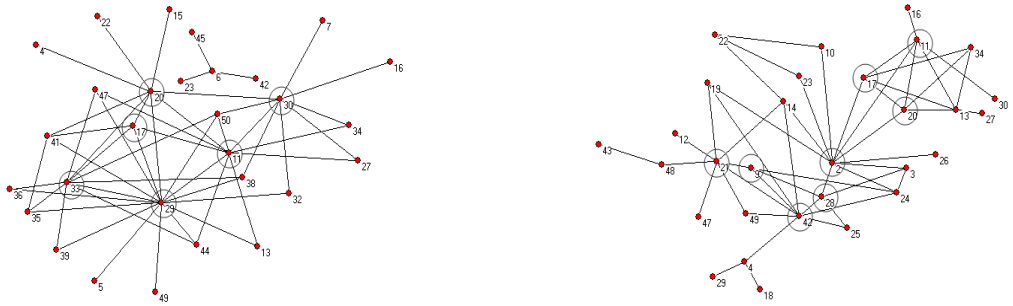


Figure 3.2: The analysis of binding triples based on the ovarian MS dataset. A+C is the 520 binding triples in normal samples. B+C is the 269 binding triples in cancer samples. C is the overlap, containing 80 triples.

suggested by the dependence model that feature i and feature j are closely related. And we will count once for D_{ij}, the connection between feature i and feature j. Basically, we count the appearance of all feature pairs in binding triples, and form an adjacency matrix D. Then, the adjacency matrix D is normalized by the total number of binding triples. Each element D_{ij} is a confidence value, which indicates the importance and strength of the connection between feature i and feature j. We call network D the dependence network. The dependence networks can be visualized as shown in Figure 3.3, where strong dependence relationship is reflected in small distance between connected nodes. The length of each connection is defined to be inverse proportional to the confidence value. Because the confidence values are normalized, through $1/D_{ij}$, features with strong dependence relationship will locate close to each other, while features with weak dependence relationship will be far apart. From Figure 3.3, we are able to see the importance of each node and identify potential biomarkers.



(a) Normal case (b) Cancer case

Figure 3.3: Dependence networks for normal and cancer cases in ovarian cancer dataset. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features.

3.3 Biomarker Identification

In a certain disease, biomarkers are defined as the alternations of patterns at the cellular, molecular or genetic level. These biomarkers normally serve as the indicators of diseases. Biomarker identification is a topic of great importance, because it provides new insights into the early detection, diagnosis and treatments of cancer. In this section, we study and compare two biomarker identification criteria derived from the dependence model and network: the classification-performance-based criterion and the dependence-network-based criterion. The two criteria are applied to both gene and protein data to identify biomarkers.

3.3.1 Dependence-Network-Based Biomarkers

As described earlier, given a dataset containing cancer and normal samples, dependence networks can be constructed for both normal and cancer cases. For each case, we examine three features at one time. Through an exhaustive search, the de-

pendence relationship of all feature triples are examined to find binding triples. From normal samples, the binding triples of normal case are found, and we build a dependence network for the normal case D_{normal}. From cancer samples, the binding triples of cancer case are found, and we build a dependence network for the cancer case D_{cancer}. By examining the norm of all the columns of the matrix $D_{normal} - D_{cancer}$, we are able to see which features go through a large topology change from normal to cancer, and identify them as dependence-network-based biomarkers. Therefore, the basic idea behind the dependence-network-based criterion is that: from normal to cancer, the features with large topology change are identified as biomarkers.

3.3.2 Classification-Performance-Based Biomarkers

In the literature, one popular biomarker identification criterion is the classification-performance-based criterion. For the classification-performance-based criterion, features are examined three at one time. A dependence-model-based classifier is build upon the three features to examine their classification power. Through an exhaustive search, the classification performances of all possible feature triples are examined. Triples with classification accuracy higher than 95% are considered to be informative triples. And, features frequently appear in the informative triples are regarded as important cancer biomarkers. These biomarkers are called the classification classification-performance-based biomarkers. The basic idea behind the classification-performance-based criterion is that: the features that have high discrimination power are identified as biomarkers.

3.4 Biomarker Identification Results

In the previous section, we propose the dependence-network-based biomarker identification criterion, and introduce the classification-performance-based criterion in the existing literature. In this section, we compare the two criteria, and show the superiority of the dependence-network-based criterion.

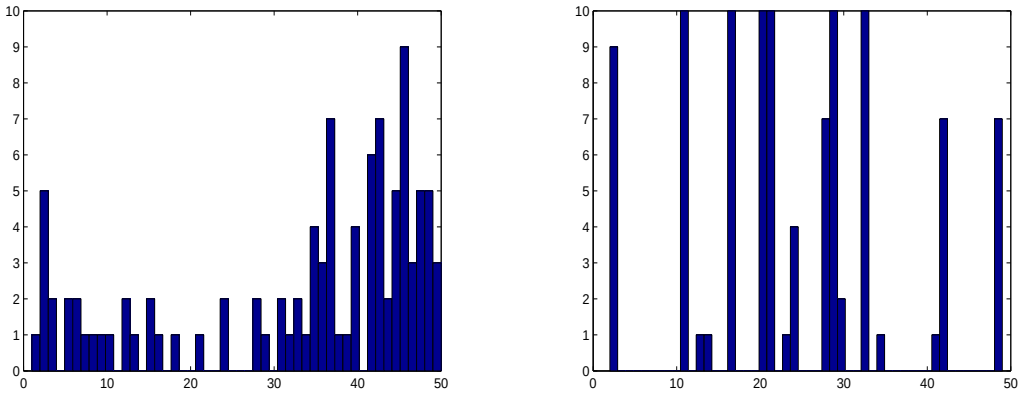
To assess the reproducibility of the identified biomarkers, we apply the strategy similar with 10-fold cross-validation, where the entire dataset is divided into 10 parts; 9 parts are used for model learning (training) and the one left is used for validation (testing). In each of the 10 iterations, we search for biomarkers based on different choices of training and testing samples. The desired reproducibility means that the same biomarkers are consistently identified in the 10 iterations. Without reproducibility, the results from a particular method and particular dataset can not be generalized to other datasets, and the validity of both the results and the method are questionable. Therefore, a consistent and reproducible result is a necessary condition of a successful method. In the following, we show the superior reproducibility of the dependence-network-based criterion in several gene and protein datasets.

There are three protein mass spectrum datasets and two microarray gene datasets under investigation. The protein datasets are: an ovarian cancer dataset, with 25 normal samples and 24 cancer samples [65], a prostate cancer dataset, with 81 normal samples, 84 early stage cancer samples and 84 late stage cancer samples [66], and a liver cancer dataset, with 176 cancer samples and 181 normal samples [14]. The gene datasets are, a gastric cancer microarray dataset [46] and a liver

cancer microarray dataset [47]. For the protein MS datasets, proper pre-processing is needed to convert the spectra data into m/z peak features. The pre-processing is the same as that in Chapter 2, with details presented in Appendix B. Because of the computational complexity of the exhaustive search, we first apply T-test to perform feature selection, and we limit our attention to the top 50 ranking features in T-test. Biomarkers are identified from these 50 candidate features.

Ovarian cancer protein MS dataset

First, we examine the ovarian cancer MS dataset. For the classification-performance-based criterion, the dataset is divided into 10 parts; 9 parts are used for model learning (training) and the one left is used for validation (testing). In each of the 10 iterations, based on different choices of training and testing samples, we search for classification-performance-based biomarkers. In each iteration, through an exhaustive search, the classification performance of all possible feature triples are examined to find informative triples, and the top 10 most frequently appeared features are considered as biomarkers. Therefore, in each of the 10 iterations, based on different training and testing set, 10 biomarkers are identified. We examine the biomarkers identified by different subsets of the whole dataset to assess the consistency of the identification criterion. The result is that, only 3 features are commonly identified as biomarkers by 7 or more out of the 10 iterations. They are features $37, 43, 46$. Figure 3.4(a) shows the histogram of the identified biomarkers, where the horizontal axis is the feature index, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. From the widely spread histogram, we can conclude that the result is not quite consistent.



(a) (b)

Figure 3.4: Fig (a) is the histogram of classification-performance-based biomarkers in the ovarian cancer MS dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the ovarian cancer MS dataset. In both figures, the horizontal axis is the feature index, and the vertical axis shows how many times one feature is identified during the 10-fold iterations.

For the dependence-network-based criterion, we also apply 10-fold cross validation. In each iteration, based on different subset of the whole datasets, D_{normal} and D_{cancer} are constructed and compared. By examining the norm of all the columns of the matrix $D_{normal} - D_{cancer}$, we are able to see which features go through a large topology change from normal to cancer, and identify them as biomarkers. In each iteration, we identify 10 features with large topology changes as biomarkers. The result shows that, 10 features are commonly identified as biomarkers by 7 or more out of the 10 iterations. They are features $2, 11, 17, 20, 21, 28, 29, 33, 42, 49$. Figure 3.4(b) shows histogram of the identified biomarkers. From this figure, we can see that the dependence-network-based criterion yields much more consistent results, compared with the classification-performance-based criterion. Another observation is that, if we apply a simple differential method for biomarker identification, such as T-test, the identified biomarkers will be features with indexes 40$\sim$50 (since the

pre-selection 50 features are based on T-test). From Figure 3.4, we can see that the classification-performance-based biomarkers have high correlation with the simple differential method. However, the dependence-network-based criterion identifies many biomarkers that are not simply the most differentially expressed features. The results indicate that, the dependence-network-based biomarker identification criterion yields much more information than the simple differential method and the classification-performance-based criterion.

In Figure 3.3, the dependence networks for normal and cancer cases are drawn, where we can see the important features in the normal and cancer dependence networks through visual inspection. In the normal case, features 11, 17, 20, 29, 30 and 33 are important core features. They have rich dependence relationships with lots of other features. However, in the cancer case, there are more core features 2, 9, 11, 17, 20, 21, 28, and 42. From normal case to cancer case, some unimportant features in normal case become core features in cancer case, especially features 2, 21, 28 and 42; while some core features in normal case become deactivated in cancer case, such as features 29, 30 and 33. These core features are strongly suggested to be biomarkers in ovarian cancer. It is our intention to investigate the origin and identity of these features.

Prostate cancer protein MS dataset

We further examine the prostate MS dataset for two cases: normal samples vs early stage cancer samples, and normal samples vs late stage cancer samples. Our main purpose in analyzing this dataset is to examine the possible difference between dominant biomarkers in early cancer stage and late cancer stage.

The two biomarker identification criteria are examined under 10-fold cross validation for two tasks, normal vs early stage cancer, and normal vs late stage cancer. The classification-performance-criterion is applied to identify biomarkers for early stage cancer and late stage cancer, respectively. Regarding the dependence-network-based criterion, since there are three stages of data in this dataset, we can build three dependence networks, one for each stage, D_{normal}, D_{early} and D_{late}. Based on D_{normal} and D_{early}, we identify biomarkers for early stage cancer samples; based on D_{normal} and D_{late}, we identify biomarkers for late stage cancer samples.

The histograms of the identified biomarkers from two criteria for two tasks are shown in Figure 3.5. Figure 3.5(a) shows the histogram of the classification-performance-based biomarkers for the task of normal vs early stage cancer. Figure 3.5(b) shows the histogram of the dependence-network-based biomarkers for the task of normal vs early stage cancer. For the other task, normal vs late stage cancer, Figure 3.5(c) and Figure 3.5(d) show the histograms of the classification-performance-based biomarkers and dependence-network-based biomarkers respectively. Consistent with the results in the ovarian cancer dataset, the dependence-network-based criterion gives more consistent results for both early stage cancer and late stage cancer than the classification-performance-based criterion. Again, it is observed that, the dependence-network-based criterion yields more information than the classification-performance-based criterion and a simple differential method, such as T-test.

The dependence networks for normal, early cancer stage and late cancer stage are drawn in Figure 3.6. From this figure, we can see some interesting behaviors of

the identified dependence-network-based biomarkers through visual inspection. For example, feature 34 is not important in normal stage. However, in cancer stages, it plays a more important role in the dependence network. Features 20 and 24 are more interesting. They are important network nodes in both normal stage and late cancer stage. However, they are deactivated in early cancer stage. Features 12, 13 and 16 behave oppositely: they are activated in early cancer stage only. These features might be the key to early stage of cancer development, and deserve to be further investigated.

Liver cancer protein MS dataset

We also examined a liver cancer protein MS dataset. Similar with the above analysis, 10-fold cross-validation is applied to compare the classification-performance-based criterion and the dependence-network-based criterion. The histograms are shown in Figure 3.7. From the results, we again observe the superiority of the dependence-network-based criterion over the classification-performance-based criterion.

In Figure 3.8, dependence networks of normal and cancer cases are shown. We see that the difference between normal and cancer is not as obvious as the previous examples. The connection among several core features are almost unchanged for both cases. However, when going into the details by examining the adjacency matrixes D_{normal} and D_{cancer}, we are still able to identify biomarkers and yield consistent results.

Gastric cancer gene microarray dataset

For the gastric cancer microarray dataset, in order to identify the classification-

performance-based biomarkers, we exhaustively examine all possible feature triples, and apply the dependence model for classification. Triples with classification accuracy higher than 95% are considered to be informative triples. Gene features that frequently appear in the informative triples are regarded as cancer biomarkers. 10-fold cross-validation is applied to examine the consistency of the identified biomarkers. In each of the 10 iterations, 10 biomarkers are identified based on different training and testing sets. The histogram of the identified biomarkers are shown in Figure 3.9(a). Only 1 feature is commonly identified as biomarkers by 7 or more out of the 10 iterations. The widely spread histogram shows the lack of consistency of classification-performance-based criterion in the gastric gene microarray data.

The dependence-network-based criterion is also examined under 10-fold cross-validation. For each of the 10 iterations, from a subset of normal samples, we build a dependence network for normal case D_{normal}; from a subset of cancer samples, we build a dependence network for cancer case D_{cancer}; then, biomarkers are identified based on the difference between D_{normal} and D_{cancer}. As shown in Fig.3.9(b), 7 features are commonly identified as biomarkers by 7 out of the 10 iterations. They are features $10, 26, 31, 37, 41, 42, 50$. From this figure, we again observe that the dependence-network-based criterion yields much more consistent results than the classification-performance-based criterion. Also, the dependence-network-based criterion yields more information than the classification-performance-based criterion, with respect to the simple differential method T-test. Compared with the results from protein MS data, the results from gene microarray data exhibit less consistency. This may be because gene microarray experiments have larger noise than the

protein MS experiments.

From Figure 3.10, we can see the important features in the normal and cancer dependence networks through visual inspection. In the normal case, features 8 and 50 are important core features. However, in the cancer case, there are much more core features 10, 26, 31, 37, 41 and 42. From normal to cancer, some unimportant features in the normal case become core features in the cancer case, while some core features in the normal case become deactivated in the cancer case. These gene features are strongly suggested to be biomarkers in gastric cancer.

Liver cancer gene microarray dataset

Finally, we analyzed a liver cancer gene microarray dataset. The histograms of the classification-performance-based criterion and the dependence-network-based criterion are shown in Figure 3.11. The result is consistent with that of the other datasets.

In Figure 3.12, the dependence networks for normal and cancer cases are drawn for this liver gene dataset. We can see that the important features in the normal and cancer dependence networks are quite different. Those important feature are potentially important biomarkers.

3.5 Biological Relevant of the Identified Biomarkers

In the previous section, we have presented the classification-performance-based biomarkers and dependence-network-based biomarkers from several gene and protein datasets. Since consistency and reproducibility is a necessary condition of a successful method, and the classification-performance-based criterion does not meet the necessary condition, so the classification-performance-based criterion is not a successful one. One the other hand, the proposed dependence-network-based criterion survives the necessary condition. In this section, we will further illustrate the merit of the dependence-network-based criterion, by analyzing its biological relevance.

This section in contributed by our collaborators in Georgetown University. The reason I include this part in the thesis is to make the argument more complete, by including support and validation from the biology point of view.

Take the gastric cancer gene microarray dataset for examples. 7 gene biomarkers are identified from 50 candidates. The 50 top-score genes we analyzed represent the most significant changes of gene expression patterns across different cancer pathological types, and correspond to four distinct gene clusters in the hierarchical clustering result [46]. Table 3.1 summarizes the 7 identified gene biomarkers, 6 with significantly increased expression levels and 1 with decreased expression. Interestingly, the six up-regulated genes all correspond to the same ECM (extracellular matrix) cluster, which has highly similar expression pattern across most pathological types. The down-regulated SIDT2 gene, on the other hand, belongs to a cluster

Gene Name	Protein Name [UniPortKB Accession]*	Feature (Node) #	Expression Level in Cancer Samples
SPARC	Osteonectin, SPARC precursor [P09486]	42	Up
COL3A1	Type III collagen alpha-1 chain precursor [P02461]	26	Up
SULF1	Extracellular sulfatase Sulf-1 precursor [Q8IWU6]	50	Up
YARS	Tyrosyl-tRNA synthetase, cytoplasmic (TyrRS) [P54577]	10	Up
ABCA5	ATP-binding cassette A5 [Q8WWZ7/Q9NY14]	41	Up
THY1	Thy-1 membrane glycoprotein precursor [P04216]	31	Up
SIDT2	SID1 transmembrane family member 2 precursor [Q8NBJ9]	37	Down

Table 3.1: Identified biomarkers based on dependence network modeling for gastric cancer. The marker genes are mapped to the protein accession numbers in UniProt Knowledgebase (UniProtKB) [67]

with no assigned function.

The ECM cluster of genes, including many that encode extracellular matrix components, tends to be more highly expressed in tumors of the diffuse histological type than in those of the intestinal type. This is consistent with greater propensity of this group of tumors for invasive growth, often provoking a dense fibrous reaction, and a reflection of reciprocal interactions between tumor and stromal cells that play important roles in tumor biology [46]. In fact, three of the six biomarker genes we identified (SPARC, COL3A1, and THY1) encode proteins of extracellular matrix component or of mediating cell-matrix interactions. In addition, SULF1 and YARS are either extracellular sulfatase or secreted cytokine and both are implicated in tumor growth and progression.

Osteonectin, also known as SPARC, is a non-structural component of extra-

cellular matrix-associated matricellular glycoprotein. Matricellular proteins mediate interactions between cells and their extracellular environment. Osteonectin is involved in the regulation of tumor cell growth, differentiation, and metastasis. It is produced at high levels in many types of cancers, especially by cells associated with tumor stroma and vasculature [68]. Osteonectin was suggested as a prognostic marker for several cancers, including invasive differentiated stomach adenocarcinoma [69], gastric cancer [70], and malignant melanoma [71], and was correlated with metastasis in prostate cancer [72]. Furthermore, osteonectin and type III collagen alpha-1, another marker gene predicted by the dependence network, were highly expressed in gastric cancer tissue [73]. Marked increases in expression of osteonectin and six other extracellular matrix proteins, including collagen type III, were also observed in rat gastric cancer models [74].

SULF1 is an extracellular endosulfatase that desulfates cell surface heparan sulfate proteoglycans (HSPG), thus regulating the cellular signaling cascades. Dynamic regulation of HSPGs by sulfatases within the tumor microenvironment can have a dramatic impact on the growth and progression of malignant cells. SULF1 has been implicated in promoting cell proliferation in bladder cancer and repression of differentiation in the muscle-invasive tumors, and was suggested as one of the top predictors for the bladder cancer outcome [75]. SULF1 was also shown to inhibit tumor growth in hepatocellular carcinoma [76].

The human tyrosyl-tRNA synthetase (TyrRS) is a synthase that produces two distinct cytokines from the N- and C-terminal fragments [77]. It may be involved in a coordinated mechanism for regulating angiogenesis with a related synthetase,

tryptophanyl-tRNA synthetase (TrpRS), which also generates two fragments in a similar fashion. While fragments of TyrRS stimulate angiogenesis, those of TrpRS inhibit this process [78]. TyrRS and TrpRS are proinflammatory cytokines with multiple activities during apoptosis, angiogenesis and inflammation. They also play important roles in cancer progression, modulating tumor angiogenesis and its escape from surveillance by immune system [79].

ABCA5 is a transmembrane protein in the ABC transporter family, and has been shown to reside in lysosomes. ABCA5 gene knockout mice develop lysosomal disease-like symptoms [80]. ABCA5 was also identified as a tissue and urine diagnostic marker for prostate intraepithelial neoplasia.

Thy-1 (CD90) is a small GPI-anchored protein abundant on the surface of mouse thymocytes and peripheral T cells. Thy-1 is involved in the maintenance of T cell homeostasis in the absence of TCR triggering, as well as potentiating antigen-induced T cell responses induced through TCR [81]. Thy-1 is also an important regulator of cell-cell and cell-matrix interactions, with important roles in nerve regeneration, metastasis, inflammation, and fibrosis [82].

The only down-regulated marker gene is SIDT2, which is a cell membrane protein that enhances cell uptake of small interfering RNA (siRNA) [83], resulting in increased siRNA-mediated gene silencing efficacy. However, its cellular functions and roles in cancer are unclear. As a central node in the dependence network (node 37 in Fig. 3.10), the cellular functions and roles of SIDT2 in gastric cancer are worth further investigation.

Taken together, the 7 gastric cancer biomarker genes that are consistently

identified by the dependence network modeling approach have been shown to be biologically relevant in gastric and other cancers. Of special note is that both SPARC and COL3A1 are concurrently observed in this study (as connected core nodes 42 and 26 in Fig. 3.10) as well as in several other studies as valuable biomarkers for gastric cancers. We therefore conclude that our network modeling approach have provided a novel and consistent mathematic model to define potential cancer biomarkers, which imply functional associations or interactions that are important for the underlying cancer biology.

3.6 Chapter Summary

In Chapter 2, we develop the dependence model and apply it to examine the big picture, which is the ensemble dependence among groups of genes. In this chapter, we zoom in to examine the dependence relationship among individual features. The dependence network is constructed, with the individual features being the nodes of the network. The connections of the dependence network is constructed based on the eigenvalue pattern of the dependence model. From the results of the gene microarray datasets and the protein MS datasets, we can see clear difference between the dependence networks for normal and cancer cases. Biomarkers can be identified based on the difference between dependence networks for normal and cancer cases. And the biomarkers are proved to be consistent, reproducible and relevant from the biology point of view. Details of this chapter is published in [84, 85].

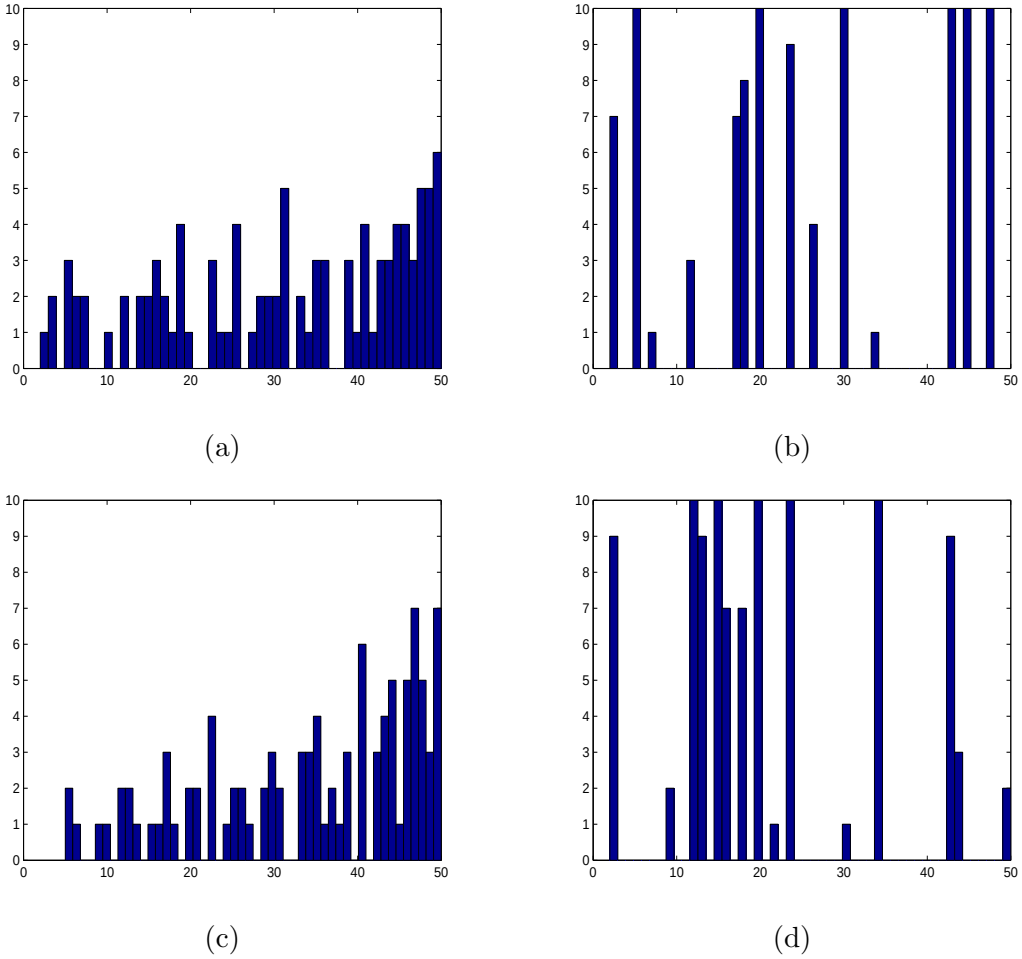
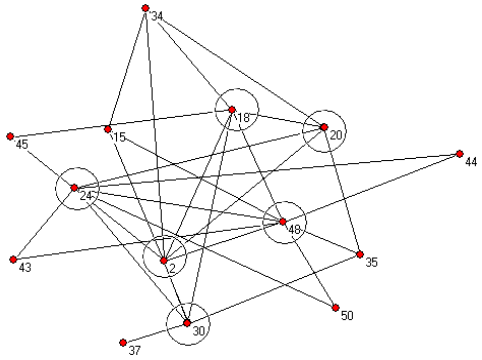
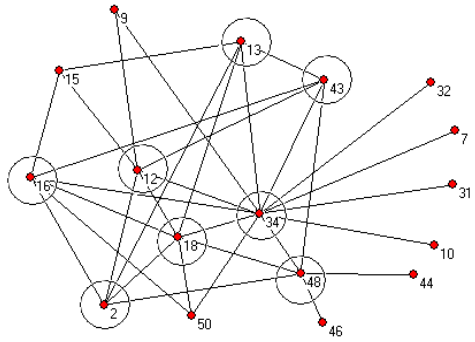


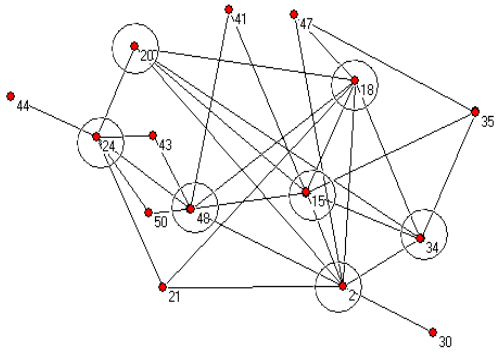
Figure 3.5: Fig (a) shows the histogram of the performance-based biomarkers for the task of normal vs early stage cancer. Fig (b) shows the histogram of the network-based biomarkers for the task of normal vs early stage cancer. Fig (c) shows the histogram of the performance-based biomarkers for the task of normal vs late stage cancer. Fig (d) shows the histogram of the network-based biomarkers for the task of normal vs late stage cancer.



(a) Normal case

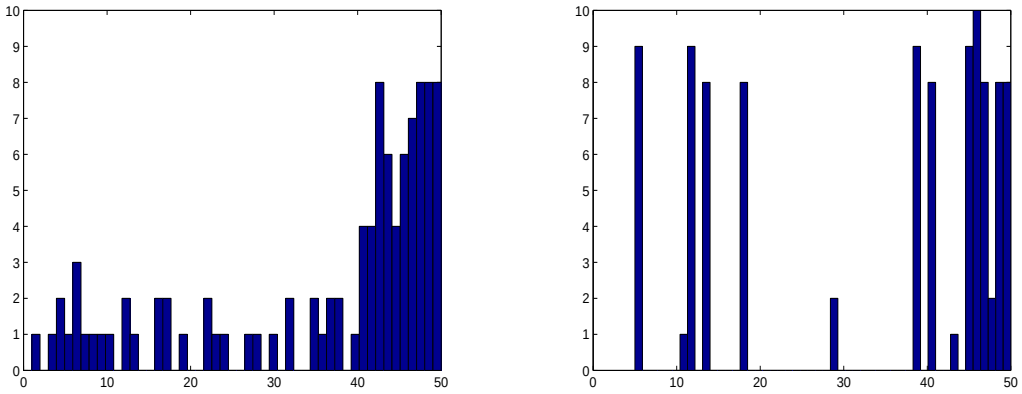


(b) Early cancer case



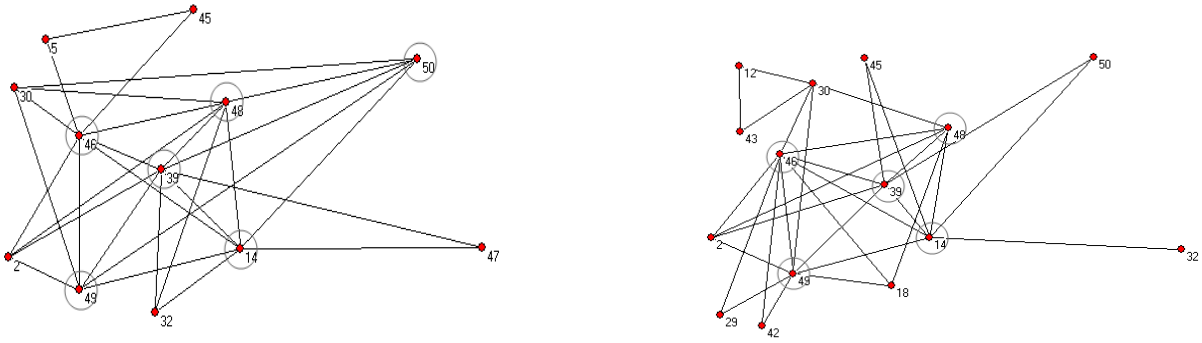
(c) Late cancer case

Figure 3.6: Dependence networks for the prostate cancer dataset: normal, early and late cancer cases. The circles are used to indicate the core features, which are identified through visual inspection.



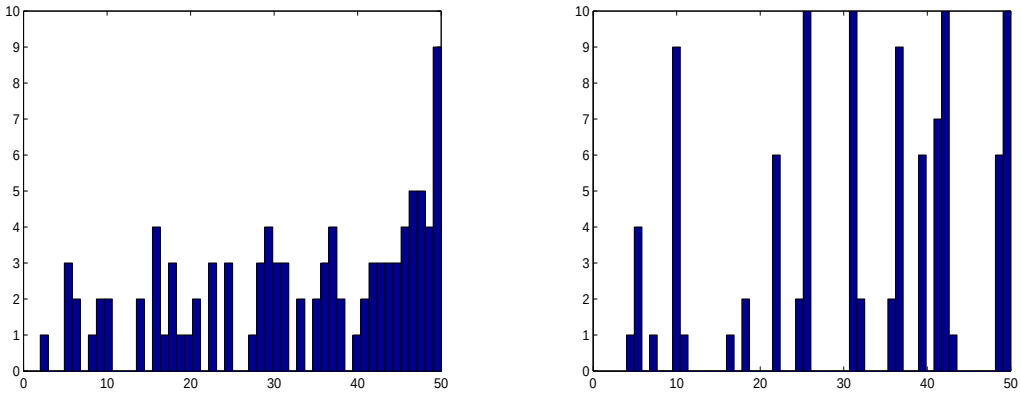
(a) (b)

Figure 3.7: Fig (a) is the histogram of the classification-performance-based biomarkers in the liver cancer MS dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the liver cancer MS dataset.



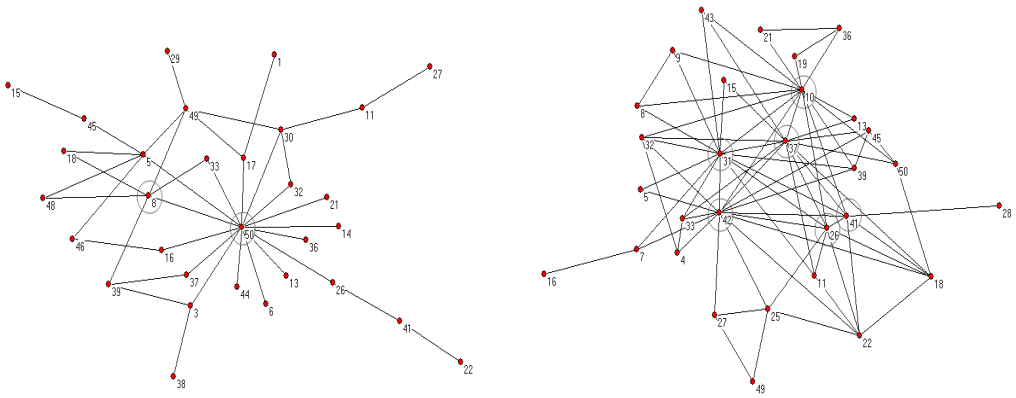
(a) Normal case (b) Cancer case

Figure 3.8: Dependence networks for normal and cancer cases in liver cancer MS dataset. The circles are used to indicate the core features, which are identified through visual inspection.



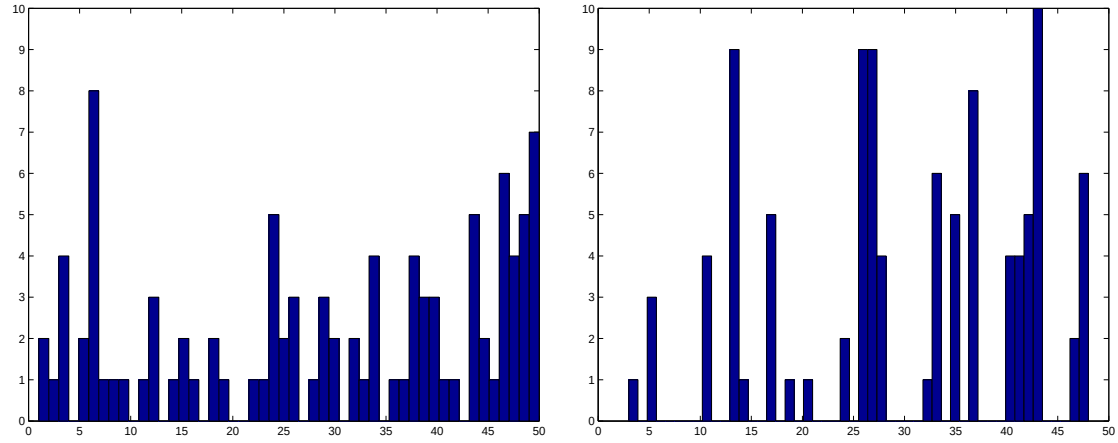
(a) (b)

Figure 3.9: Fig (a) is the histogram of the classification-performance-based biomarkers in the gastric cancer microarray dataset. Fig (b) is the histogram of the dependence-network-based biomarkers of the gastric cancer microarray datasets.



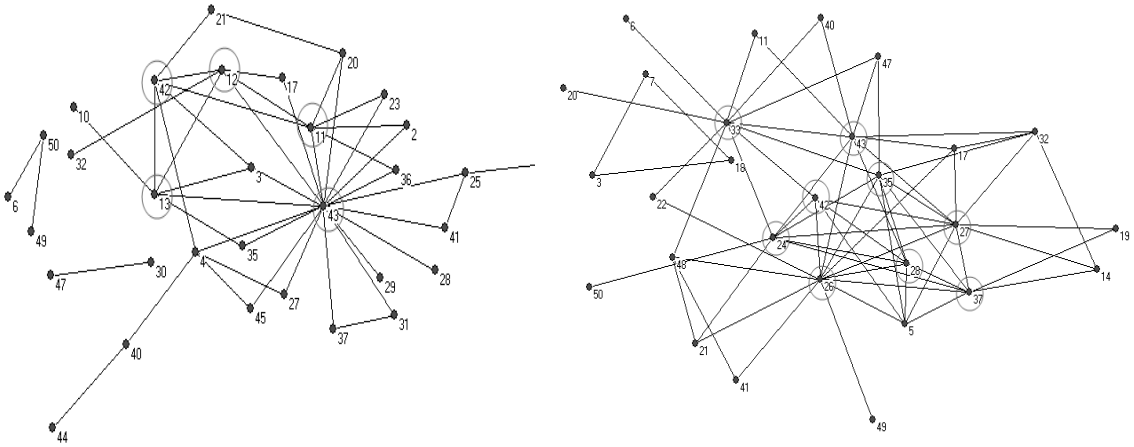
(a) Normal case (b) Cancer case

Figure 3.10: Dependence networks for normal and cancer cases in the gastric cancer microarray dataset. The circles are used to indicate the core features, which are obtained through visual inspection.



(a) (b)

Figure 3.11: Fig (a) is the histogram of the classification-performance-based biomarkers in the liver cancer microarray dataset. Fig (b) is the histogram of dependence-network-based biomarkers of the liver cancer microarray datasets.



(a) Normal case (b) Cancer case

Figure 3.12: Dependence networks for normal and cancer cases in the liver cancer microarray dataset. The circles are used to indicate the core features, which are identified through visual inspection.

Chapter 4

Resynchronization of Microarray Time-Series

4.1 Motivation

Besides the direct comparison between normal and cancer cases, the understanding of the fundamental cell-cycle system could also contribute a lot to cancer research. Starting from this chapter, we shift our attention to the analysis of time-series data of the cell-cycle system. The time-series data is obtained by measuring one sample at multiple time points during a certain biology process, such as cell-cycle. Therefore, from time-series data, we are able to see how the sample evolves along time. The first critical task in understanding the cell-cycle system is to identify the genes which are periodically expressed during the cell-cycle. In the current technologies, most expression data are measured based on a population of cells which are synchronized to exhibit similar behaviors [16]. However, even with the most

advanced synchronization method, maintaining a tightly synchronized population even over a couple of cycles is a challenging research issue, since continuous synchronization loss is gradually observed due to the diversity of individual cell growth rates [17]. Because of the synchronization loss, the gene expression data observed from a population of cells is different from the gene expression data of a single cell. Therefore, in addition to the noise effect on the measurements, a significant difficulty in identifying cell-cycle regulated genes from time-series microarray data arises from synchronization loss. Direct periodicity test on the expression measurements could be misleading or fail due to the fact that the expression measurements are contributed by a mixture of cell populations growing at different rates.

Several approaches for identifying cell-cycle regulated genes, when taking into consideration the issue of synchronization loss, have been proposed in the literature. They can be divided into two major categories, differentiated by the absence or presence of other complementary information besides gene expression data. Most studies in the literature belong to the former category, which relies solely on the expression data. Fourier analysis is employed for periodicity test in [17, 18, 19]. The authors present an exact statistical test to identify periodically expressed genes by distinguishing periodicity from random processes in [20]. In [21], a periodic-normal mixture (PNM) model is proposed to fit the transcription profiles of periodically expressed genes. In the second category, an algorithm combining budding index and gene expression data is recently proposed to deconvolve expression data in [22]. Regardless these developments, efforts are still needed to accurately identify cyclic genes and recover a more accurate single cell time-series expression compared with

the current expression measurements.

The goal of this chapter is to develop an efficient scheme to identify periodically expressed genes and reconstruct the underlying single cell gene expression profiles, by estimating the effect of synchronization loss. The main contributions of this chapter are two fold.

- We propose a synchronization loss model by representing the gene expression measurements as a superposition of different cell populations growing at different rates, because the model can mimic the synchronization loss observed in microarray experiments, and is easy to implement. Also, we develop a model-based estimation algorithm to reconstruct the underlying single cell gene expression profiles. In previous studies, the single cell expression profiles are often assumed to be sinusoids. However, the proposed algorithm does not require such assumption. It is able to handle a much larger variety of single cell expression profiles.

- Using the fitting residue error as criteria, we explore a supervised learning scheme to identify the cell-cycle regulated genes. The performance of the proposed scheme are examined via both simulations and real microarray time-series data of *Saccharomyces Cerevisiae*.

In the following, we start by introducing a synchronization loss model. After that, a cyclic gene identification scheme is proposed and applied on both simulated data and real microarray time-series data. The resulting identified cyclic genes are compared with two previous studies. From the results, we show that the proposed

scheme is promising in improving the quality of gene microarray time-series data.

4.2 System Model for Synchronization Loss

Even with the best currently available synchronization method, cells begin to lose their synchronization shortly, due to the diversity of individual cell growth rates. Therefore, we propose to model the observed gene expression data as a superposition from a mixed population of cells growing at slightly different rates, as

$$y_i(t) = \sum_{m=0}^{N} \beta_m x_i(\rho_m t), \tag{4.1}$$

where $y_i(t)$ is the observed expression of gene i at the time t; $x_i(t)$ is the underlying single cell expression profile; ρ_m represents the relative growth rates of cells with respect to standard cell-cycle; β_m represents the percentage of cells with a growth rate ρ_m, and it is assumed to be constant in one time-series of measurements. Although ρ_m can take continuous values in experiment, due to the limited size of microarray data, ρ_m is approximated by $N + 1$ components. Because of the different growth rates in experiment cell population, for a cell-cycle regulated gene, the measured expression may not exhibit clear periodicity. Therefore, it is difficult to accurately detect periodically expressed genes and distinguish them from non-periodically expressed genes based on the time-series microarray data.

Note that in equation (4.1), for gene i, from the underlying expression profile $x_i(t)$ to the observation $y_i(t)$, the distortion is dictated by β_m and ρ_m, which describe the synchronization loss status of the whole cell populations. Note that the distortion is the same for all the genes. We propose to utilize this common information of

all the genes to estimate the distortion and extract the underlying single cell gene expression profiles from the observations.

4.3 Polynomial-Model-Based Resynchronization

4.3.1 Inverse Formulation of Synchronization Loss Model

Since the underlying expression profile $x_i(t)$ is unknown, the right hand side of equation (4.1) is totally unknown. It is difficult to estimated $x_i(t)$ and the other parameters. Therefore, we propose to re-write equation (4.1) into the following form,

$$x_i(t) = \sum_{m=0}^{M} a_m y_i(c_m t) = [a_0, a_1, \cdots, a_M] \begin{bmatrix} y_i(c_0 t) \\ y_i(c_1 t) \\ \vdots \\ y_i(c_M t) \end{bmatrix}, \tag{4.2}$$

where the underlying single cell expression $x_i(t)$ is represented by the superposition of $M+1$ multiple scaled versions of the observation $y_i(t)$. Parameters a_m's and c_m's describe the coefficient and scaling factor of each component. An intuitive explanation for equation (4.2) is motivated by the inverse relationship between Finite Impulse Response (FIR) filters and Infinite Impulse Response (IIR) filters, since the structure of (4.1) is quite similar to that of FIR filters. Equation (4.1) describes an FIR-like operation which transforms $x_i(t)$ to $y_i(t)$. In order to perform the inverse transformation, an IIR-like operation is required. If the range of c_m is properly

chosen, equation (4.2) can be regarded as a truncated IIR-like operation, which is an approximate inverse of the FIR-like operation in equation (4.1). Therefore, equation (4.1) and equation (4.2) relates $x_i(t)$ and $y_i(t)$ in approximately the same way.

It is worth mentioning that the parameters a_m's and c_m's depend solely on β_m's and ρ_m's. They are common constants for all the genes. Thus, we propose to utilize this common property of all the genes to extract underlying single cell expression profiles.

Equations (4.1) and (4.2) are not mathematically equivalent in general. However, if $x_i(t)$ is polynomial, equations (4.1) and (4.2) can be equivalent. In this study, we are particularly interested in the case of polynomials, since polynomial is a common tool for data fitting [86]. Shown in the literature, polynomials are often successfully used to fit the time-series gene expression data [22].

Suppose $x_i(t)$ is a polynomial of order K such that

$$x_i(t) = \sum_{k=0}^{K} b_k t^k = [1, 1, \cdots, 1] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{4.3}$$

with b_k's being the polynomial coefficients. Then, according to equation (4.1), $y_i(t)$

can be expressed as,

$$y_i(t) = \sum_{m=0}^{N} \beta_m x_i(t\rho_m) = [\underline{\beta}^T \underline{\rho}^0, \underline{\beta}^T \underline{\rho}^1, \cdots, \underline{\beta}^T \underline{\rho}^K] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{4.4}$$

where $\underline{\beta} = [\beta_0, \beta_1, \cdots, \beta_N]^T$, and $\underline{\rho}^k = [\rho_0^k, \rho_1^k, \cdots, \rho_N^k]^T$. Similarly, since

$$y_i(ct) = [\underline{\beta}^T \underline{\rho}^0 c^0, \underline{\beta}^T \underline{\rho}^1 c^1, \cdots, \underline{\beta}^T \underline{\rho}^K c^K] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{4.5}$$

if we pick up multiple scaled version $y_i(c_m t)$ of the observation $y_i(t)$, we can write them together into the following matrix form,

$$\begin{bmatrix} y_i(c_0 t) \\ y_i(c_1 t) \\ \vdots \\ y_i(c_M t) \end{bmatrix} = \begin{bmatrix} \underline{\beta}^T \underline{\rho}^0 c_0^0 & \underline{\beta}^T \underline{\rho}^1 c_0^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_0^K \\ \underline{\beta}^T \underline{\rho}^0 c_1^0 & \underline{\beta}^T \underline{\rho}^1 c_1^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_1^K \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^T \underline{\rho}^0 c_M^0 & \underline{\beta}^T \underline{\rho}^1 c_M^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_M^K \end{bmatrix} \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}. \tag{4.6}$$

Now, if we want to find a set of coefficients a_m's to represent the underlying single cell expression profile $x_i(t)$ as in equation (4.2), based on equations (4.3) and (4.6), we will require coefficients a_m's to satisfy the following equation,

$$[a_0, a_1, \cdots, a_M] \begin{bmatrix} \underline{\beta}^T \underline{\rho}^0 c_0^0 & \underline{\beta}^T \underline{\rho}^1 c_0^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_0^K \\ \underline{\beta}^T \underline{\rho}^0 c_1^0 & \underline{\beta}^T \underline{\rho}^1 c_1^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_1^K \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^T \underline{\rho}^0 c_M^0 & \underline{\beta}^T \underline{\rho}^1 c_M^1 & \cdots & \underline{\beta}^T \underline{\rho}^K c_M^K \end{bmatrix} = [1, 1, \cdots, 1]. \tag{4.7}$$

Note that in the matrix in equation (4.7), every element in one column shares a common factor. If we pull out the common factor, the remaining part will be a Vandermonde matrix, as shown in equation (4.8).

$$\begin{bmatrix} \underline{\beta}^T\underline{\rho}^0 c_0^0 & \underline{\beta}^T\underline{\rho}^1 c_0^1 & \cdots & \underline{\beta}^T\underline{\rho}^K c_0^K \\ \underline{\beta}^T\underline{\rho}^0 c_1^0 & \underline{\beta}^T\underline{\rho}^1 c_1^1 & \cdots & \underline{\beta}^T\underline{\rho}^K c_1^K \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^T\underline{\rho}^0 c_M^0 & \underline{\beta}^T\underline{\rho}^1 c_M^1 & \cdots & \underline{\beta}^T\underline{\rho}^K c_M^K \end{bmatrix} =$$

$$\begin{bmatrix} c_0^0 & c_0^1 & \cdots & c_0^K \\ c_1^0 & c_1^1 & \cdots & c_1^K \\ \vdots & \vdots & \ddots & \vdots \\ c_M^0 & c_M^1 & \cdots & c_M^K \end{bmatrix} \begin{bmatrix} \underline{\beta}^T\underline{\rho}^0 & 0 & \cdots & 0 \\ 0 & \underline{\beta}^T\underline{\rho}^1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underline{\beta}^T\underline{\rho}^K \end{bmatrix}. \tag{4.8}$$

The Vandermonde matrix is of full rank, $min\{M, K\}$, as long as different scaled (c_m) observations are considered. We can show that, as long as M is greater than or equal to K, there exists at least one solution to equation (4.7). That is, there exists at least one set of coefficients a_m's that satisfies equation (4.7). In this case, equation (4.1) and equation (4.2) are mathematically equivalent.

In the above argument, the underlying expression $x_i(t)$ does not assume periodicity. However, in this study, the most interested expression signal is cell-cycle regulated, i.e. periodic,

$$x_i(t) = \sum_{k=0}^{K} b_k(t \ mod \ T)^k \tag{4.9}$$

where *mod* means the modulus operator that gives the reminders after division. In microarray time-series experiment, the range of relative growth rate ρ_m is not large.

With c_m carefully chosen, although periodic, the above argument holds for most of the cell-cycle data. In the following section of simulation results, we will demonstrate that, under such a periodic condition equation (4.2) is a fine approximation of the inverse of equation (4.1).

4.3.2 Estimation of Model Parameters

For cell-cycle regulated genes, because of the periodicity, $x_i(t) = x_i(t+T)$, from equation (4.2), the observations and parameters a_m's, c_m's are related as follows:

$$\sum_{m=0}^{M} a_m[y_i(c_m t) - y_i(c_m(t+T))] = 0. \tag{4.10}$$

Denote $\underline{y_i}(t) = [y_i(c_0 t) - y_i(c_0(t+T)), \cdots, y_i(c_M t) - y_i(c_M(t+T))]^T$, and $\underline{a} = [a_0, \cdots, a_M]^T$. Equation (4.10) can be re-written as,

$$\underline{y_i}(t)^T \underline{a} = 0. \tag{4.11}$$

Note that, we can evaluate equation (4.11) at different time points (as long as the time-series data allows). Also, all cell-cycle regulated genes satisfy equation (4.11). So, the estimation of a_m parameters can be formulated as a constrained least square problem,

$$[\underline{y_i}(t_1), \cdots, \underline{y_i}(t_n), \underline{y_j}(t_1), \cdots, \underline{y_j}(t_n), \cdots]^T \underline{a} = 0, \tag{4.12}$$

$$\text{subject to } \sum_{m=0}^{M} a_m = 1 \tag{4.13}$$

where genes $i, j, \cdots$ are cell-cycle regulated genes; $t_1, \cdots, t_n$ are the measurement time points that satisfies $(t_n+T)c_m < 2T$, for all $m = 0, \cdots, M$. Since in the current

Saccharomyces Cerevisiae time-series gene expression data, only two cell-cycles are available, the value of n in equation (4.12) is quite small, e.g. 4 or 5, depending on parameters c_m and the experiment sampling rate. Therefore it is important to use many cell-cycle regulated genes together to estimate the coefficients a_m's reliably. In this formulation, c_m are assumed known. Since in real experiment, the growth rate of different cells differ slightly, the range of the relative growth rate ρ_m is not large. In later simulations, we will show that, it is accurate enough to choose c_m to cover the range from 0.6 to 1.4. The constraint in equation (4.13) is chosen to avoid the trivial 0-vector solution, i.e. $\underline{a} = \underline{0}$.

4.3.3 Fitting Residue Criterion

After estimating a_m's, the model in (4.2) is used to reconstruct the underlying periodical component $x_i(t)$ for every gene. In order to detect cell-cycle regulated genes, a criterion is needed to answer the question whether the extracted signal is the underlying periodical expression profile of a cell-cycle regulated gene, or it is the periodical component from a non-cell-cycle regulated gene. We propose a criterion based on the model in (4.1), using the extracted periodical signal to fit the observations. The fitting residue will serve as the criterion in detecting cell-cycle regulated genes. For a particular gene, if the fitting residue is sufficiently small, compared with a threshold, then the extracted signal could lead to the measurements due to synchronization loss, which means the gene is highly likely to be cell-cycle regulated. On the other hand, if the fitting residue is large, then the extracted periodical signal

is likely to be the periodical component of a non-periodically expressed gene, which means the gene is more likely to be non-cell-cycle regulated. In the proposed identification schmem, the threshold of fitting residue is dynamically determined during iterations. Details are described in sections 4.3.4 and 4.5.

4.3.4 Cyclic Genes Identification Scheme

Based on the synchronization loss model and estimation approach described above, we further proceed to identify the cyclic genes. The scheme described here is a supervised learning scheme, since it requires an initial training set which consists of cell-cycle regulated genes previously identified by traditional biology experiments. Specifically, we propose an iterative framework to purify the training set and detect cyclic genes simultaneously. The main steps in the proposed iterative framework is described as follows:

1. Define initial training set as cell-cycle regulated genes previously identified by traditional methods.

2. Apply the proposed model on training set to estimate the parameters a_m's, and extract the underlying periodical signal for every gene in the training set.

3. Based on the extracted signal $x_i(t)$, fit it to the observation model in (4.1). According to the fitting residue criterion, remove some non-periodically expressed genes from the training set. Then, re-estimate the parameters a_m's using the training set and use the estimated a_m's to extract periodical signal for every gene in the testing set.

4. According to the fitting residue criterion, include some periodically expression genes into the training set. Then go back to Step 2.

Note that,under this framework, in order to purify the training set and detect the periodically expressed genes correctly, the criteria for removing and including genes in Step 2 and Step 4 should be carefully designed and fine tuned for each dataset. It is a difficult optimization problem. The proposed scheme, although heuristic in updating the training set, yields satisfactory results as will be demonstrated in section 4.5.

4.4 Simulation Results

In this section, we simulate time-series expression data with synchronization loss for both periodically expressed genes and non-periodically expressed genes. The proposed method is used to resynchronize the simulated data and identify periodically expressed genes. To evaluate the performance of the proposed method, we compare it with the methods studied in [16, 21]. We also perform sensitivity analysis to examine the robustness of the proposed method.

4.4.1 Simulations based on sinusoids

In this subsection, we simulate time-series expression data for 100 periodically expressed genes and 600 non-periodically expressed genes. The underlying single-cell periodical expression profile for cyclic gene i is generated by a linear combination

of 4 sinusoids with random phases,

$$x_i(t) = \sum_{j=1}^{4} \lambda_{ij} sin(\frac{2\pi j}{T} t + \phi_{ij}), \tag{4.14}$$

where the period T is set to be 60 minutes, same as the cell-cycle duration in the *alpha* experiment in [16]. The parameter λ_{ij} is randomly chosen, different for each gene. ϕ_{ij} represents the random phase, which is uniformly distributed on $[0, 2\pi)$. For the 600 non-cyclic genes, their underlying expressions are obtained through random permutations of expressions of cyclic genes.

For each gene, we simulate the synchronization loss by

$$y_i(t) = \beta_1 x_i(t * s) + \beta_2 x_i(t) + \beta_3 x_i(t * f) + v, \tag{4.15}$$

where $f = 1.3$ and $s = 0.7$ represent the relative growth rates. β_m is randomly generated, representing the percentage of cells growing at different rates. v represents the microarray measurement noise, which is modeled as a zero-mean Gaussian random variable. Its variance is chosen to make the signal to noise ratio (SNR) to be 5.716 dB, which is close to the SNR value estimated from the *alpha* dataset in [16]. Equation (4.15) is applied to all genes, representing the common synchronization status of the cell populations. In the simulations, measurements are taken every 6 minutes from 0 to 120 minutes, yielding 21 time points in total.

In the simulation, 50 cyclic genes are assumed known, in order to form the initial training set. The testing set contains the remaining 650 genes. For a particular choice of c_m, by applying the proposed model, a_m parameters are estimated, the underlying periodical signals for all genes are extracted, and the fitting residue criterion is examined.

range of c_m	avg fitting residue cyclic genes	avg fitting residue non-cyclic genes	diff
$[0.9, 1.1]$	0.4424	0.9373	0.4949
$[0.8, 1.2]$	0.4164	0.9560	0.5396
$[0.7, 1.3]$	0.3993	0.9583	0.5590
$[0.6, 1.4]$	0.3917	0.9355	0.5437
$[0.5, 1.5]$	0.4052	0.9478	0.5426
$[0.4, 1.6]$	0.4354	0.9865	0.5511

Table 4.1: For the simulation based on sinusoids, comparison of the normalized average fitting residues for cyclic and non-cyclic genes.

The parameter M is set to be $M = 7$. As mentioned in Section 4.3.1, we need to choose M to be larger than or equal to K. With $M = 7$, the proposed method can handle all polynomials with $K \leq 7$. And we know, 7th order polynomials can generate a large variety of curves, with up to 6 peaks and valleys. We believe the current parameters-setting can sufficiently model gene expression profiles.

As mentioned earlier, c_m should be chosen properly, in order to extract underlying single cell expression profiles accurately. In table 4.1, different choices of c_m are examined. To ensure a fair comparison, with M set to be 7, the values of c_m are chosen to be uniformly spaced in tested range. In the fitting residue criterion, ρ_m is set to be $[0.7, 0.8, \cdots, 1.3]$. From table 4.1, we can see that, different choice of c_m leads to different fitting residues for both cyclic genes and non-cyclic genes. As the range of c_m increases, the fitting residues for cyclic genes tend to decrease first, and then increase. This observation can be intuitively explained by the trade-off between errors due to the model-complexity and the data size. First, from the implication of FIR and IIR filters, the larger range of c_m considered, the smaller truncation error there will be. However, if the range of c_m is too large, due to the limited size of time series data, the number of available time points n in equation (4.12) will be small.

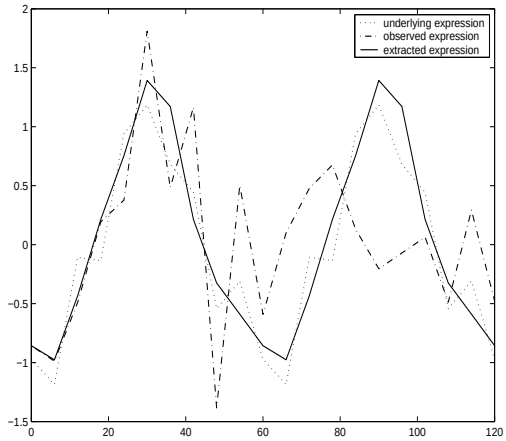
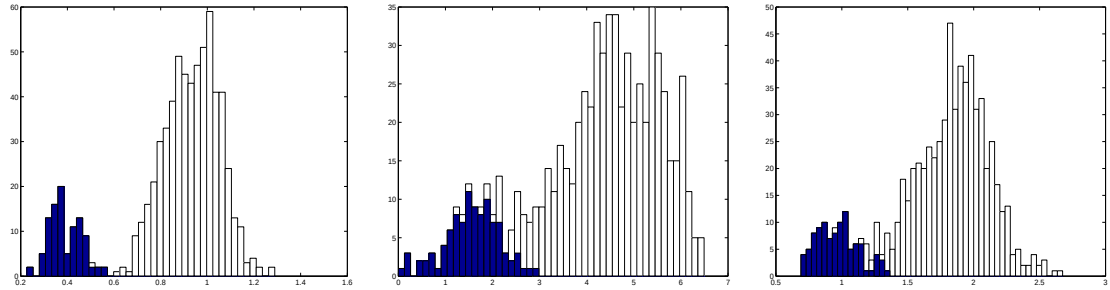


Figure 4.1: For an example of a simulated gene: the simulated sinusoid underlying periodical expression, experiment observation and extracted expression.

Less training data will cause the fitting residues increase. Therefore, based on Table 4.1, we choose the range of c_m to be $[0.6, 1.4]$, since with this choice, the average fitting residue for cyclic genes is small and the difference between cyclic genes and non-cyclic genes is large.

After determining the choice of c_m's, the proposed model is applied to estimate parameters a_m's based on the training set, and extract the underlying periodical signals for genes in the training set. Figure 4.1 gives a typical example of genes in the training set. Although there is clear difference between the underlying periodical expression and the simulated observation, based on the proposed method, the extracted expression is quite similar to the underlying periodical expression.

Based on the a_m and c_m parameters, the proposed model is applied to extract periodical signal components for all genes, and the fitting residue criterion is examined. In Figure 4.2(a), the histogram of all genes' fitting residues is shown, where the shaded part corresponds to the 100 cyclic genes. We can see that the cyclic genes have smaller fitting residues, while non-cyclic genes yield larger fitting



(a) proposed method (b) method in [16] (c) method in [21]

Figure 4.2: For the simulated data based on sinusoids: Fig (a) show the histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. Fig (b) is the result of the Fourier analysis used in [16]. Fig (c) shows the upper bound of results from method in [21].

residues. Therefore, this clear separation between these two groups of genes leads to the accurate identifications of cyclic genes.

In order to examine the identification performance of the proposed method, we compare it with two previous works [16, 21], by applying them to the same simulated time series data. In [16], Fourier analysis is applied to calculate the energy of the periodical components for each gene. The energy serves as a metric to identify cyclic genes. From Figure 4.2(b), we can see that, this method can identify cyclic genes with small outage. However, its performance is worse than that of the proposed method. In [21], a periodic-normal mixture (PNM) model is proposed, where a probabilistic (Gaussian) distribution and Fourier analysis are combined to model the synchronization loss. In [21], before identifying cyclic genes, the parameters of the Gaussian distribution have to be estimated. In our implementation, we skip the parameter estimation step by feeding the actual parameter values into the PNM model. Therefore, Figure 4.2(c) shows the performance upper method in [21], which is close to that of the proposed proposed method. However, it is worth mentioning

Probability of detection	False positive of proposed method	False positive of [16]	False positive of [21]
0.75	0	0.0741	0.0132
0.80	0	0.0805	0.0123
0.85	0	0.1053	0.0116
0.90	0	0.1262	0.0217
0.95	0	0.1518	0.1121
1.00	0.01	0.2857	0.1597

Table 4.2: For the simulation based on sinusoids, we compare the proposed method and two previous studies. When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive.

that the PNM-based method is admittedly sensitive to the parameter estimation of the Gaussian distribution.

In Table 4.2, we present the results in Figure 4.2 in a more quantitative fashion. We employ the Neyman-Pearson framework in detection theory [87]. During comparison, we fix the probability of correctly detecting cyclic genes, and examine the probability of false positive of different methods. That is, under the condition that certain amount of cyclic genes are correctly detected, how many non-cyclic genes will be falsely detected as cyclic genes. From Table 4.2, we can see that, when fixing the probability of detection, the proposed method has much less false positives, compared with the two previous studies.

In this subsection, the time series are simulated with the underlying single cell expression $x_i(t)$ being sinusoids. Together with the fact that Fourier analysis is employed, both previous studies have nice performance in identifying cyclic genes. However, if the underlying signal is based on polynomials, the result could be different.

4.4.2 Simulation based on polynomials

In this subsection, we simulate time-series expression data based on polynomial models. Again, 100 cyclic genes and 600 non-cyclic genes are simulated. The underlying single cell periodical expression profile for cyclic gene i is generated by a polynomial of order $K = 6$,

$$x_i(t) = \sum_{k=0}^{K=6} a_k (t \ mod \ T)^k, \tag{4.16}$$

where the period T is set to be 60 minutes, same as the cell-cycle duration in the *alpha* experiments. The parameter a_k is randomly chosen in $[-1, 1]$, different for each gene. For the 600 non-cyclic genes, the underlying expressions are obtained through random permutations of the expressions of the cyclic genes.

For each gene, we simulate the synchronization loss by equation (4.15). All parameters are set to be the same as the previous subsection. 50 cyclic genes are assumed known, forming the training set. For a particular choice of c_m, by applying the proposed model, a_m parameters are estimated based on the training set, the underlying periodical signals for all genes are extracted, and the fitting residue criterion is examined. Again, M is set to be 7, and different choices of c_m are examined. From Table 4.3, similar result with the previous subsection is observed. We choose the range of c_m to be $[0.6, 1.4]$. Because the average fitting residue for cyclic genes is small, and the difference between cyclic genes and non-cyclic genes is large.

Figure 4.3 is a typical example of genes in the training set. We can see that the

range of c_m	avg fitting residue cyclic genes	avg fitting residue non-cyclic genes	diff
$[0.9, 1.1]$	0.2240	0.9439	0.7199
$[0.8, 1.2]$	0.2164	1.0267	0.8104
$[0.7, 1.3]$	0.2157	1.0256	0.8099
$[0.6, 1.4]$	0.2284	1.0567	0.8283
$[0.5, 1.5]$	0.2455	1.1299	0.8845
$[0.4, 1.6]$	0.3411	1.0531	0.7120

Table 4.3: For polynomial based simulation, we compare the normalized average fitting residues for cyclic and non-cyclic genes.

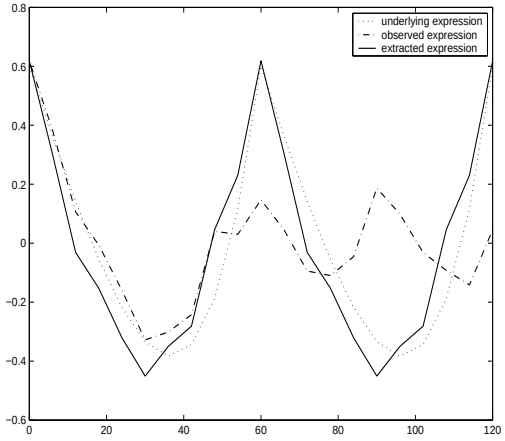
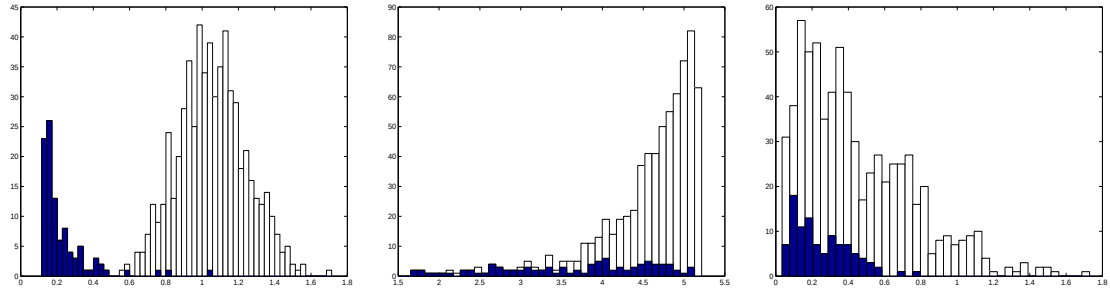


Figure 4.3: For an example of a simulated gene: the simulated polynomial underlying periodical expression, experiment observation and extracted expression.

simulated observations is quite different from the underlying periodical expression profile. Due to synchronization loss, the observed time-series does not exhibit a clear periodicity, especially in the second cycle. From poorly synchronized observations, the proposed method can successfully recover the underlying single cell periodical expression profile.

Based on the estimates of a_m's and c_m's, the proposed model is applied to extract periodical components for all genes, and the fitting residue criterion is examined. In Figure 4.4 (a), the histogram of residues shows that, the cyclic genes and non-cyclic genes are well separated, meaning that the proposed method can suc-



(a) proposed method (b) method in [16] (c) method in [21]

Figure 4.4: For the simulated data based on polynomials: Fig (a) show the histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. Fig (b) is the result of the Fourier analysis used in [16]. Fig (c) shows the upper bound of results from method in [21].

Probability of detection	False positive of proposed method	False positive of [16]	False positive of [21]
0.75	0	0.6622	0.7768
0.80	0	0.6887	0.7838
0.85	0	0.7028	0.7870
0.90	0	0.7443	0.7897
0.95	0	0.7765	0.7894
1.00	0.7375	0.8415	0.8353

Table 4.4: For the polynomial based simulation, we compare the proposed method and two previous studies. When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive.

cessfully identify the cyclic genes. The methods in [16] and [21] are also examined, with results shown in Figure 4.4 (b) and (c). From these figures, we note that both previous methods failed to separate cyclic and non-cyclic genes in the case when the underlying single cell expression profiles are polynomials. Similar with previous subsection, the result is shown in a more quantitative way, in Table 4.4. Easy to see, the proposed method out performs previous studies in the simulation based on polynomials. It is encouraging to see that the proposed method works well for a much larger variety of the underlying single cell expressions.

4.4.3 Sensitivity analysis

In our discussions so far, the standard cell-cycle duration T is assumed to be known as a prior knowledge. However, the cell-cycle duration may vary due to various environmental and experimental factors. In this subsection, we examine the performance of the proposed method when inexact prior knowledge of the cell-cycle duration T is considered.

The sensitivity analysis is conducted based on the simulated data by sinusoids. In the simulated data, the true cell-cycle length is $T = 60$. However, when applying the proposed method, we do not know the correct cell-cycle length. In Figure 4.5, we can see that, when the prior knowledge is inexact, the separation of fitting residues between cyclic and non-cyclic genes is not affected much. In Table 4.5, we quantitatively examine the sensitivity of the proposed method in terms of probability of detection and false positive. In Table 4.5, each row corresponds a certain requirement of probability of detection; each column corresponds to a case where certain value of T is taken as prior knowledge; and each element is the probability of false positive. From this table, as long as we do not require probability of detection to be extremely high (i.e., 100%), only when the prior knowledge is significantly different from the truth (i.e. the prior $T \leq 40$ or $T \geq 70$), will the performance degrade severely. This simulation result demonstrates the robustness of the proposed method with respect to the prior knowledge of cell-cycle duration.

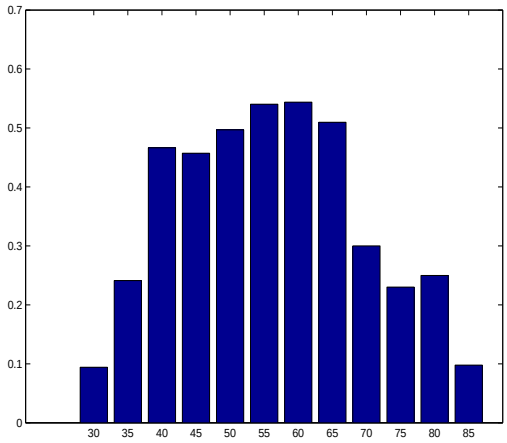
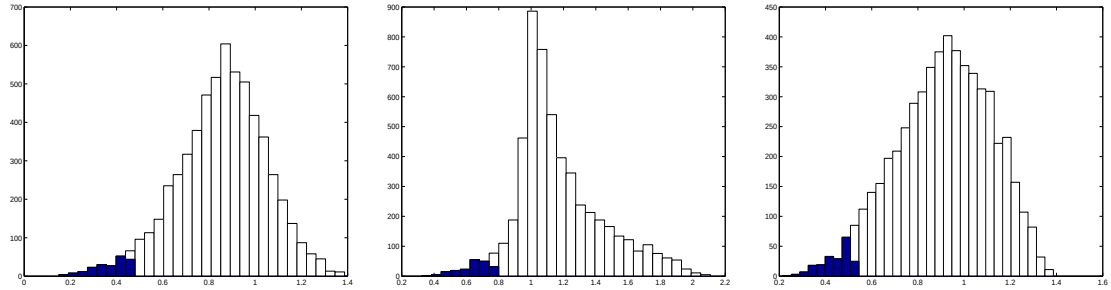


Figure 4.5: The horizontal axis is the prior knowledge of cell-cycle length, though it may not be the true cell-cycle length $T = 60$. The vertical axis is the difference of fitting residues between cyclic and non-cyclic genes.

PD \ T	35	40	45	50	55	60	65	70	75	80
0.75	0.10	0	0.01	0	0	0	0	0.05	0.12	0.16
0.80	0.10	0	0.01	0	0	0	0	0.05	0.17	0.15
0.85	0.12	0	0.01	0	0	0	0	0.08	0.18	0.18
0.90	0.18	0	0.01	0	0	0	0	0.11	0.24	0.24
0.95	0.38	0.01	0.01	0.02	0.01	0	0.01	0.14	0.29	0.25
1.00	0.60	0.09	0.12	0.05	0.01	0.01	0.01	0.22	0.39	0.52

Table 4.5: The performance sensitivity to inexact prior knowledge of cell-cycle length. When the probability of correctly detecting cyclic genes (PD) is fixed, we compare the probability of false positive, under different prior knowledge of cell-cycle T.



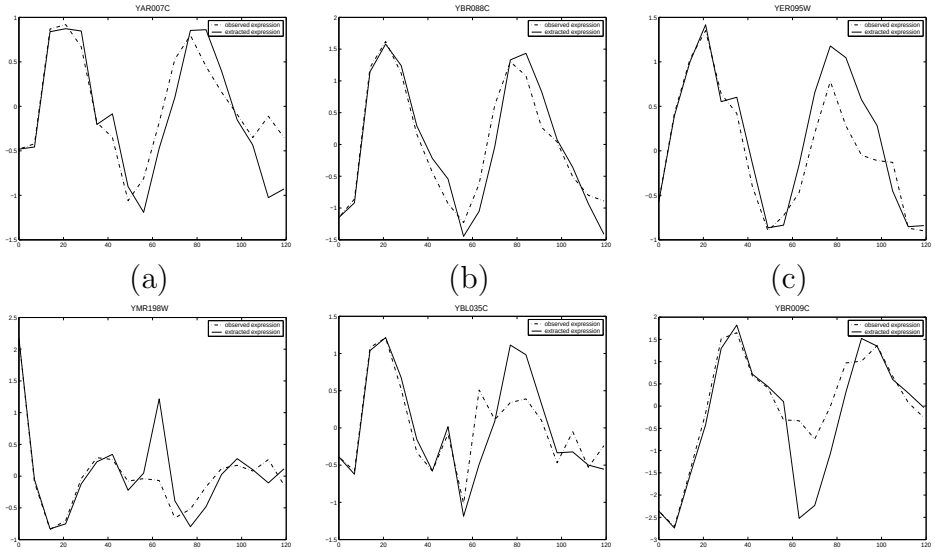
(a) alpha (b) cdc15 (c) cdc28

Figure 4.6: Histogram of fitting residues for the cdc28 dataset. Shaded part represents the histogram of fitting residues for the identified cyclic genes.

4.5 Results on Real Microarry Datasets

In this study, three microarray time-series datasets are investigated, alpha, cdc15 in [16] and cdc28 in [88]. From [16], 93 cell-cycle regulated genes previously identified by traditional methods are selected as initial training set. Since there is no guarantee that all those 93 genes will behave periodically in a particular experiment, we employ the iterative framework to purify the training set and identify cyclic genes simultaneously. During each iteration, we adopt simple removing and including criterion in step 2 and step 4. In step 2, the size of training set is reduced to half in order to purify the training set. In step 4, 200 genes with smallest fitting residues are included into the training set. In this way, we hope to purify the training set. Although the including and removing criteria are heuristic, the algorithm can converge within several iterations ($5 \sim 10$). The resulting histograms of fitting residues for the alfa, cdc15 and cdc28 datasets are shown in Figure 4.6, where the identified cyclic genes in the training set has small fitting residues.

Since both [16] and [21] identify about 800 cyclic genes, to make a fair comparison, we choose 800 genes with smallest fitting residues as identified cyclic genes. In



(a) (b) (c)

(d) (e) (f)

Figure 4.7: The experiment observed expression and the extracted periodical expression of genes identified in both the proposed scheme, the previous studies, and traditional methods. The title of each figure represents the gene's ORF name.

the identified cyclic genes, the intersection between the proposed method and [16] is 403; the intersection between proposed method and [21] is 433; the intersection between [16] and [21] is 541; the intersection among all three studies is 355. It is encouraging to see the large overlaps, an indication of consistency of the proposed method with the previous studies. In Figure 4.7, we show some examples of genes identified by both the proposed method, [16], and traditional experimental methods. Both the observed expression and extracted expression are shown. We can see that, for the cyclic genes that already exhibit periodical expression, the extracted expression is closed to experiment observed expression. And for the cyclic genes that do not exhibit periodical expression, the proposed method can recover the periodicity.

Although the genes identified by the proposed method have large overlap with those of the previous studies, it is interesting to examine the non-overlapping genes identified by the proposed method, but not identified in the previous studies, neither

[16] nor [21]. Some examples are shown in Figure 4.8. Since both previous studies rely on Fourier analysis, genes without clear periodicity may not be identified. However, the proposed method may be able to identify them, because synchronization loss is estimated and recovered. Besides technical improvement, we need to further investigate the biological relevance of the genes identified by the proposed method. One possible validation method is to validate the biological relevance of the identified cell-cycle genes by semantic analysis based on the gene ontology (GO) terms. To achieve this purpose, an online tool is applied, the SGD Gene Ontology Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder). We analyze the set of non-overlapping genes which are identified by one method, but not by the other two methods. The top GO terms associated with each method's results can be found in the Appendix D. For the proposed method, in the top 25 GO terms, there are several cell-cycle related terms, such as "M phase", "cell-cycle", "mitotic cell cycle", and "M phase of mitotic cell cycle". It suggests that some genes identified by the proposed method but not by the other two methods are cell-cycle related. For the sets of non-overlapping genes identified by the two reference methods, it is noted that none of the above four cell-cycle related GO terms appears in the top 25 GO terms. Details of top GO terms associated with results of each method can be found in the Appendix D. These encouraging observations demonstrate that the proposed method is promising for identifying cyclic genes.

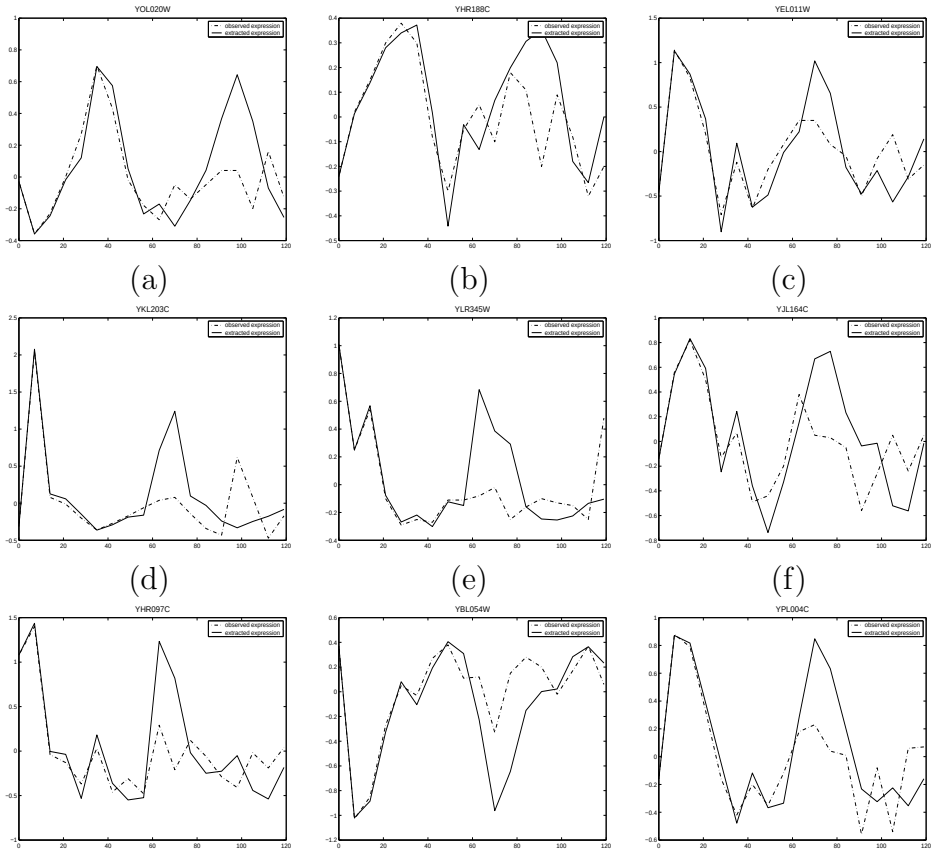


Figure 4.8: The experiment observed expression and the extracted periodical expression of genes identified in the proposed scheme, but not identified by previous studies. The title of each figure represents the gene's ORF name.

4.6 Chapter Summary

Synchronization loss is a major concern in identifying cyclic genes to understand the fundamental cell-cycle systems. To address the issue of synchronization loss, we consider a synchronization loss model where the gene time-series measurements are regarded as superpositions of mixed cell populations with different growth rates. We develop a polynomial-model-based framework to identify cell-cycle regulated genes and reconstruct the underlying gene expression profiles, which represent the single cell behavior more accurately. The proposed scheme is shown feasible and robust via simulations. Results from real microarray time-series data show that the proposed scheme is effective in reconstructing single cell time-series expression and identifying cell-cycle regulated genes. Moreover, the propose scheme removes the effect of synchronization loss, thus is a good pre-processing method that improves the quality of microarray time-series data. Details of this chapter is published in [89, 90]

Chapter 5

Discovering Regulator Network from Microarray Time-Series

5.1 Motivation

In Chapter 3, we discuss the dependence network, which is a co-regulation network. The connected genes in the dependence network are likely to be co-regulated (affected by the same factor). In this chapter, we will discuss a more useful network, the gene regulatory network (GRN). The gene regulatory network describes the complex relationship about which gene affects which gene, thus describes how a gene system evolve along time. Discovering and identifying such regulatory network will greatly improve our understanding of biological systems at the gene level. The knowledge of regulatory network will lead to the discovery of the signaling pathway of different biological processes and different diseases, which will greatly facilitate the development of effective drugs.

In the literature, there are many existing studies to infer a gene regulatory network based on microarray expression data. In [23], the boolean network is introduced to model the gene regulatory network as boolean relationship in combinatorial logic circuits. A boolean network is defined as $G(V, F)$ in [24]. It contains a set $V = \{x_1, x_2, ..., x_n\}$ of nodes representing genes, and a set $F = \{f_1, f_2, ..., f_n\}$ of boolean functions. In a boolean network, the gene expression data is quantized into two levels, binary values $\{0, 1\}$. Each boolean function f_i takes genes $x_{i1}, x_{i2}, ..., x_{ik}$ as input and generates the output as the expression level of x_i. Therefore, each function f_i describes how gene i is related with other genes in terms of combinatorial logic. Here, gene i is called the regulated gene, or the target gene; genes $x_{i1}, x_{i2}, ..., x_{ik}$ are called the regulators, or the predictors, or the parent genes. In boolean network, all genes are assumed to be updated synchronously. The dynamics of such a boolean network can be characterized as follows $x_i(t + 1) = f_i(x_{i1}(t), x_{i2}(t), ..., x_{ik}(t))$. Such system will eventually transitions into one of a number of attractor states, or transitions into periodical cycles among several attractors. In [25], the boolean network is extended to a probabilistic boolean network (PBN). In PBN, for each node, there are a number of associated logic functions that can predict the expression of this node. There is a probability distribution which describes how the logic functions compete to predict the node. Basically, PBN is a probabilistic mixture of several boolean networks. For the purpose of learning a PBN from expression data, an influence measure is proposed to determine the strength of regulatory relationship between a set of regulators and a regulated gene. In [26], the influence measure is applied to grow a regulatory network from a predetermined

seed.

The Bayesian network models the relationships among genes in terms of conditional probability distributions and joint probability distributions. Recently, Bayesian network has been used to analyze gene microarray data [27]. The Bayesian network models a gene regulatory network as a directed acyclic graph, where each vertex corresponds to a random variable (the expression of a gene). For each vertex, there is a conditional distribution, describing the probability of this random variable given its parent vertices. The Bayesian network approach has several limitations. For example, the Bayesian network has difficulty in determining directions, which is the regulator and which is the regulated gene; different time points are treated as different samples without utilizing the time information; and the Bayesian network assumes acyclic structure. To address these limitations, Dynamic Bayesian Network (DBN) is proposed to study the gene regulatory network in [28], followed by a number of studies [29, 30, 31, 32]. Different from Bayesian network, the conditional probabilities in DBN are the conditional probability of a regulated gene's expression at current time point given the expression of its regulators at a previous time point. In this way, the time information is incorporated and cyclic structures are allowed.

Differential equations are also used to model gene regulatory networks in the literature. In [33], the relationships among genes, mRNAs and proteins are modeled as differential equations. In [34, 35], differential equations are used to model the regulatory relationships among genes, and the parameters are computed through evolutionary programming. In [36, 37], maximum likelihood criterion is applied to determine the parameters of the differential equations. [38] proposed to model gene

regulatory network using stochastic differential equations. Similar with boolean network and Bayesian network, the differential equations also examine the relationship between a set of regulators and one regulated gene.

There are several other methods for inferring gene regulatory networks. [39] examines the relationship between several regulators and one regulated gene with different time lags, using a data-driven approach. [40] examines pairwise mutual information between gene pair's time-lagged expressions, and compare with a threshold determined by minimum description length (MDL) to infer existence of a connection. In [41], fuzzy logic is applied to model gene expression data.

There is a common property among existing methods, boolean network, Bayesian network, differential equations, etc. The relationship under investigation is always the relationship between one or several regulators and one regulated gene. To our knowledge, there is no method that directly examines the regulatory relationship between one or several regulators and several regulated genes. In our study, we will address this issue and provide a tool to examine several regulated genes together.

In Chapter 2 and Chapter 3, we propose the dependence model. We have shown that the dependence model and its eigenvalue pattern are consistent indicators that describe the group dependence behavior of several genes. In this chapter, we propose to use the eigenvalue pattern to infer regulatory relationships. We will infer the relationship between one regulator and a group of regulated genes from the relationship between the regulator and the regulated genes' group behavior (eigenvalue pattern). Therefore, we are able to examine regulatory relationships in a novel way, compared with the existing literature.

In the rest of this chapter, we will first take a detour to build a mathematical foundation for the dependence model, proving several properties of the eigenvalue pattern. After that , we show how a regulator can affect the regulated genes' eigenvalue pattern in several cases. Then, the proposed idea will be tested on cell-cycle microarray time-series data and compared with an existing method. Finally, some conclusion and summary will be presented.

5.2 Analytical Form of Eigenvalue Pattern of the Dependence Model

In previous chapters, the dependence model is applied for cancer classification, cancer prediction, and biomarker identification, where the results are closely related to the eigenvalue pattern of the dependence model. An interesting observation is that, the eigenvalues of the dependence model are always real-valued, from both simulated data and microarray experiment data. In this section, we will build a mathematical foundation for the dependence model and its eigenvalue pattern. We will mathematically prove that the eigenvalues of dependence model are always real, and given the analytical form of the eigenvalue pattern.

5.2.1 2-Dimensional Case

We start with a trivial 2-dimensional dependence model, where the dependence relationship between two genes are studied. Assume that we have two genes, whose

expression levels are random variables, denoted as x_1, x_2. We further assume the second order statistics of the two random variables are as follows, in equation (5.1).

$$\begin{cases} E[x_1^2] = \sigma_1^2 \\ E[x_2^2] = \sigma_2^2 \\ E[x_1 x_2] = \sigma_{12} \end{cases} \tag{5.1}$$

where $\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2$, because of the Cauchy−Schwarz inequality. Note that the terms $\sigma_1^2, \sigma_2^2, \sigma_{12}$ are not necessary to be the variances and covariance of random variables x_1, x_2, since x_1, x_2 are not assumed to be of zero mean.

Equation (5.2) shows the dependence model that examines the relationship between these two genes,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} \\ a_{21} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}, \tag{5.2}$$

where, the elements of the dependence matrix (the values of a_{12}, a_{21}) are chosen such that the noise terms n_1, n_2 are minimized separately. For example,

$$\begin{aligned} E[n_1^2] &= E[(x_1 - a_{12}x_2)^2] \\ &= E[x_1^2] - 2a_{12}E[x_1x_2] + a_{12}^2E[x_2^2] \\ &= \sigma_1^2 - 2a_{12}\sigma_{12} + a_{12}^2\sigma_2^2 \end{aligned} \tag{5.3}$$

Equation (5.3) is a quadratic function of a_{12}, and the second order term has position coefficient. Therefore, $E[n_1^2]$, as a function of a_{12}, has a unique minimum. In order to choose a proper value of a_{12} to minimize $E[n_1^2]$, we take derivative of equation (5.3), and set it to be 0.

$$\frac{\partial E[n_1^2]}{\partial a_{12}} = \frac{\partial(\sigma_1^2 - 2a_{12}\sigma_{12} + a_{12}^2\sigma_2^2)}{\partial a_{12}}$$

$$= \quad 2a_{12}\sigma_2^2 - 2\sigma_{12}$$

$$= \quad 0$$

Thus, choosing $a_{12} = \sigma_{12}/\sigma_2^2$ minimizes $E[n_1^2]$. Similarly, following the same argument, in order to minimize $E[n_2^2]$, we choose $a_{21} = \sigma_{12}/\sigma_1^2$.

Therefore, for this 2-dimensional case, the dependence matrix is of the following form,

$$A_2 = \begin{bmatrix} 0 & a_{12} \\ a_{21} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{\sigma_{12}}{\sigma_2^2} \\ \frac{\sigma_{12}}{\sigma_1^2} & 0 \end{bmatrix}, \tag{5.4}$$

with eigenvalues being $\pm\sqrt{\frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}}$. It is clear that in the 2-dimensional case, the two eigenvalues of the dependence model are both real-valued in the range of $[-1, 1]$.

5.2.2 3-Dimensional Case

In this subsection, we discuss a much more challenging case, a 3-dimensional dependence model. We first derive the analytical form of the 3×3 dependence matrix. Then, we prove the eigenvalues are real-valued. And finally, we derive the analytical form of the eigenvalues.

In the 3-dimensional case, we study the dependence relationship among three genes. The expression level of the three gene are considered to be random variables, denoted as x_1, x_2, x_3. The second order statistics of the three random variables are

as follows, in equation (5.5).

$$\begin{cases} E[x_1^2] = \sigma_1^2 \\ E[x_2^2] = \sigma_2^2 \\ E[x_3^2] = \sigma_3^2 \\ E[x_1 x_2] = \sigma_{12} \\ E[x_1 x_3] = \sigma_{13} \\ E[x_2 x_3] = \sigma_{23} \end{cases} \tag{5.5}$$

where $\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2$, $\sigma_1^2\sigma_3^2 \geq \sigma_{13}^2$, and $\sigma_2^2\sigma_3^2 \geq \sigma_{23}^2$, because of the Cauchy−Schwarz inequality. Again, x_1, x_2, x_3 are not assumed to be zero mean. Therefore, the terms $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_{12}, \sigma_{13}, \sigma_{23}$ are not necessary to be variances and covariances.

The dependence model for this 3-dimensional case can be written as follows,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}, \tag{5.6}$$

We first determine the values of a_{12}, a_{13} in the dependence matrix through minimizing $E[n_1^2]$. It is easy to see that,

$$\begin{aligned} E[n_1^2] &= E[(x_1 - a_{12}x_2 - a_{13}x_3)^2] \\ &= E[x_1^2] + a_{12}^2 E[x_2^2] + a_{13}^2 E[x_3^2] - 2a_{12}E[x_1x_2] - 2a_{13}E[x_1x_3] + 2a_{12}a_{13}E[x_2x_3] \\ &= \sigma_1^2 + a_{12}^2\sigma_2^2 + a_{13}^2\sigma_3^2 - 2a_{12}\sigma_{12} - 2a_{13}\sigma_{13} + 2a_{12}a_{13}\sigma_{23} \end{aligned} \tag{5.7}$$

Equation (5.7) is a quadratic function of a_{12} and a_{13}, with position coefficients for the second order terms. Therefore, $E[n_1^2]$ has a unique minimum with respect to a_{12}

and a_{13}. In order to minimize $E[n_1^2]$, we take the partial derivatives of (5.7) with respect to a_{12} and a_{13}, and set the derivatives to be 0.

$$\begin{aligned}
\frac{\partial E[n_1^2]}{\partial a_{12}} &= \frac{\partial(\sigma_1^2 + a_{12}^2\sigma_2^2 + a_{13}^2\sigma_3^2 - 2a_{12}\sigma_{12} - 2a_{13}\sigma_{13} + 2a_{12}a_{13}\sigma_{23})}{\partial a_{12}} \\
&= 2a_{12}\sigma_2^2 - 2\sigma_{12} + 2a_{13}\sigma_{23} \\
&= 0 \qquad (5.8) \\
\frac{\partial E[n_1^2]}{\partial a_{13}} &= \frac{\partial(\sigma_1^2 + a_{12}^2\sigma_2^2 + a_{13}^2\sigma_3^2 - 2a_{12}\sigma_{12} - 2a_{13}\sigma_{13} + 2a_{12}a_{13}\sigma_{23})}{\partial a_{13}} \\
&= 2a_{13}\sigma_3^2 - 2\sigma_{13} + 2a_{12}\sigma_{23} \\
&= 0 \qquad (5.9)
\end{aligned}$$

From the above two equations, the values of a_{12} and a_{13} can be solved,

$$\begin{cases} a_{12} = \frac{\sigma_{12}\sigma_3^2 - \sigma_{13}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \\ a_{13} = \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \end{cases} \qquad (5.10)$$

Following the same argument, we can compute the values of a_{21}, a_{23}, a_{31} and a_{32} through minimizing $E[n_2^2]$ and $E[n_3^2]$. The analytical form of the dependence matrix is as follows,

$$\begin{aligned}
A_3 &= \begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \frac{\sigma_{12}\sigma_3^2 - \sigma_{13}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} & \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \\ \frac{\sigma_{12}\sigma_3^2 - \sigma_{13}\sigma_{23}}{\sigma_1^2\sigma_3^2 - \sigma_{13}^2} & 0 & \frac{\sigma_{23}\sigma_1^2 - \sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_3^2 - \sigma_{13}^2} \\ \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} & \frac{\sigma_{23}\sigma_1^2 - \sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} & 0 \end{bmatrix} \qquad (5.11)
\end{aligned}$$

where, it is easy to see that $a_{13}a_{21}a_{32} = a_{12}a_{23}a_{31}$.

Compared with the dependence matrix for the previous 2-dimensional case, the dependence matrix for the 3-dimensional case is much more complex. In order to prove the eigenvalues are real-valued and obtain the analytical form of the eigenvalues, we examine the characteristic polynomial of A_3.

$$\begin{aligned} det(A_3 - \lambda I) &= \begin{bmatrix} -\lambda & a_{12} & a_{13} \\ a_{21} & -\lambda & a_{23} \\ a_{31} & a_{32} & -\lambda \end{bmatrix} \\ &= -\lambda^3 + (a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})\lambda + 2a_{13}a_{21}a_{32} \end{aligned} \quad (5.12)$$

Let $f(\lambda) = det(A_3 - \lambda I)$, which is a 3rd order polynomial of the shape shown in Figure 5.1. The eigenvalues of A_3 are the roots of equation $f(\lambda) = 0$. If we can prove that the two extremals of $f(\lambda)$ (local maximal and local minimal) are of different signs, then we can argue that equation $f(\lambda) = 0$ has 3 real roots, and thus, all eigenvalues of A_3 are real.

The two extremal points can be obtained by setting $\frac{df(\lambda)}{d\lambda} = 0$. It is easy to see that the $\lambda = \pm\sqrt{\frac{1}{3}(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})}$ achieve the extremals of $f(\lambda)$. And the extremals of $f(\lambda)$ are:

$$\begin{cases} f_{Local\_max} = 2\sqrt{\frac{1}{3}(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})}^3 + 2a_{13}a_{21}a_{32} \\ f_{Local\_min} = -2\sqrt{\frac{1}{3}(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})}^3 + 2a_{13}a_{21}a_{32} \end{cases} \quad (5.13)$$

Taking into consideration the fact that $a_{13}a_{21}a_{32} = a_{12}a_{23}a_{31}$, the product of f_{Local_max} and f_{Local_min} can be expressed as follows,

$$\begin{aligned} f_{Local\_max}f_{Local\_min} &= -4\left(\frac{1}{3}(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})\right)^3 + 4\left(a_{13}a_{21}a_{32}\right)^2 \\ &= -4\left(\frac{1}{3}(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})\right)^3 + 4(a_{13}a_{21}a_{32})(a_{12}a_{23}a_{31}) \end{aligned}$$

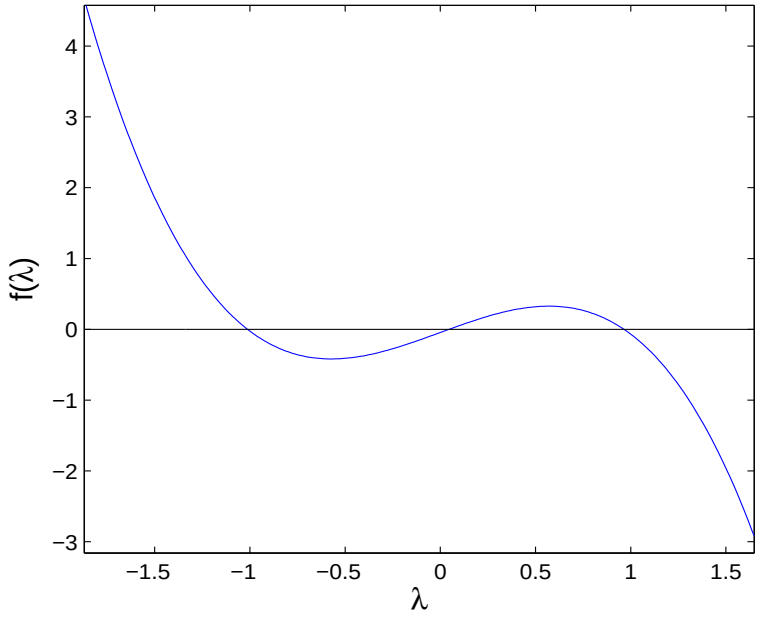


Figure 5.1: Shape of the characteristic polynomial of a dependence model with dimension being 3.

$$= \quad -4\left((\frac{1}{3}(a_{12}a_{21}+a_{13}a_{31}+a_{23}a_{32})\right)^3+4(a_{12}a_{21})(a_{13}a_{31})(a_{23}a_{32})$$

Denote $a = a_{12}a_{21}$, $b = a_{13}a_{31}$, and $c = a_{23}a_{32}$. Because of the structures shown in equation (5.11), a, b and c are all positive quantities. For positive quantities, the arithmetic mean is no less than the geometric mean,

$$\frac{1}{3}(a+b+c) \geq \sqrt[3]{abc} \tag{5.14}$$

and thus,

$$(\frac{1}{3}(a+b+c))^3 \geq abc \tag{5.15}$$

Therefore, it is easy to see that

$$\left((\frac{1}{3}(a_{12}a_{21}+a_{13}a_{31}+a_{23}a_{32})\right)^3 \geq (a_{12}a_{21})(a_{13}a_{31})(a_{23}a_{32}) \tag{5.16}$$

which implies $f_{Local\_max}f_{Local\_min} \leq 0$. When $f_{Local\_max}f_{Local\_min}$ is strictly less than

0, equation $f(\lambda) = 0$ has 3 different real roots. When $f_{Local\_max} f_{Local\_min}$ equals to 0, the roots are still real, but some of the roots may have more than 1 multiplicity. Therefore, all the three eigenvalues of the dependence matrix are real-valued.

Now, we have proved that in the 3-dimensional case, the eigenvalues of the dependence models are all real-valued. In the following, we will proceed to derive the analytical forms of the eigenvalues. Recall the characteristic equation of the dependence matrix A_3,

$$det(A_3 - \lambda I) = -\lambda^3 + (a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})\lambda + 2a_{13}a_{21}a_{32} = 0 \tag{5.17}$$

Denote $p = -(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})$, $q = 2a_{13}a_{21}a_{32}$, and make the Vieta's substitution $\lambda = w - \frac{p}{3w}$, equation (5.17) becomes,

$$(w^3)^2 - q(w^3) - (\frac{p}{3})^3 = 0 \tag{5.18}$$

which is a quadratic equation of w^3, with roots $w^3 = \frac{q}{2} \pm \sqrt{(\frac{q}{2})^2 + (\frac{p}{3})^3}$. From inequality (5.16), it follows that $(\frac{q}{2})^2 + (\frac{p}{3})^3 \leq 0$. Thus, the magnitudes of the two roots are equal, $|w^3| = \sqrt{(-\frac{p}{3})^3}$, and w^3 can be written in the following form,

$$w^3 = \begin{cases} \sqrt{(-\frac{p}{3})^3} e^{i\theta_1} \\ \sqrt{(-\frac{p}{3})^3} e^{i\theta_2} \end{cases} \tag{5.19}$$

where,

$$\begin{cases} cos(\theta_1) = cos(\theta_2) = \frac{q}{2} \\ sin(\theta_1) = \sqrt{-(\frac{q}{2})^2 - (\frac{p}{3})^3} \\ sin(\theta_2) = -\sqrt{-(\frac{q}{2})^2 - (\frac{p}{3})^3} \end{cases} \tag{5.20}$$

Therefore,

$$w = \begin{cases} \sqrt{-\frac{p}{3}} e^{i\frac{\theta_1}{3}} \\ \sqrt{-\frac{p}{3}} e^{i(\frac{\theta_1}{3}+\frac{2\pi}{3})} \\ \sqrt{-\frac{p}{3}} e^{i(\frac{\theta_1}{3}+\frac{4\pi}{3})} \\ \sqrt{-\frac{p}{3}} e^{i\frac{\theta_2}{3}} \\ \sqrt{-\frac{p}{3}} e^{i(\frac{\theta_2}{3}+\frac{2\pi}{3})} \\ \sqrt{-\frac{p}{3}} e^{i(\frac{\theta_2}{3}+\frac{4\pi}{3})} \end{cases} \tag{5.21}$$

Recall the Vieta's substitution $\lambda = w - \frac{p}{3w}$, we can calculate the eigenvalues,

$$\lambda = \begin{cases} 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3}) \\ 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3} + \frac{2\pi}{3}) \\ 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3} + \frac{4\pi}{3}) \\ 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_2}{3}) \\ 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_2}{3} + \frac{2\pi}{3}) \\ 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_2}{3} + \frac{4\pi}{3}) \end{cases} \tag{5.22}$$

Because θ_1 and θ_2 satisfy equation (5.20), the first 3 solutions of λ are the same as the last 3 solutions. Since $\theta_1 \in [0, \pi)$, it is easy to order the first 3 solutions of λ,

$$\begin{cases} \lambda_{max} &= 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3}) \\ \lambda_{mid} &= 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3} + \frac{4\pi}{3}) \\ \lambda_{min} &= 2\sqrt{-\frac{p}{3}} cos(\frac{\theta_1}{3} + \frac{2\pi}{3}) \end{cases} \tag{5.23}$$

where $p = -(a_{12}a_{21} + a_{13}a_{31} + a_{23}a_{32})$, and θ_1 is defined in equation (5.20).

As a summary, in this subsection, we examine the dependence model for the 3-dimensional case in detail. We prove that the eigenvalues for the 3-dimensional case are real, and we derive the analytical forms of the eigenvalues, as shown in

equation (5.23). It is difficult to tell the numerical range of the eigenvalues from equation (5.23). In the next subsection, we will discuss higher dimensional cases, where general results for higher dimensional cases will be presented. For now, we just present the result that the eigenvalues for the 3-dimensional cases belong to the range of $[-2, 1]$. The proof will be presented in the next subsection.

5.2.3 High Dimensional Case

In the previous subsections, we prove the the eigenvalues are real-valued in the 2-dimensional case and the 3-dimensional case. The proof is complicated, and the 3-dimensional case can not be generalized from the 2-dimensional case. In this subsection, we will take a slightly different approach that can be generalized to high dimensional cases. We will prove for a high dimensional case (M-dimensional case) that, all the eigenvalues of the dependence model are real-valued, belonging to the range $[-(M-1), 1]$.

Before we discuss high dimensional cases, let's first re-visit the 3-dimensional case. The first row of the dependence matrix is determined through minimizing $E[n_1^2]$, that is, by solving equations (5.8) and (5.9), which can be written in a matrix form as follows,

$$\begin{bmatrix} \sigma_2^2 & \sigma_{23} \\ \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a_{12} \\ a_{13} \end{bmatrix} = \begin{bmatrix} \sigma_{12} \\ \sigma_{13} \end{bmatrix} \tag{5.24}$$

Similarly, through minimizing $E[n_2^2]$ and $E[n_3^2]$, the second and third row of the

dependence matrix can be determined from the following two matrix equations,

$$\begin{bmatrix} \sigma_1^2 & \sigma_{13} \\ \sigma_{13} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{23} \end{bmatrix} = \begin{bmatrix} \sigma_{12} \\ \sigma_{23} \end{bmatrix} \tag{5.25}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a_{31} \\ a_{32} \end{bmatrix} = \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \tag{5.26}$$

Matrix equations (5.24), (5.25) and (5.26) share a common structure, which might be more obvious in the following extended versions,

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 0 \\ a_{12} \\ a_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ \sigma_{12} \\ \sigma_{13} \end{bmatrix} \tag{5.27}$$

$$\begin{bmatrix} \sigma_1^2 & 0 & \sigma_{13} \\ 0 & \sigma_2^2 & 0 \\ \sigma_{13} & 0 & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a_{21} \\ 0 \\ a_{23} \end{bmatrix} = \begin{bmatrix} \sigma_{12} \\ 0 \\ \sigma_{23} \end{bmatrix} \tag{5.28}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a_{31} \\ a_{32} \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \\ 0 \end{bmatrix} \tag{5.29}$$

The similar structures of matrix equations (5.27), (5.28) and (5.29) motivate us to examine the following,

$$\begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} a_{12}\sigma_{12} + a_{13}\sigma_{13} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & a_{21}\sigma_{21} + a_{23}\sigma_{23} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & a_{31}\sigma_{13} + a_{32}\sigma_{23} \end{bmatrix} \tag{5.30}$$

which can be denoted as,

$$A_3 R_3 = B_3 \tag{5.31}$$

where, R_3 is a symmetric positive semi−definite matrix, and B_3 is a symmetric matrix.

The characteristic polynomial of dependence matrix A_3 can be related with R_3 and B_3 as follows,

$$\begin{aligned} & det(A_3 - \lambda I) \\ = \; & det((A_3 - \lambda I)R_3)/det(R_3) \\ = \; & det(A_3 R_3 - \lambda R_3) det(R_3^{-1}) \\ = \; & det(B_3 - \lambda R_3) det(R_3^{-\frac{1}{2}}) det(R_3^{-\frac{1}{2}}) \\ = \; & det(R_3^{-\frac{1}{2}}(B_3 - \lambda R_3)R_3^{-\frac{1}{2}}) \\ = \; & det(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - \lambda R_3^{-\frac{1}{2}} R_3 R_3^{-\frac{1}{2}}) \\ = \; & det(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - \lambda I) \end{aligned} \tag{5.32}$$

where, $R_3^{-\frac{1}{2}} = VD^{-\frac{1}{2}}V'$. V and D are obtained from the eigen-decomposition of $R_3 = VDV'$. From equation (5.32), we can see that matrix A_3 and matrix $R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}}$ share the same set of eigenvalues. Since R_3 is symmetric positive semi−definite and B_3 is symmetric, $R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}}$ is a symmetric matrix, whose eigenvalues are all real-valued. Therefore, the eigenvalues of A_3 are all real-valued.

In order to determine the range of the eigenvalues of A_3, we examine the following quantity, where z is an arbitrary vector,

$$\begin{aligned}
& z'(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - I)z \\
= \; & z'(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - R_3^{-\frac{1}{2}} R_3 R_3^{-\frac{1}{2}})z \\
= \; & z' R_3^{-\frac{1}{2}} (B_3 - R_3) R_3^{-\frac{1}{2}} z \\
= \; & (R_3^{-\frac{1}{2}} z)'(B_3 - R_3)(R_3^{-\frac{1}{2}} z) \\
= \; & (R_3^{-\frac{1}{2}} z)' \begin{bmatrix} a_{12}\sigma_{12} + a_{13}\sigma_{13} - \sigma_1^2 & 0 & 0 \\ 0 & a_{21}\sigma_{21} + a_{23}\sigma_{23} - \sigma_2^2 & 0 \\ 0 & 0 & a_{31}\sigma_{13} + a_{32}\sigma_{23} - \sigma_3^2 \end{bmatrix} (R_3^{-\frac{1}{2}} z)
\end{aligned} \tag{5.33}$$

Since R_3 is positive semi−definite, $y' R_3 y \geq 0$ holds for any vector y. Let $y = [-1, a_{12}, a_{13}]'$,

$$\begin{aligned}
[-1, a_{12}, a_{13}] R_3 \begin{bmatrix} -1 \\ a_{12} \\ a_{13} \end{bmatrix} &\geq 0 \\
-(a_{12}\sigma_{12} + a_{13}\sigma_{13} - \sigma_1^2) &\geq 0 \\
a_{12}\sigma_{12} + a_{13}\sigma_{13} - \sigma_1^2 &\leq 0
\end{aligned}$$

With similar method, we can show that all diagonal elements in equation (5.33) are smaller than or equal to 0. So, for any vector z, $z'(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - I)z$ is less than or equal to 0, meaning that $(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - I)$ is negative semi−definite, with all eigenvalues less than or equal to 0. Thus, all the eigenvalues of $R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}}$ are less than or equal to 1. Therefore, all the eigenvalues of A_3 are less than or equal to 1.

On the other hand, the sum of A_3's eigenvalues equals to $tr(A_3)$, which is 0. From these two facts, we can conclude that the eigenvalues of A_3 belong to the range $[-2, 1]$. Examples where eigenvalues take the boundary values (-2 and 1) can be found in Appendix C.

In the high M-dimensional case, i.e. $M > 3$, there is an equality similar with equation (5.30).

$$\begin{bmatrix} 0 & a_{12} & \cdots & a_{1M} \\ a_{21} & 0 & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & 0 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sigma_M^2 \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{i\neq 1} a_{1i}\sigma_{1i} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sum_{i\neq 2} a_{2i}\sigma_{2i} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sum_{i\neq M} a_{Mi}\sigma_{iM} \end{bmatrix} \tag{5.34}$$

which can be denoted as,

$$A_M R_M = B_M \tag{5.35}$$

where, R_M is a symmetric positive semi$-$definite matrix, and B_M is a symmetric matrix. Through the same argument in equation (5.32), we can show that,

$$det(A_M - \lambda I) = det(R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}} - \lambda I) \tag{5.36}$$

Thus, A_M and $R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}}$ share the same set of eigenvalues. Since $R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}}$ is symmetric with real eigenvalues, the eigenvalues of A_M are all real-valued. Similar

with equation (5.33), for an arbitrary vector z,

$$
\begin{aligned}
& z'(R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}} - I)z \\
= \; & (R_M^{-\frac{1}{2}} z)' \begin{bmatrix} -\sigma_1^2 + \sum_{i \neq 1} a_{1i}\sigma_{1i} & 0 & \cdots & 0 \\ 0 & -\sigma_2^2 + \sum_{i \neq 2} a_{2i}\sigma_{2i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\sigma_M^2 + \sum_{i \neq M} a_{Mi}\sigma_{iM} \end{bmatrix} (R_M^{-\frac{1}{2}} z)
\end{aligned}
\tag{5.37}
$$

where each diagonal element of the above is less than or equal to 0. For any vector z, $z'(R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}} - I)z$ is less than or equal to 0, meaning that $(R_3^{-\frac{1}{2}} B_3 R_3^{-\frac{1}{2}} - I)$ is negative semi−definite. Thus, all the eigenvalues of $R_M^{-\frac{1}{2}} B_M R_M^{-\frac{1}{2}}$ are less than or equal to 1, and all the eigenvalues of A_M are less than or equal to 1. Together with the fact that, the sum of A_M's eigenvalues equals to $tr(A_M)$, which is 0, we can conclude that the eigenvalues of A_M belong to the range $[-(M-1), 1]$. Examples where eigenvalues take the boundary values can be found in Appendix C.

As a summary, in this subsection, we present a solid mathematically foundation of the dependence model, proving that for an M-dimensional dependence model, the eigenvalues are all real-valued and belong to the range of $[-(M-1), 1]$.

5.3 Regulatory Relationships vs. Eigenvalue Pattern

In this section, we discuss how regulatory relationships can affect the eigenvalue pattern in several cases. We propose to use the eigenvalue pattern as a feature to study regulatory relationships. In the literature, all existing methods examine the regulatory relationship between one or more regulators and one regulated gene. However, using the proposed dependence model and the eigenvalue pattern, we are able to examine the regulatory relationship between one or more regulators and a set of regulated genes. In the following, we discuss several cases, and show why the eigenvalue pattern can be used to identify regulatory relationships.

Case 1:

Suppose we are interested in the regulatory relationships among 4 genes, gene 1, 2, 3, and 4, where the ground truth is gene 4 regulates gene 1. Assume that without the presence of gene 4, the expression of gene 1, 2, 3 during some certain biological process (i.e. cell-cycle) are $x_1(t)$, $x_2(t)$, $x_3(t)$ respectively, which are stationary random processes. Their second order statistics are assumed to be stationary (5.5), and their dependence relationships can be described by the dependence matrix (5.11). Assume that gene 4's expression, $x_4(t)$, is orthogonal to gene 1, 2, 3. With the presence of gene 4, the expression of gene 1 becomes $x_1'(t) = x_1(t) + s_1(t)$, where $s_1(t) = \alpha x_4(t-1)$; while the presence of gene 4 does not affect gene 2 and 3, i.e. $x_2'(t) = x_2(t)$, $x_3'(t) = x_3(t)$. Then, with the presence of gene 4, the dependence

model among gene 1, 2, and 3 becomes,

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t) \end{bmatrix} = \begin{bmatrix} 0 & a_{12}' & a_{13}' \\ a_{21}' & 0 & a_{23}' \\ a_{31}' & a_{32}' & 0 \end{bmatrix} \begin{bmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t) \end{bmatrix} + \begin{bmatrix} n_1'(t) \\ n_2'(t) \\ n_3'(t) \end{bmatrix}$$

that is,

$$\begin{bmatrix} x_1(t)+s_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} 0 & a_{12}' & a_{13}' \\ a_{21}' & 0 & a_{23}' \\ a_{31}' & a_{32}' & 0 \end{bmatrix} \begin{bmatrix} x_1(t)+s_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} n_1'(t) \\ n_2'(t) \\ n_3'(t) \end{bmatrix}$$

where the dependence matrix is,

$$A_3'(t) = \begin{bmatrix} 0 & \frac{\sigma_{12}\sigma_3^2-\sigma_{13}\sigma_{23}}{\sigma_2^2\sigma_3^2-\sigma_{23}^2} & \frac{\sigma_{13}\sigma_2^2-\sigma_{12}\sigma_{23}}{\sigma_2^2\sigma_3^2-\sigma_{23}^2} \\ \frac{\sigma_{12}\sigma_3^2-\sigma_{13}\sigma_{23}}{(\sigma_1^2+\sigma_{s_1}^2(t))\sigma_3^2-\sigma_{13}^2} & 0 & \frac{\sigma_{23}(\sigma_1^2+\sigma_{s_1}^2(t))-\sigma_{12}\sigma_{13}}{(\sigma_1^2+\sigma_{s_1}^2(t))\sigma_3^2-\sigma_{13}^2} \\ \frac{\sigma_{13}\sigma_2^2-\sigma_{12}\sigma_{23}}{(\sigma_1^2+\sigma_{s_1}^2(t))\sigma_2^2-\sigma_{12}^2} & \frac{\sigma_{23}(\sigma_1^2+\sigma_{s_1}^2(t))-\sigma_{12}\sigma_{13}}{(\sigma_1^2+\sigma_{s_1}^2(t))\sigma_2^2-\sigma_{12}^2} & 0 \end{bmatrix} \tag{5.38}$$

and $\sigma_{s_1}^2(t) = \alpha^2\sigma_4^2(t-1) = \alpha^2 E[x_4^2(t-1)]$. From (5.38), we can see that with the presence of gene 4, the regulatory relationship between gene 4 and gene 1 directly affects the dependence matrix for gene 1, 2, 3. The eigenvalue pattern, $\lambda(A_3'(t))$, is a nonlinear function of $\sigma_{s_1}^2(t)$, or a nonlinear function of the statistics of $x_4(t-1)$.

In the case where gene 4 regulates two or three genes among genes 1, 2, 3, the dependence matrix will have a more complicated form, and the eigenvalues are still a nonlinear function of the statistics of the regulator $x_4(t-1)$. Although $\lambda(A_3'(t))$ does not change linearly with the change of $x_4(t-1)$, the nonlinear relationship between $\lambda(A_3'(t))$ and $x_4(t-1)$ can be approximately quantified by linear measures such as correlation coefficient or nonlinear measures such as mutual information.

Therefore, in this case, the regulator and the eigenvalue pattern of the regulated genes are nonlinearly correlated.

Case 2:

Suppose we are examining the same set of genes as in case 1, where gene 4 is orthogonal to gene 1, 2, 3. Again, without the presence of gene 4, the expression of gene 1, 2, 3 are assumed to be stationary processes, whose statistics are stationary (5.5). Different from previous case, in this case, gene 4 does not directly regulate gene 1, 2, 3. Instead, gene 4 works as a switch, which regulates the strength of how gene 1 is correlated with gene 2. In particular, with the presence of gene 4, the expression of gene 1 becomes, $x_1''(t) = x_1(t) - \frac{E[x_1(t)x_2(t)]}{E[x_2^2(t)]}x_2(t)s''(t) = x_1(t) - \frac{\sigma_{12}}{\sigma_2^2}x_2(t)s''(t)$, where $s''(t) = \beta x_4(t-1)$; while the expression of gene 2 and 3 are not affected, i.e. $x_2''(t) = x_2(t)$, $x_3''(t) = x_3(t)$. Then, with the presence of gene 4, the dependence model for gene 1, 2, and 3 becomes,

$$\begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} = \begin{bmatrix} 0 & a_{12}'' & a_{13}'' \\ a_{21}'' & 0 & a_{23}'' \\ a_{31}'' & a_{32}'' & 0 \end{bmatrix} \begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} + \begin{bmatrix} n_1''(t) \\ n_2''(t) \\ n_3''(t) \end{bmatrix}$$

that is,

$$\begin{bmatrix} x_1(t) - \frac{\sigma_{12}}{\sigma_2^2}x_2(t)s''(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} 0 & a_{12}'' & a_{13}'' \\ a_{21}'' & 0 & a_{23}'' \\ a_{31}'' & a_{32}'' & 0 \end{bmatrix} \begin{bmatrix} x_1(t) - \frac{\sigma_{12}}{\sigma_2^2}x_2(t)s''(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} n_1''(t) \\ n_2''(t) \\ n_3''(t) \end{bmatrix}$$

where the dependence matrix A_3'' has a structure similar with (5.11),

$$A_3'' = \begin{bmatrix} 0 & \frac{\sigma_{12}''\sigma_3''^2 - \sigma_{13}''\sigma_{23}''}{\sigma_2''^2\sigma_3''^2 - \sigma_{23}''^2} & \frac{\sigma_{13}''\sigma_2''^2 - \sigma_{12}''\sigma_{23}''}{\sigma_2''^2\sigma_3''^2 - \sigma_{23}''^2} \\ \frac{\sigma_{12}''\sigma_3''^2 - \sigma_{13}''\sigma_{23}''}{\sigma_1''^2\sigma_3''^2 - \sigma_{13}''^2} & 0 & \frac{\sigma_{23}''\sigma_1''^2 - \sigma_{12}''\sigma_{13}''}{\sigma_1''^2\sigma_3''^2 - \sigma_{13}''^2} \\ \frac{\sigma_{13}''\sigma_2''^2 - \sigma_{12}''\sigma_{23}''}{\sigma_1''^2\sigma_2''^2 - \sigma_{12}''^2} & \frac{\sigma_{23}''\sigma_1''^2 - \sigma_{12}''\sigma_{13}''}{\sigma_1''^2\sigma_2''^2 - \sigma_{12}''^2} & 0 \end{bmatrix} \tag{5.39}$$

and,

$$\begin{aligned} \sigma_1''^2 &= \sigma_1^2 + \frac{\sigma_{12}^2}{\sigma_2^2}\sigma_{s''}^2 \\ \sigma_2''^2 &= \sigma_2^2 \\ \sigma_3''^2 &= \sigma_3^2 \\ \sigma_{12}'' &= (1 - m_{s''})\sigma_{12} \\ \sigma_{13}'' &= \sigma_{13} - \frac{\sigma_{12}}{\sigma_2^2}\sigma_{23}m_{s''} \\ \sigma_{23}'' &= \sigma_{23} \end{aligned}$$

where and $\sigma_{s''}^2(t) = \beta^2\sigma_4^2(t-1) = \beta^2 E[x_4^2(t-1)]$, and $m_{s''} = \beta m_{x_4} = \beta E[x_4(t)]$. Again, the regulatory relationship between the regulator (gene 4) and the regulated genes (gene 1, 2) will affect the second order statistics of genes 1, 2, 3, and thus affect the dependence matrix and the eigenvalue pattern. Therefore, there exists a nonlinear relationship between $\lambda(A_3''(t))$ and $x_4(t-1)$.

As a summary of this section, for a pair of regulator and regulated genes, there exists time-lagged correlation between the expression of the regulator and the eigenvalues associated to the regulated genes. In the case where the time-lagged correlation between the expression of the a regulator and the expression of a regulated gene is week, the using the eigenvalues can serve as an alternative way to discover the regulatory relationship.

5.4 Discovery of Regulatory Relationships

In the previous section, we show that the eigenvalues of a small group of genes is a nonlinear function of the external factors that regulates one or more genes in the group. Thus, the expression of the regulator and the eigenvalues associated to the regulated genes are nonlinearly correlated. In this section, we examine a yeast cell-cycle microarray time-series dataset, and show how to discover regulatory relationships from the eigenvalue pattern.

5.4.1 Yeast Regulatory Network (Dataset and Prior Knowledge)

In this study, we examine the *alpha* time-series dataset in [16], which is also studied in Chapter 4. The dataset contains 18 time points for 6178 genes. The 18 time points cover two cell-cycles, about 120 minutes, with sampling time interval being 7 minutes. The time-serious for each gene is normalized to zero-mean. Our goal is to identify regulatory relationships from the time-series dataset.

In order to set up a performance evaluation criterion, we employ the partial model of the yeast transcriptional regulatory network [32, 91]. The partial model is derived by [91], based on the findings of [15]. As shown in Figure 5.2, the partial model describes 58 regulatory relationships among 30 genes. In this partial model, the topology is fixed, but there is no associated rules, i.e., the regulatory relationships may not be boolean, Bayesian, etc. The topology of the partial model is considered as the ground truth in our study. Identification method that discovers

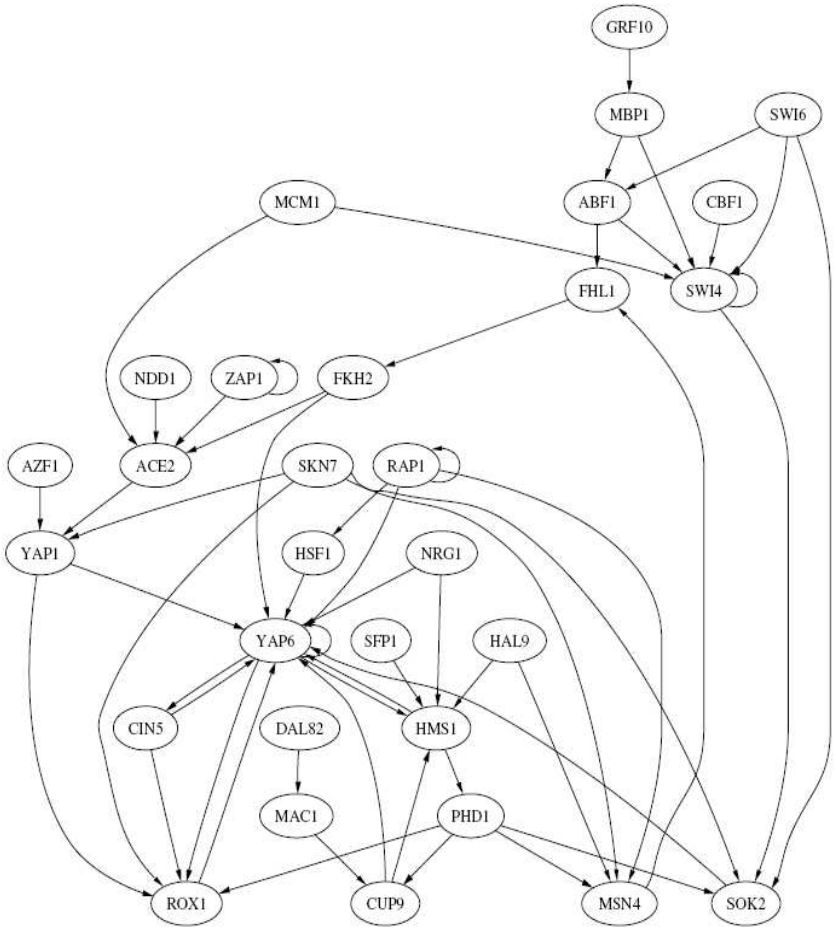


Figure 5.2: Partial model of the gene regulatory network of yeast cell-cycle.

large portion of the interactions in the partial model is considered to have good performance.

Similar with the construction of the dependence networks in Chapter 3, the eigenvalue-based discovery of regulatory relationships involves exhaustive search. To limit the computational complexity, it is desirable to limit the number of genes under investigation and focus on a small gene system. (Details will be presented in the following subsections). Since a partial model of 30 genes is employed as per-

formance evaluation criterion, in this study, we will focus on discovering regulatory relationships among the 30 genes.

5.4.2 Correlation Between Gene Expressions

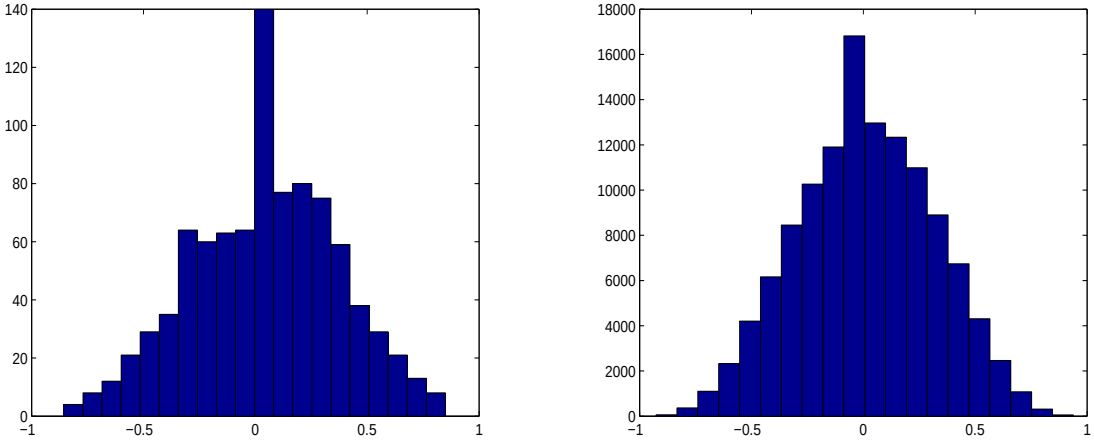
To infer and identify a regulatory relationship between a pair of genes, the simplest way is to test the correlation between the time-lagged expression time-series of the two genes, either the original continuous expression data [92] or the quantized discrete expression data [93]. If the time-lagged correlation is larger than certain threshold, there exists a regulatory relationship. The threshold can be determined by permutation analysis or graph-theoretic transitivity measure such as clustering coefficient [94]. The statistical significance of the time-lagged correlation can be assessed by p-value, which is obtained through random permutation. In this section, we pair-wisely examine the correlation coefficient and p-value of the 30 genes in the partial model in Figure 5.2.

Define $\mathbf{x}_{(i)}{}_a^b$ as a column vector containing gene i's time-series expression data from time point a to time point b, i.e., $\mathbf{x}_{(i)}{}_a^b = [x_{i,a}, x_{i,a+1}, ..., x_{i,b}]^T$, where $x_{i,a}$ represents gene i's expression at time a. i takes value from 1 to 30, because we are examining 30 genes. Since the dataset under investigation [16] contains 18 time points, a and b take values from 1 to 18, and $a \leq b$. In order to infer whether gene i regulates gene j, we calculate the time-lagged correlation between genes i and j,

$$c_{i,j} = \frac{< \mathbf{x}_{(i)}{}_1^{17}, \mathbf{x}_{(j)}{}_2^{18} >}{|\mathbf{x}_{(i)}{}_1^{17}| \cdot |\mathbf{x}_{(j)}{}_2^{18}|} \tag{5.40}$$

where $< \cdot, \cdot >$ is the vector inner product, and $|\cdot|$ represents the L_2 vector norm.

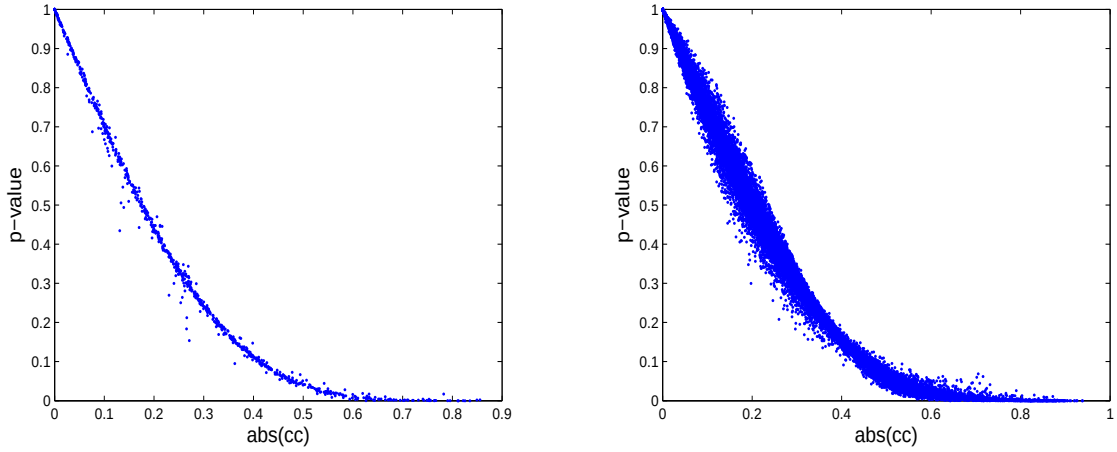
$c_{i,j}$ belongs to the range of $[-1, 1]$. A positive $c_{i,j}$ means gene i up-regulates gene j; while a negative $c_{i,j}$ means gene i down-regulates gene j. The absolute value of $c_{i,j}$ represents the strength of the regulatory relationship. The histogram of all $c_{i,j}$'s are shown in Figure 5.3(a).



(a) (b)

Figure 5.3: Fig (a) shows the histogram of time-lagged correlation between gene pairs $c_{i,j}$ for all $i, j = 1, 2, ..., 30$. Fig (b) shows the histogram of time-lagged correlation between gene expression and eigenvalue of gene triple, $c_{i;(j,k,l)}$ for all $i, j, k, l = 1, 2, ..., 30$ and $j \neq k$, $j \neq l$, $k \neq l$.

The statistical significance of $c_{i,j}$ is assessed by permutation analysis. We randomly permute the elements of vector $\mathbf{x}_{(i)}{}_1^{17}$ 10000 times, and compute the correlation between $\mathbf{x}_{(j)}{}_2^{18}$ and the permuted versions of $\mathbf{x}_{(i)}{}_1^{17}$. The p-value $p_{i,j}$ is defined as the probability that the absolute value of correlation of random data is greater than that of the original un-permuted data. The smaller the p-value is, the more confident we are about the correlation $c_{i,j}$. Intuitively, larger absolute value of $c_{i,j}$ will lead to smaller $p_{i,j}$. In Figure 5.4(a), we plot the $c_{i,j}$ and $p_{i,j}$ of all gene pairs, where we can see that larger absolute value of $c_{i,j}$ in general corresponds to smaller $p_{i,j}$.



(a) (b)

Figure 5.4: Fig (a) shows the p-value vs absolute value of time-lagged correlation of all gene pairs i, j. Fig (b) shows the p-value vs absolute value of time-lagged correlation of all pairs of regulator and regulated triple $i; (j, k, l)$.

For the purpose of identifying and discovering regulatory relationships, we examine the time-lagged correlation and the p-value. For a pair of genes i and j, if the absolute value of $c_{i,j}$ is large and $p_{i,j}$ is small, we will infer that gene i regulates gene j. Since the absolute value of the correlation and the p-value changes monotonically with each other in Figure 5.4(a), we only need to apply a correlation threshold c^* and ignore the p-value. For a pair of genes i and j, if $|c_{i,j}| \geq c^*$ is satisfied, we will identify the regulatory relationship that gene i regulates gene j.

Given the ground truth of the regulatory relationships among the 30 genes under investigation, we can evaluate the detection performance of particular choices of the correlation threshold c^*, in terms of the probability of mis-detection and the probability of false-positive. By varying the correlation threshold c^*, we are able to examine the ROC (receiver operating characteristic) of identifying regulatory relationships by thresholding $c_{i,j}$. As mentioned in the previous subsection, the

partial model of yeast describes 58 regulatory relationships among 30 genes. If we consider the 58 regulatory relationships as ground truth, we can obtain one realization of the ROC curve for thresholding $c_{i,j}$ as the dashed line in Figure 5.5(a), where the horizontal axis is the probability of false-positive, and the vertical asix is the probability of correct detection which is 1 minus the probability of mis-detection. Note that an ROC curve is a statistical characterization of a detection scheme. Given a particular dataset, we are only able to obtain one realization of the ROC curve.

Note that multi-hop regulatory relationships could also be picked up by the time-lagged correlation $c_{i,j}$, and the 58 regulatory relationships may not be exactly the ground truth to be discovered by thresholding $c_{i,j}$. Thus, the realization of ROC curve in Figure 5.5(a) may not be able to characterize the detection performance. Therefore, we may need to consider multi-hop regulatory relationships as the ground truth. From the partial model in Figure 5.2, we can easy derive all the multi-hop regulatory relationships. If we consider both one-hop and two-hop regulatory relationships as the ground truth, which contains 100 regulatory relationships, we will obtain the realization of the ROC curve as the dashed curve in Figure 5.5(b). If we consider all the one-hop, two-hop and three-hop regulatory relationships as the ground truth, which contains 156 regulatory relationships, we will obtain the dashed curve in Figure 5.5(c). If we consider the 214 relationships less or equal to four-hop, we can observe the dashed curve in Figure 5.5(d). It is observed that the 4 cases do not exhibit much difference.

5.4.3 Correlation Between Expression and Eigenvalues

In this subsection, we study the time-lagged correlation between regulator's expression and the eigenvalues of the regulated genes. As discussed in section 5.3, if there exists gene i that regulates one or more components of the gene triple j, k, l, there is likely to exist strong time-lagged correlation between gene i's expression and the eigenvalues of the triple. We will use this correlation to infer the existence of regulatory relationships. Based on the observation from Chapter 2 and Chapter 3, the smallest eigenvalue of the dependence model is most sensitive to the change of dependence relationship. Therefore, we choose the smallest eigenvalue to be the representative of all eigenvalues, and focus on the correlation between the regulator's expression and the smallest eigenvalue of the regulated genes.

Define $\lambda_{(j,k,l);t}$ be the smallest eigenvalue of the dependence model for genes j, k, l at time t. In our analysis, we choose a time window of 5 points to estimate the dependence model and the smallest eigenvalue. For example, in order to estimate $\lambda_{(j,k,l);t}$, we use the expression data ${\mathbf{x}_{(j)}}_{t-2}^{t+2}$, ${\mathbf{x}_{(k)}}_{t-2}^{t+2}$, ${\mathbf{x}_{(l)}}_{t-2}^{t+2}$ to estimate the dependence model of genes j, k, l for the time window $[t-2, t+2]$, and compute the smallest eigenvalue. For a particular choice of gene triple j, k, l, we are able to estimate $\lambda_{(j,k,l);t}$ for different time points, and see how the eigenvalue pattern changes along time. Similar with the previous subsection, define ${\lambda_{(j,k,l)}}_a^b = [\lambda_{(j,k,l);a}, \lambda_{(j,k,l);a+1}, ..., \lambda_{(j,k,l);b}]^T$. In order to determine whether gene i regulates one or more components of gene triple j, k, l, we examine the following correlation,

$$c_{i;(j,k,l)} = \frac{< {\mathbf{x}_{(i)}}_2^{15}, {\lambda_{(j,k,l)}}_3^{16} - {\overline{\lambda}_{(j,k,l)}}_3^{16} >}{|{\mathbf{x}_{(i)}}_2^{15}| \cdot |{\lambda_{(j,k,l)}}_3^{16} - {\overline{\lambda}_{(j,k,l)}}_3^{16}|} \tag{5.41}$$

where $\overline{\lambda}_{(j,k,l)_3^{16}}$ is the mean of the elements in vector $\lambda_{(j,k,l)_3^{16}}$. In Figure 5.3(b), the histogram of $c_{i;(j,k,l)}$ is shown, for all $i, j, k, l = 1, 2, ..., 30$ where $j \neq k$, $j \neq l$, $k \neq l$.

Similar with the previous subsection, the statistical significance of $c_{i;(j,k,l)}$ can be assessed by permutation analysis. We randomly permute the elements of $\mathbf{x}_{(i)_2^{15}}$ 10000 times, and compute the correlation between $\lambda_{(j,k,l)_3^{16}}$ and the permuted versions of $\mathbf{x}_{(i)_2^{15}}$, as in equation (5.41). The p-value, $p_{i;(j,k,l)}$, is defined as the probability that the absolute value of correlation of random data is greater than that of the original un-permuted data. Again, as shown in Figure 5.4(b), we observed that higher absolute value of the correlation will lead to smaller p-value. The reason for showing Figure 5.3(b) and Figure 5.4(b) is to demonstrate that among all possible pairs of regulator and regulated triple, the correlation derived from eigenvalues is centered around zero without bias.

As mentioned earlier, if gene i regulated gene j, the regulatory relationship will contribute to the correlation between gene i and the eigenvalue pattern of a regulated triple, genes j, k, l. Therefore we argue that, if we remove the effect of gene i from the expression of gene j, the correlation from the eigenvalue pattern will be reduced. Mathematically, define $\mathbf{x}_{(j_{(-i)})_a^b}$ as a column vector containing the expression of gene j from time point a to time point b, after the effect of gene i is removed from gene j:

$$\mathbf{x}_{(j_{(-i)})_a^b} = \mathbf{x}_{(j)_a^b} - \frac{< \mathbf{x}_{(j)_a^b}, \mathbf{x}_{(i)_{a-1}^{b-1}} >}{|\mathbf{x}_{(i)_{a-1}^{b-1}}|^2} \mathbf{x}_{(i)_{a-1}^{b-1}} \tag{5.42}$$

Based on the data $\mathbf{x}_{(j_{(-i)})_1^{18}}$, $\mathbf{x}_{(k)_1^{18}}$ and $\mathbf{x}_{(l)_1^{18}}$, we are able to obtain the eigenvalue

pattern $\lambda_{(j_{(-i)},k,l)}{}_3^{16}$. Define

$$c_{i;(j_{(-i)},k,l)} = \frac{< \mathbf{x}_{(i)}{}_2^{15}, \lambda_{(j_{(-i)},k,l)}{}_3^{16} - \overline{\lambda}_{(j_{(-i)},k,l)}{}_3^{16} >}{|\mathbf{x}_{(i)}{}_2^{15}| \cdot |\lambda_{(j_{(-i)},k,l)}{}_3^{16} - \overline{\lambda}_{(j_{(-i)},k,l)}{}_3^{16}|} \tag{5.43}$$

Then, if $c_{i;(j,k,l)}$ is greater than $c_{i;(j_{(-i)},k,l)}$, we argue that gene i is likely to regulate gene j.

In order to characterize the regulatory relationship of $i \rightarrow j$, we examine the reduction of eigenvalue pattern correlation $c_{i;(j,k,l)} - c_{i;(j_{(-i)},k,l)}$. Since we exhaustively examine all possible regulated triples, for a particular regulatory relationship ($i \rightarrow j$), the correlation reduction can be evaluated 406 times, because of different choices of k and l ($406 = \frac{29 \times 28}{2}$). We propose to use the mean of the 406 values of correlation reduction as a metric to characterize the possible regulatory relationship of $i \rightarrow j$.

We evaluate the proposed correlation reduction metric for all pairs of genes, and obtain a 30×30 matrix denoted as Cr, with the Cr_{ij} element characterizing the strength of the regulatory relationship of $i \rightarrow j$. Similar with the previous subsection, we apply a correlation threshold cr^*. If $Cr_{ij} \geq cr^*$ is satisfied, we will identify the regulatory relationship that gene i regulates gene j.

Given the ground truth of the regulatory relationships, we can evaluate the detection performance of particular choices of the threshold cr^*. By varying the value of cr^*, we are able to examine the ROC curve of identifying regulatory relationships from eigenvalue pattern. Similar with the previous subsection, if the 58 regulatory relationships in the partial model of yeast are considered as the ground truth, we can obtain a realization of the ROC curve as the solid line in Figure 5.5(a). If we consider both the one-hop and two-hop regulatory relationships as the ground

truth, which contains 100 regulatory relationships, we will obtain the solid curve in Figure 5.5(b). If we consider all the one-hop, two-hop and three-hop regulatory relationships as the ground truth, which contains 156 regulatory relationships, we will obtain the solid curve in Figure 5.5(c). If we consider the 214 relationships less or equal to four-hop, we can observe the solid curve in Figure 5.5(d).

From Figure 5.5, we can observe that, when we consider one-hop and two-hop regulatory relationships, identification from eigenvalue pattern and direct correlation test have comparable performance. However, in the case where we consider regulatory relationships up to three-hop or four-hop, identification from eigenvalue pattern yields higher probability of correct detection, when the probability of false positive is required to be low.

The performance of both methods in Figure 5.5 is not satisfactory. One reason is, we only have one dataset in our current study. From this dataset, we are only able to obtain one realization of the ROC curve, which cannot fully characterize the statistics of the detection methods. In the future study, more datasets should be examined to statistically evaluate the effectiveness of the proposed method. Another reason is as follows. For both methods, we are considering only one regulator at one time. If there are two regulators affecting the same regulated gene (or the same set of regulated genes), the correlation between either one of the regulators and the regulated gene (or the eigenvalue pattern) could be weak. In the future study, we will enlarge the search space of regulatory relationships by considering several regulators simultaneously.

5.5 Chapter Summary

In this chapter, we briefly review the literature of modeling and discovering regulatory relationships, where the existing models only study the relationship between a set of regulators and one regulated gene. Motivated by the dependence model, we propose to use the eigenvalue pattern to characterize the relationship between one or more regulator and a set of regulated genes. We take a detour to build a mathematical foundation for the dependence model, proving several properties of the eigenvalue pattern and showing how a regulator can affect the regulated genes' eigenvalue pattern. Then, the proposed method is applied on cell-cycle microarray time-series data to identify regulatory relationships. The results are compared with identification using correlation test based on expression data only. The comparison shows that identifying regulatory relationships from eigenvalue pattern is able to better pick up evidence of multi-hop regulatory relationships.

This chapter is a preliminary effort of the identification of regulatory relationships. The main contribution is that we provide a new way to examine regulatory relationships, which considers several regulated genes simultaneously. Much future effort is required for improvement and verification of the proposed method.

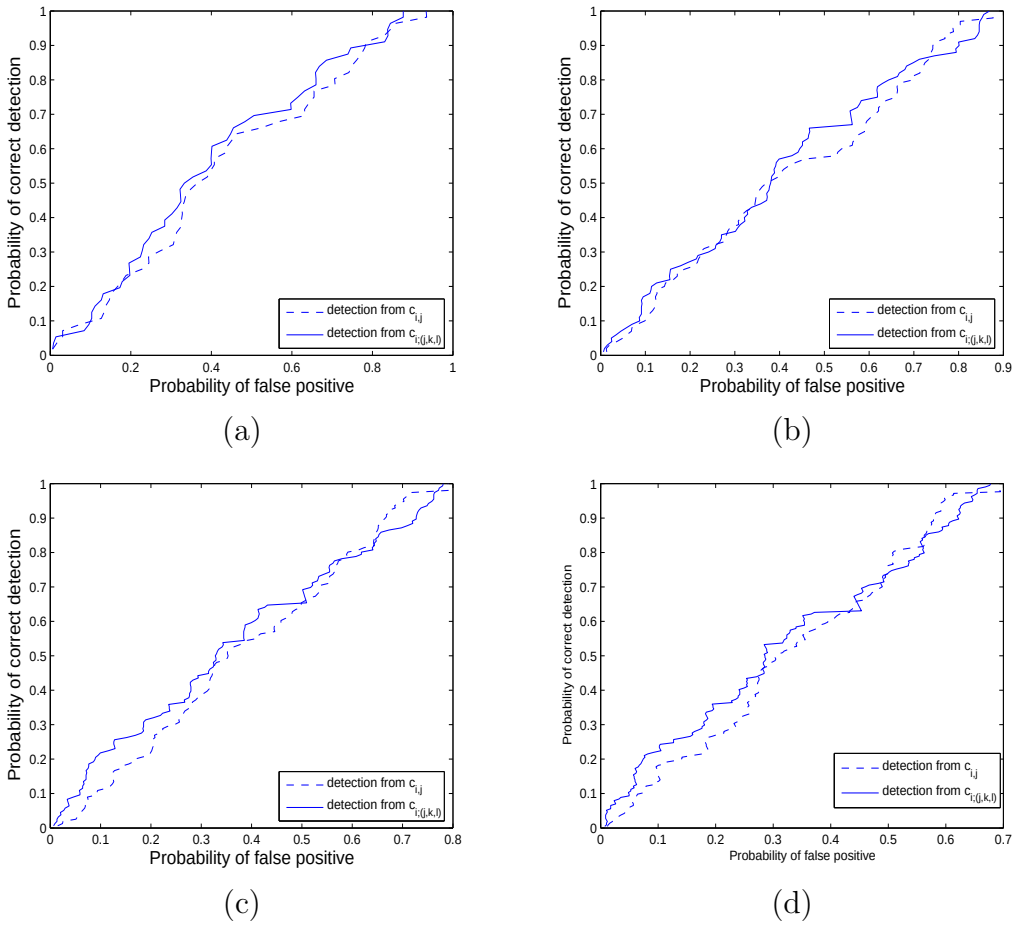


Figure 5.5: Comparison of ROC curves for two schemes: detecting regulatory relationships from time-lagged correlation between genes' expressions, detecting regulatory relationship from eigenvalue pattern. Fig (a) shows the case where 58 one-hop regulatory relationships from the partial model are considered as the ground truth. Fig (b) shows the cases where 100 regulatory relationships are regarded as the ground truth, containing both one-hop and two-hop regulatory relationships. Fig (c) considers 156 regulatory relationships as the ground truth, containing all relationships less or equal to three-hop. Fig (d) considers 214 regulatory relationships as the ground truth, containing all relationships less or equal to four-hop in the partial model.

Chapter 6

Conclusions and Future Research

6.1 Conclusions

Throughout this thesis, we are focusing on model-driven approaches for genomic and proteomic signal processing. In Chapter 2, we develop the dependence model, which is able to describe gene and protein's group behavior. With this model, we build a supervised classifier to study the big picture, the ensemble dependence relationship among gene clusters. The dependence model based classifier is used to classify normal and cancer samples based on gene clusters. Although we only examine the global dependence among clusters and throw away detailed information of individual genes, we are still able to obtain excellent classification performance. For the purpose of comparison, we examine the widely applied support vector machine algorithm. Although these two algorithms exhibit comparable performance, our algorithm presents a fundamental departure from the existing SVM approach. Because the dependence model develops a more plausible model by taking genes group

behaviors and interactions into account, and thus may have potential to classify intransigent data on which other classifiers balk.

An interesting observation is noted in the eigenvalue domain of the dependence model. Two distinguishing eigenvalue patterns of the dependence models are noted for the normal and cancer cases. By examining one prostate cancer dataset, we illustrate that the eigenvalue pattern goes through a continuous change from the perfect healthy case, to the normal case, to the early stage cancer case, and further to the late stage cancer case. The continuous change of eigenvalue pattern indicates that from normal case to cancer case, the ensemble dependence among gene clusters becomes weaker and weaker. Therefore, we conclude that the dependence model carries certain biology meaning that, the gene clusters are working more cooperatively in the normal case, while the gene clusters are working less cooperatively in the cancer case. This is the uniqueness of our approach, while the data-driven approaches cannot offer such biology meaning. Furthermore, since the eigenvalue pattern goes through a continuous change from normal to cancer, the eigenvalue pattern is promising for the early prediction of cancer development, and thus for potential cancer diagnosis usage.

After studying the big picture in Chapter 2, we zoom in to examine the detail relationship among individual gene and protein features in Chapter 3, where the dependence network is proposed. In building the dependence network, the dependence relationship among features can be indicated by the eigenvalue pattern. From binding triples found via the desired eigenvalue pattern, the dependence networks for both normal and cancer cases are built. These dependence networks

are essentially co-regulation networks, where connected features are co-regulated by some common factor. From the difference between dependence networks for normal and cancer cases, biomarkers are identified, which is called the dependence-network-based biomarkers. For the purpose of comparison, we examine a popular biomarker identification criterion in the literature, the classification-performance-based criterion. Based on results from both gene and protein expression data, we observe that the dependence-network-based approach provides much more consistent and reproducible results then the classification-performance-based approach. Another observation is that, the classification-performance-based biomarkers have high correlation with the simple differential method, such as T-test. However, the dependence-network-based criterion identifies many biomarkers that are not simply the most differentially expressed features. The results indicate that, the dependence-network-based biomarker identification criterion yields much more information than the simple differential method and the classification-performance-based criterion. Further, with the help of our collaborators in Georgetown University, we analyze the biological relevance of the identified gene biomarkers from a gastric cancer microarray dataset. Results show that the identified biomarkers are indeed biologically relevant. Several identified biomarkers have been shown to be valuable gastric cancer biomarkers in the literature. The encouraging results demonstrate that the proposed dependence model and network framework is able to facilitate discovery of better biomarkers for cancer research.

Chapter 2 and Chapter 3 both study the static gene and protein expression data, where different normal and cancer samples are taken from different individual

subjects while the time information of the normal and cancer samples is unknown. In Chapter 4 and Chapter 5, we address some challenges in microarray time-series data, where the data is obtained by measuring one sample at multiple time points. From time-series data, we are able to see how a gene system behave along time, and thus discovery the gene regulatory network, which represents causal relationships among genes.

In time-series experiment, the measurement is based on a population of synchronized cells, so that the observed expression is proportional to the single cell expression. However, even with the most advanced biochemical synchronization method, continuous synchronization loss is observed due to the diversity of individual cell growth rates. Therefore, there is an inherent problem of synchronization loss, which degrades the quality of the time-series data. In Chapter 4, we develop a model-based resynchronization framework to remove the effect of synchronization loss and reconstruct the underlying gene expression profiles, which represent single cell behavior more accurately. We consider a synchronization loss model where the gene expression measurements are regarded as superposition of mixed cell populations with different growth rates. The proposed scheme is shown feasible, promising and robust via simulations. Results from real microarray time-series data reveal that the proposed scheme is able to resynchronize the data. Comparisons with existing literature show that we are able to better discover cell-cycle regulated genes based on the resynchronized data. The significance of Chapter 4 is that, it presents an effective pre-processing method that greatly improves the quality of the time-series data.

In Chapter 5, we analyze time-series data to discover gene regulatory network. In the literature, there is a common property among existing methods, boolean network, Bayesian network, differential equations, etc. The relationship under investigation is always the relationship between one or several regulators and one regulated gene. To our knowledge, there is no method that considers several regulated genes together. In our study, we address this issue and provide a tool to examine the relationship between one or several regulators and several regulated genes. We propose to infer regulatory relationships based on the correlation between regulators and the regulated genes' group behavior (eigenvalue pattern). The proposed method is applied on cell-cycle microarray time-series data. The results are compared with identification using pair-wise correlation test based on expression data. The comparison shows that identifying regulatory relationships from eigenvalue pattern is able to better pick up evidence of multi-hop regulatory relationships. Chapter 5 is a preliminary effort of the identification of regulatory relationships. The main contribution is that we provide a new way to examine regulatory relationships, which considers several regulated genes simultaneously. Much future effort is required for improvement and verification of the proposed method.

As a summary, in this thesis, we propose novel model-driven approaches to address several topics in bioinformatics and cancer research, which are cancer classification and prediction, biomarker identification, time-series resynchronization, and regulatory network discovery. Excellent performance is obtained from the proposed dependence model, polynomial model, and their variations. Different from existing data-driven methods, the proposed dependence model carries certain biology

meaning and it has the potential for early prediction of cancer. The dependence-network-based biomarker identification criterion generates much more consistent and reproducible results than the popular classification-performance-based criterion in the literature. The polynomial approach is able to greatly improve the quality of microarray time-series data. Using the dependence model and its eigenvalues, we are able to examine regulatory relationships in a novel way that involve multiple regulated genes together.

6.2 Future Research

The research area of bioinformatics is developing fast. Every year, huge amount of new data is generated from high throughput technologies such as gene microarray and protein mass spectrum. These data are of various forms for different research purposes, and they require different computational tools. There are many interesting research directions that need to be further investigated.

First, for the purpose of cancer classification, cancer prediction and biomarker identification, we have developed the dependence model and the dependence network, and we have verified the proposed model in several gene and protein datasets. In the dependence model, we observed the continuous change of the eigenvalue pattern from normal cases to cancer cases. Such continuous change implies that the eigenvalue pattern has the potential for the early prediction of cancer. For our future work, we will obtain more datasets to further verify the proposed model, especially the eigenvalue pattern for early prediction of cancer. Our goal is to find a set of

meaningful biomarkers, based on which the dependence model can predict cancer development in the early stage.

Second, our current work for time-series resynchronization only utilize the microarray gene expression data. There is actually some side information available, such as the budding index data. Incorporating the side information may help to better resynchronize the time-series data and further improve the data quality.

Third, for the purpose of discovering a gene regulatory network, the dependence model and its eigenvalues are applied to study regulatory relationships that cannot be examined by the existing literature. Our main contribution is that we proposed a noval method to examine regulatory relationships that involves several regulated genes simultaneously. This is a preliminary effort of the identification of regulatory relationships. Much future effort is required for further improvement and verification. After the successful identification of regulatory relationships, the next question is how to integrate and assemble the identified regulatory relationships into one regulatory network in a biologically meaningful way. The knowledge of gene regulatory network will lead to the discovery of the signaling pathways of different biological processes and different diseases. Through pathway analysis, the discovered gene regulatory network and functional annotations could be integrated, so that our understanding of the mechanism of biology systems will be greatly improved.

Also, there are many other interesting bioinformatics problems that are closely related to this thesis. For example, the sequence information of genes and the locations of genes in the whole genome are closely related to their functionalities

and regulatory relationships. How to infer genes' functions and regulations from sequence information represents another horizon of the genomics research. In proteomic research, the sequence information and the secondary structures of proteins also provide valuable information of protein functionalities and interactions. How to infer protein secondary structures from sequence information, and how to infer functional interactions from protein secondary structures are both important problems.

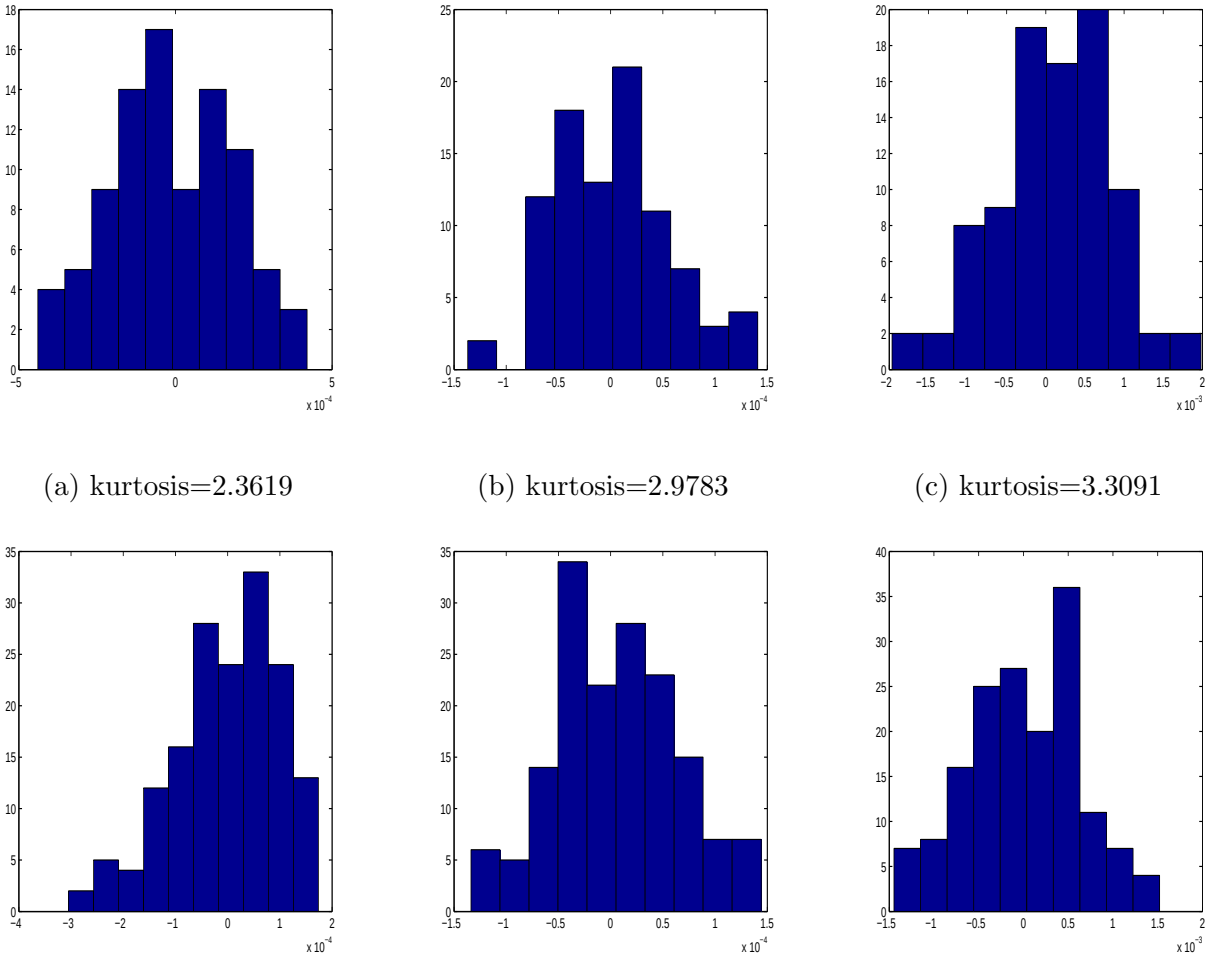
Overall, bioinformatics is an exciting interdisciplinary research area. A lot of research work is being conducted by biologist, computational scientists and statisticians, at different levels of abstractions. At each level of abstraction, signal processing and computational tools are needed to analyze and archive data effectively and systematically. Moreover, there is need for an effective way to integrate information from different data. With the background of electrical and computer engineering, we believe that signal processing will greatly facilitate and expedite bioinformatics research.

Appendix A

Gaussian Assumption in the Dependence Model

In the dependence model, the statistics of the noise-like term is assumed to be Gaussian. In order to evaluate this assumption, we examine the histograms and kurtosis of the noise-like term based on real gene and protein expression data. We use a protein mass spectrum dataset as an example, the ovarian cancer dataset. Here, we choose the number of clusters to be 3. Therefore, the protein features are clustered into 3 groups. In the two hypotheses in equation (2.7), the two noise-like terms are assumed to be two gaussian random vectors, both of dimension 3. If we look at these two gaussian random vectors element by element, draw the histogram and calculate the kurtosis (using matlab), we can get the following result in Figure A.1. As we know, the kurtosis of a gaussian random variable should be 3. From the histograms and kurtosis values, we can see that the noise term is roughly gaussian. Therefore, the gaussian assumption is used in our approach for its simplicity and

attractive properties. Note that the Gaussian assumption only affects the form of the maximum likelihood (ML) criterion of classification. It does not affect the theory of the dependence model.



(a) kurtosis=2.3619 (b) kurtosis=2.9783 (c) kurtosis=3.3091

(d) kurtosis=3.0729 (e) kurtosis=2.5098 (f) kurtosis=2.5770

Figure A.1: Fig (a) (b) (c) show the histograms for the noise term in normal case. Fig (d) (e) (f) show the histograms for the noise term in cancer case.

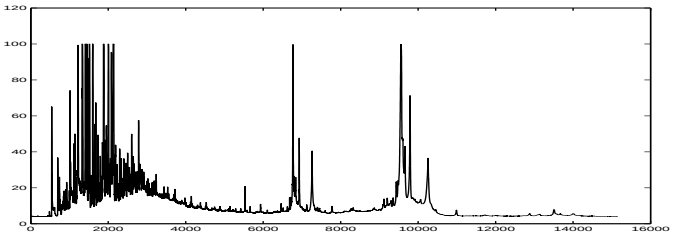
Appendix B

Pre-processing of Protein Mass Spectrum Data

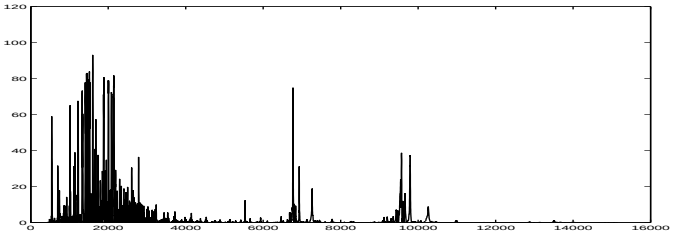
Because of the noisy nature of mass spectrum (MS) datasets, proper pre-processing of MS data such as spectrum smoothing, baseline identification and correction, and peak selection, is needed before applying the proposed model. Among the pre-processing steps, peak selection is essentially important, because peak selection aims at choosing a subset of peaks that are mostly associated with the phenotypes of interest. This step is similar to the common step of feature selection in classification applications. In this section, we briefly describe our preprocessing procedures.

Fig.B.1(a) is the raw MS data of one particular sample from the ovarian cancer dataset. The horizontal axis is mass-to-charge ratio (m/z), and the vertical axis corresponds to intensity. Similar to [59], we carried out the spectrum smoothing and baseline correction. We smooth the spectrum through the wavelet technique (using

matlab wavelet toolbox). Baseline correction removes baseline drift and background spectrum. In this step, the baseline is generated by a sliding window, which finds the lowest intensity within the sliding window. With window length being 10, the corrected spectrum is shown in Fig.B.1(b). After baseline correction, the spectra data are normalized by the mean intensity in each spectrum. Because of the spectra-shifting problem along the mass axis, we cannot use the intensities of particular mass values as features. Instead, peaks in the spectra are selected as features. Moreover, each peak does not correspond to a particular mass value. It corresponds to a certain range of mass values. In order to compare spectra from different experiments, peaks of spectra from different experiments must be aligned to the same mass ranges. In this study, since the spectra-shifting problem is not severe in the investigated datasets, peak detection and peak alignment are performed by detecting peaks from the average of normal spectra and the average of cancer spectra, respectively. In the ovarian dataset, there are 15,154 mass features. After peak detection and peak alignment, around 500 peaks are detected.



(a) Raw spectrum



(b) Spectrum after baseline correction

Figure B.1: Pre-processing of MS data.

Appendix C

Proof – Enginvalue of Ideal-case Dependence Model

In section 2.6, the eigenvalue pattern is discussed. It is claimed that for a more general noise free case where we have M clusters, the eigenvalues of the M-by-M matrix $\mathbf{A_{ideal}}$ are $\{1, 1, ..., 1, -(M-1)\}$, no matter what are the values of $\alpha_i, i = 1, 2, ..., M$. The proof is as following.

For the noise free case, the general form of dependence matrix for M clusters is as follows,

$$\mathbf{A_{ideal}} = \begin{pmatrix} 0 & \alpha_1 & \alpha_2 & \cdots & \alpha_{M-1} \\ \frac{1}{\alpha_1} & 0 & -\frac{\alpha_2}{\alpha_1} & \cdots & -\frac{\alpha_{M-1}}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & 0 & \cdots & -\frac{\alpha_{M-1}}{\alpha_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\alpha_{M-1}} & -\frac{\alpha_1}{\alpha_{M-1}} & -\frac{\alpha_2}{\alpha_{M-1}} & \cdots & 0 \end{pmatrix} \tag{C.1}$$

In order to solve for eigenvalues, we need to solve the determinant of $bfA_{ideal} - \lambda I$, and set it to be zero, which is

$$det \begin{pmatrix} -\lambda & \alpha_1 & \alpha_2 & \cdots & \alpha_{M-1} \\ \frac{1}{\alpha_1} & -\lambda & -\frac{\alpha_2}{\alpha_1} & \cdots & -\frac{\alpha_{M-1}}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & -\lambda & \cdots & -\frac{\alpha_{M-1}}{\alpha_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\alpha_{M-1}} & -\frac{\alpha_1}{\alpha_{M-1}} & -\frac{\alpha_2}{\alpha_{M-1}} & \cdots & -\lambda \end{pmatrix} = 0 \tag{C.2}$$

As we know, common factor from a column or a row can be pulled out of the determinant operator. So, we can pull the common factor α_1 from the second column, pull the common factor $\frac{1}{\alpha_1}$ from the second row, and thus get rid of parameter α_1. Then, we can pull the common factor α_2 from the third column, pull the common factor $\frac{1}{\alpha_2}$ from the third row, to get rid of parameter α_2. In the same way, all α parameters can be eliminated, and the left side can be simplified into

$$det \begin{pmatrix} -\lambda & \alpha_1 & \alpha_2 & \cdots & \alpha_{M-1} \\ \frac{1}{\alpha_1} & -\lambda & -\frac{\alpha_2}{\alpha_1} & \cdots & -\frac{\alpha_{M-1}}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & -\lambda & \cdots & -\frac{\alpha_{M-1}}{\alpha_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\alpha_{M-1}} & -\frac{\alpha_1}{\alpha_{M-1}} & -\frac{\alpha_2}{\alpha_{M-1}} & \cdots & -\lambda \end{pmatrix}$$

$$= \; det \begin{pmatrix} -\lambda & 1 & 1 & \cdots & 1 \\ 1 & -\lambda & -1 & \cdots & -1 \\ 1 & -1 & -\lambda & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{pmatrix} \tag{C.3}$$

After that, through some determinant invariant operation, such as adding column one to all other columns and subtracting rows 2 to $M-1$ from row one, the above matrix can be simplified.

$$det \begin{pmatrix} -\lambda & 1 & 1 & \cdots & 1 \\ 1 & -\lambda & -1 & \cdots & -1 \\ 1 & -1 & -\lambda & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & -1 & \cdots & -\lambda \end{pmatrix}$$

$$= det \begin{pmatrix} -\lambda & 1-\lambda & 1-\lambda & \cdots & 1-\lambda \\ 1 & 1-\lambda & 0 & \cdots & 0 \\ 1 & 0 & 1-\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1-\lambda \end{pmatrix}$$

$$
\begin{aligned}
&= det \begin{pmatrix} -(M-1)-\lambda & 0 & 0 & \cdots & 0 \\ 1 & 1-\lambda & 0 & \cdots & 0 \\ 1 & 0 & 1-\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1-\lambda \end{pmatrix} \\
&= -(M-1+1-\lambda)(1-\lambda)^{M-1}
\end{aligned} \tag{C.4}
$$

Easy to see, the eigenvalues of the dependence matrix $\mathbf{A_{ideal}}$ are $\{1, 1, ..., 1, -(M-1)\}$. Proof complete.

Appendix D

GO Terms of Identified Cell-Cycle Regulated Genes

Though genes identified by the proposed method have large overlap with previous studies, it is interesting to examine non-overlapping genes identified by the proposed method, but not identified in previous studies, neither [16] nor [21]. We examine the non-overlapping genes identified by the each scheme, through the semantic analysis based on the gene ontology (GO) terms. To achieve this purpose, we apply an online tool, the SGD Gene Ontology Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder). Given a list of genes, the GO Term Finder can output a list of associated GO terms, ranked by a p-value for each GO term.

We first analyze the set of genes identified by the proposed scheme only. From the results, we note that in the top 25 associated GO terms, there are several terms related to cell-cycle, such as "M phase", "cell-cycle", "mitotic cell cycle", and "M phase of mitotic cell cycle", associated with 84 genes. Moreover, the cell-cycle

related GO terms appears in top positions of the GO terms list. It suggests that some genes identified by the proposed scheme but not by other two schemes are cell-cycle related. For the other two schemes [16] and [21], it is noticed that none of the above cell cycle related GO terms appears in the top 25 GO terms list.

GO terms of cyclic genes identified by the proposed scheme only.

Gene Ontology term	Genes annotated to the term
chromosome segregation | AmiGO	BRN1, CSM1, SCC2, MCM21, LRS4, SMC2, CDH1, DOC1, ESP1, OKP1, CTF8, RSC2, CTF3, SGS1, RFC3
M phase | AmiGO	BRN1, CSM1, MSH5, SCC2, LRS4, SMC2, CDH1, SAE2, IME4, DOC1, ESP1, RIM4, KEL1, CTF8, MAD3, TOR1, MSC3, SGS1, DMA2, CLA4, MEI5, UME1, CCL1
cell cycle | AmiGO	FUS3, BRN1, PPS1, CSM1, MSH5, CLB3, SCC2, SAC7, LRS4, SMC2, CDH1, CKB1, SAE2, IME4, DOC1, ESP1, RIM4, KEL1, CTF8, MAD3, CBF1, TOR1, MSC3, SGS1, DMA2, CLA4, VHS3, MEI5, UME1, CCL1
myo-inositol metabolism | AmiGO	IPK1, SCS2, IRE1
mitotic sister chromatid segregation | AmiGO	BRN1, SCC2, SMC2, CDH1, DOC1, ESP1, CTF8, SGS1
sister chromatid segregation | AmiGO	BRN1, SCC2, SMC2, CDH1, DOC1, ESP1, CTF8, SGS1
biopolymer metabolism | AmiGO	CYC3, CNE1, FUS3, LSM2, AAR2, MNN2, MUM2, UBS1, TDP1, PRP5, PPS1, DPB3, STP22, SGF29, BUD31, TUP1, CSM1, SOL2/YCRX13W, PTC1, BPL1, MSH5, STE7, DHH1, FAP7, UFD2, SCC2, GPI8, LRS4, SPF1, MAK10, UTP7, SCS2, ATG18, CDH1, CKB1, STT3, PAN2, CEG1, SAE2, IME4, TAN1, DOC1, RAI1, GCN5, RIM4, VMA22, RPP1, PPE1, IRE1, LRP1, YHR087W, UBA4, ORC6, MPH1, MRS1, CBF1, FIP1, DBR1, MRS4, ORC3, KNS1, ARP6, MSC3, VPS34, RSC2, VPS36, NTR1, USA1, ARP9, UBX4, SGS1, RCE1, SIW14, ALG11, INP52, MGS1, KEX2, RFC3, CLA4, RCL1, DCP1, PET127, BUD21, NOP4, MEI5, SPT14, LEA1, PUF2, PRP4
cell organization and biogenesis | AmiGO	VPS8, ECM13, PET112, BRN1, ECM21, MRS5, SHE3, UBS1, DPB3, STP22, SGF29, BUD31, TUP1, SOL2/YCRX13W, PTC1, BUD30, FAP7, GLE1, CIS1, RAV2, MSS4, SAC7, LRS4, UTP7, SCS2, GLO3, COG3, ATG18, SMC2, RET2, CDH1, CKB1, PEX14, CSE1, DOC1, RAI1, YIP1, GCN5, VMA22, RPP1, LRP1, CDC12, ORC6, PEX28, KEL1, PEX18, FIS1, RHO3, MLP2, GEA1, CBF1, CCT5, TOR1, CDC11, MRS4, ORC3, SPA2, ENT4, ARP6, VPS34, RSC2, SFP1, VPS36, ARP9, STV1, SGS1, INP1, TRS130, FUS2, DYN3, HRB1, SIW14, INP52, DMA2, ATG2, CLA4, RCL1, WHI2, BUD21, SEY1, MYO2, NOP4, GYP5
mRNA metabolism | AmiGO	LSM2, AAR2, PRP5, DHH1, PAN2, CEG1, IME4, IRE1, LRP1, FIP1, NTR1, USA1, DCP1, LEA1, PUF2, PRP4
trehalose catabolism | AmiGO	NTH2, ATH1
double-strand break repair via homologous	SCC2, SAE2, LRP1, MEI5

recombination | AmiGO	
RNA processing | AmiGO	LSM2, AAR2, PRP5, BUD31, SOL2/YCRX13W, PTC1, FAP7, UTP7, PAN2, CEG1, RAI1, RPP1, IRE1, LRP1, MRS1, FIP1, MRS4, NTR1, USA1, RCL1, PET127, BUD21, NOP4, LEA1, PRP4
organelle organization and biogenesis | AmiGO	PET112, BRN1, MRS5, SHE3, DPB3, SGF29, BUD31, TUP1, PTC1, BUD30, FAP7, CIS1, MSS4, SAC7, LRS4, UTP7, SCS2, ATG18, SMC2, CDH1, CKB1, PEX14, DOC1, RAI1, GCN5, RPP1, LRP1, CDC12, ORC6, PEX28, PEX18, FIS1, RHO3, GEA1, CBF1, CCT5, TOR1, CDC11, ORC3, SPA2, ENT4, ARP6, VPS34, RSC2, SFP1, ARP9, SGS1, INP1, DYN3, SIW14, DMA2, ATG2, CLA4, RCL1, WHI2, BUD21, MYO2, NOP4
recombinational repair | AmiGO	SCC2, SAE2, LRP1, MEI5
cellular morphogenesis | AmiGO	BUD31, BUD30, CDH1, CKB1, CDC12, KEL1, RHO3, TOR1, CDC11, SPA2, SFP1, FUS2, CLA4, MYO2
morphogenesis | AmiGO	BUD31, BUD30, CDH1, CKB1, CDC12, KEL1, RHO3, TOR1, CDC11, SPA2, SFP1, FUS2, CLA4, MYO2
mitosis | AmiGO	BRN1, SCC2, SMC2, CDH1, DOC1, ESP1, KEL1, CTF8, MAD3, SGS1, DMA2, CLA4, CCL1
regulation of biological process | AmiGO	FUS3, PPS1, DPB3, TUP1, PTC1, CLB3, DHH1, RAV2, LRS4, SCS2, BUR6, CDH1, PDR1, CKB1, CEG1, DOC1, ESP1, ORC6, KEL1, MLP2, MAD3, TOR1, ORC3, SPA2, VPS36, LEU3, FUS2, DMA2, MGS1, CAF120, CLA4, DCP1, CIN5, WHI2, YRM1, ELP4, CCL1, PUF2
nucleobase, nucleoside, nucleotide and nucleic acid metabolism | AmiGO	LSM2, AAR2, MUM2, TFC1, TDP1, PRP5, DPB3, SGF29, BUD31, TUP1, CSM1, SOL2/YCRX13W, MED2, PTC1, MSH5, DHH1, FAP7, PDC2, SCC2, LRS4, UTP7, SCS2, BUR6, GNA1, PDR1, CKB1, PAN2, CEG1, SAE2, IME4, TAN1, RAI1, TFC4, GCN5, RIM4, MED6, RPP1, IRE1, LRP1, YHR087W, ORC6, DCD1, MLP2, MPH1, MRS1, CBF1, FIP1, DBR1, MRS4, ORC3, ARP6, MSC3, RSC2, SFP1, VPS36, NTR1, LEU3, USA1, ARP9, MOT3, SGS1, IDP3, MGS1, ADE12, CAF120, RFC3, RCL1, DCP1, PET127, CIN5, BUD21, YRM1, SYC1, RET1, NOP4, ELP4, MEI5, LEA1, CCL1, PUF2, PRP4
mitotic cell cycle | AmiGO	BRN1, PPS1, CLB3, SCC2, SMC2, CDH1, CKB1, DOC1, ESP1, KEL1, CTF8, MAD3, TOR1, SGS1, DMA2, CLA4, VHS3, CCL1
M phase of mitotic cell cycle | AmiGO	BRN1, SCC2, SMC2, CDH1, DOC1, ESP1, KEL1, CTF8, MAD3, SGS1, DMA2, CLA4, CCL1
monovalent inorganic cation homeostasis | AmiGO	RAV2, VMA22, TOK1, STV1, VHS3
meiosis I | AmiGO	CSM1, MSH5, LRS4, SAE2, RIM4, MSC3, MEI5
premeiotic DNA synthesis | AmiGO	MUM2, RIM4
cell ion homeostasis | AmiGO	CSG2, RAV2, SPF1, CKB1, VMA22, TOK1, STV1, MMT1, ATM1, VHS3

GO terms of cyclic genes identified by the scheme in (Spellman *et al*, 1998) only

Gene Ontology term	Genes annotated to the term
response to pheromone during conjugation with cellular fusion | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SAG1, STE3, MDG1, AGA1, RGA1, MF(ALPHA)1
response to pheromone | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SAG1, STE3, SST2, MDG1, AGA1, RGA1, MF(ALPHA)1
transport | AmiGO	SEO1, PET9, PHO89, ATG22, VCX1, PMP3, NPL3, SIT1, FTR1, NIC96, VAM7, ZRT1, TNA1, ARN1, HXT4, HXT1, SEC28, POR2, QDR2, TPM2, AVT1, CYC1, STE6, GAP1, MMP1, ERG3, FRE1, FKS1, SUR4, ATR1, NDI1, PHO84, SSO2, FAA4, PET8, PDR16, MCH4, UFE1, TRS33, VPH1, TPO4, FAA1, VMA4, FIT3, PDR12, SSO1, DIP5, KAR9, SAM3, MEP3
establishment of localization | AmiGO	SEO1, PET9, PHO89, ATG22, VCX1, PMP3, NPL3, SIT1, FTR1, NIC96, VAM7, ZRT1, TNA1, ARN1, HXT4, HXT1, SEC28, POR2, QDR2, TPM2, AVT1, CYC1, STE6, GAP1, MMP1, ERG3, FRE1, FKS1, SUR4, ATR1, NDI1, PHO84, SSO2, FAA4, PET8, PDR16, MCH4, UFE1, TRS33, VPH1, TPO4, FAA1, VMA4, FIT3, PDR12, SSO1, DIP5, KAR9, SAM3, MEP3
localization | AmiGO	SEO1, PET9, PHO89, ATG22, VCX1, PMP3, NPL3, SIT1, FTR1, NIC96, OLE1, VAM7, ZRT1, TNA1, ARN1, HXT4, HXT1, SEC28, POR2, QDR2, TPM2, AVT1, CYC1, STE6, GAP1, MMP1, ERG3, FRE1, FKS1, SUR4, ATR1, NDI1, PHO84, SSO2, FAA4, PET8, PDR16, MCH4, UFE1, TRS33, VPH1, TPO4, FAA1, VMA4, FIT3, PDR12, SSO1, DIP5, KAR9, SAM3, MEP3
sexual reproduction | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, MF(ALPHA)1
conjugation with cellular fusion | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, MF(ALPHA)1
conjugation | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, MF(ALPHA)1
interaction between organisms | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, MF(ALPHA)1
amino acid catabolism | AmiGO	ARO10, BAT1, PUT1, CAR2, GCV2, CAR1
amine catabolism | AmiGO	ARO10, BAT1, PUT1, CAR2, GCV2, CAR1
nitrogen	ARO10, BAT1, PUT1, CAR2, GCV2, CAR1

compound catabolism | AmiGO	
ion transport | AmiGO	PHO89, VCX1, PMP3, SIT1, FTR1, ZRT1, ARN1, POR2, FRE1, PHO84, MEP3
fatty acid biosynthesis | AmiGO	HTD2, ELO1, FAS1, SUR4
sulfur amino acid transport | AmiGO	MMP1, PET8, SAM3
organic acid metabolism | AmiGO	ARO10, OLE1, MET13, ASN2, HTD2, ARO9, BAT1, YIL168W, ELO1, FAS1, PUT1, SUR4, CAR2, GCV2, IDH1, LYS9, ARG1, PDR12, CAR1, MET16
amine transport | AmiGO	AVT1, GAP1, MMP1, PET8, TPO4, DIP5, SAM3
carboxylic acid metabolism | AmiGO	ARO10, OLE1, MET13, ASN2, HTD2, ARO9, BAT1, YIL168W, ELO1, FAS1, PUT1, SUR4, CAR2, GCV2, IDH1, LYS9, ARG1, PDR12, CAR1, MET16
reproductive cellular physiological process | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, SSP2, MUM3, SPS4, MF(ALPHA)1
reproductive physiological process | AmiGO	FUS1, MFA1, STE2, MF(ALPHA)2, SCW4, SAG1, STE3, SST2, MDG1, AGA1, RGA1, SSP2, MUM3, SPS4, MF(ALPHA)1
organic acid transport | AmiGO	AVT1, GAP1, MMP1, PET8, PDR12, DIP5, SAM3
di-, tri-valent inorganic cation transport | AmiGO	VCX1, SIT1, FTR1, ZRT1, ARN1, FRE1, PHO84
amino acid transport | AmiGO	AVT1, GAP1, MMP1, PET8, DIP5, SAM3
sterol metabolism | AmiGO	YEH1, ERG3, ERG27, ERG2, CYB5, PDR16
lipid metabolism | AmiGO	OLE1, HTD2, ELO1, FAS1, YEH1, ERG3, ERG27, SUR4, ERG2, FAA4, CYB5, PSD1, PDR16, IZH4, MUM3, FAA1

GO terms of cyclic genes identified by the scheme in (Lu *et al,* 2004) only.

Gene Ontology term	Genes annotated to the term
glycoprotein biosynthesis | AmiGO	MNT2, MNN5, MNN4, KTR2, YEH2, SEC59, MNT4, ALG5
glycoprotein metabolism | AmiGO	MNT2, MNN5, MNN4, KTR2, YEH2, SEC59, MNT4, ALG5
double-strand break repair via homologous recombination | AmiGO	RAD57, SCC4, YKU80, RAD50
response to endogenous stimulus | AmiGO	PIN4, RAD57, NSE3, SCC4, MLH2, PSY3, NSE5, YKU80, RAD50, AZF1, RFC1, REV1, SKS1
recombinational repair | AmiGO	RAD57, SCC4, YKU80, RAD50
biopolymer glycosylation | AmiGO	MNT2, MNN5, MNN4, KTR2, SEC59, MNT4, ALG5
protein amino acid glycosylation | AmiGO	MNT2, MNN5, MNN4, KTR2, SEC59, MNT4, ALG5
mitochondrial fission | AmiGO	MDV1, DNM1
organelle fission | AmiGO	MDV1, DNM1
protein targeting to peroxisome | AmiGO	PEX5, PEX12, PEX25
mitochondrial genome maintenance | AmiGO	RIM2, MMF1, MDV1, ABF2
DNA repair | AmiGO	RAD57, NSE3, SCC4, MLH2, PSY3, NSE5, YKU80, RAD50, RFC1, REV1
protein amino acid O-linked glycosylation | AmiGO	MNT2, MNN4, MNT4
response to DNA damage stimulus | AmiGO	PIN4, RAD57, NSE3, SCC4, MLH2, PSY3, NSE5, YKU80, RAD50, RFC1, REV1
protein import into peroxisome matrix | AmiGO	PEX12, PEX25
response to carbohydrate stimulus | AmiGO	AZF1, SKS1
response to organic substance | AmiGO	AZF1, SKS1
reproductive cellular physiological process | AmiGO	SWF1, EMI2, SPO73, SPO74, KAR2, CSN12, SPO75, FAR10, TUB1, FAR8, PRM4

reproductive physiological process | AmiGO	SWF1, EMI2, SPO73, SPO74, KAR2, CSN12, SPO75, FAR10, TUB1, FAR8, PRM4
response to stimulus | AmiGO	PIN4, KIN82, RAD57, PLP1, NSE3, SCC4, QDR1, KAR2, CSN12, MSN4, MNN4, MLH2, FAR10, PSY3, NSE5, FAR8, YKU80, NST1, ZWF1, RAD50, AZF1, RFC1, REV1, SKS1
peroxisome organization and biogenesis | AmiGO	PEX5, VPS1, PEX12, PEX25
cell cycle arrest in response to pheromone | AmiGO	FAR10, FAR8
protein amino acid palmitoylation | AmiGO	SWF1, ERF2
protein palmitoylation | AmiGO	SWF1, ERF2
organelle organization and biogenesis | AmiGO	UTP20, SMY2, RIM2, TAF5, LSB5, RAD57, SWF1, PEX5, UTP4, ESC2, HPA3, MMF1, ICE2, KAR2, NET1, MDV1, SWI3, NUP120, VPS1, SET3, LAS1, RPF2, DNM1, YPT7, TUB1, UTP14, QRI8, PEX12, ABF2, YKU80, VAC7, RAD50, RIO1, RRS1, PEX25, RRP9, RHO1

BIBLIOGRAPHY

[1] Lockhart,D. and Winzeler,E., "Genomics, gene expression and DNA arrays", *Nature*, 405(6788):827-846, 2000.

[2] Diamandis,E., "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations", *Mol. Cell Proteomics*, 3(4):367-378, 2004.

[3] Chang,J., Wooten,E., Tsimelzon,A., Hilsenbeck,S., Gutierrez,M., Elledge,R., Mohsin,S., Osborne,C., Chamness,G., Allred,D., and O'Connell,P., "Gene Expression Profiling for the Prediction of Therapeutic Response to Docetaxel in Patients with Breast Cancer", *Mechanisms of Disease*, 362(9381):362-369, 2003.

[4] Van't Veer,L., Dai,H., Van De Vijver,M., "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer", *Nature*, 415(6871):530-536, 2002.

[5] Eisen,M., Spellman,P., Brown,P., and Botstein,D., "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci. USA*, 95(25):14863-14668, 1998.

[6] Wu,X., Chen,Y., Bernard,R, Yan,A., "The local maximun clustering method and its application in microarray gene expression data analysis", *EURASIP Journal on Applied Signal Proc*, 1:51-61, 2004.

[7] Kohonen,T., *Self-organizing Maps*. Springer, Berlin, 1997.

[8] Tavazoie,S., Hughes,J., Campbell,M., Cho,R. and Church,G., "Systematic determination of genetic network architecture", *Nat. Genet.*, 22(3):281-285, 1999.

[9] Duda,R.O., Hart, P.E. and Stork, D.G. *Pattern Classification. Second ed.*, 2001.

[10] Furey,T., Cristinanini,N., Duffy,N., Bednarski,D., Schmmer,M. and Haussler,D., "Support vector machine classification and validation of cancer tissue samples using microarray expression data", *Bioinformatics*, 16(10):906-914, 2000.

[11] O'Neill,M. and Song,L., "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect", *BMC Bioinformatics*, 4:28-41, 2003.

[12] Frank,R. and Hargreaves,R., "Clinical Biomarkers in Drug Discovery and Development", *Nat Rev Drug Discov.*, 2(7):566-580, 2003.

[13] Li,J., Zhang,Z., Rosenzweig,J., Wang,Y., and Chan,D., "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer", *Clin Chem*, 48(8):1296-1304, 2002.

[14] Ressom,H., Varghese,R., Abdel-Hamid,M., Eissa,S., Saha,D., Goldman,L., Petricoin,E., Conrads,T., Veenstra,T., Loffredo,C., and Goldman,R., "Analysis of mass spectral serum profiles for biomarker selection", *Bioinformatics*, 21(21):4039-4045, 2005.

[15] Lee,T., Rinaldi,N., Robert,F., Odom,D., Bar-Joseph,Z., Gerber,G., Hannett,N., Harbison,C., Thompson,C., Simon,I., Zeitlinger,J., Jennings,E., Murray,H., Gordon,D., Ren,B., Wyrick,J., Tagne,J., Volkert,T., Fraenkel,E., Gifford,D., and Young,R., "Transcriptional Regulatory Networks in Saccharomyces cerevisiae", *Science*, 298(5594):799-804, 2002.

[16] Spellman,P., Sherlock,G., Zhang,M., Iyer,V., Anders,K., Eisen,M., Brown,P., Botstein,D., and Futcher,B., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisia by microarray hybridization", *Mol. Biol. Cell*, 9(12):3273-3297, 1998.

[17] Shedden,K., and Cooper,S., "Analysis of cell-cycle gene expression in Saccharmoyces cerevisiae using microarray and multiple synchronization methods", *Nucleic Acids Research*, 30(13):2920-2929, 2002.

[18] Johansson,D., Lindgren,P., and Berglund,A., "A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription", *Bioinformatics*, 19(4):467-473, 2003.

[19] Whitfield,M., Sherlock,G., Saldanha,A., Murray,J., Ball,C., Alexander,K., Matese,J., Perou,C., Hurt,M., Brown,P., and Botstein,D., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors", *Mol. Biol. Cell*, 13(6):1977-2000, 2002.

[20] Wichert,S., Fokianos,K., and Strimmer,K., "Identifying periodically expressed transcripts in microarray time series data", *Bioinformatics*, 20(1):5-20, 2004.

[21] Lu,X., Zhang,W., Qin,Z., Kwast,K., and Liu,J., "Statistical Resynchronization and Bayesian detection of periodically expressed genes", *Nucleic Acids Research*, 32(2):447-455, 2004.

[22] Bar-Joseph,Z., Farkash,S., Gifford,D., Simon,I., and Rosenfeld,R., "Deconvolv ing cell cycle expression data with complementary information", *Bioinformatics*, 20(Suppl.1):I23-I30, 2004.

[23] Kauffman,S., "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets", *J. Theor. Biol.,*, 22(3):429-467, 1969.

[24] Akutsu,T., Miyano,S., and Kuhara,S., "Inferring Qualitative Relations in Genetic Networks and Metabolic Pathways", *Bioinformatics*, 16(8):727-734, 2000.

[25] Shmulevich,I., Dougherty,E., Kim,S., and Zhang,W., "Probabilistic Boolean Networks: a Rule-based Uncertainty Model for Gene Regulatory Networks", *Bioinformatics*, 18(2):261-274, 2002.

[26] Hashimoto,R., Kim,S., Shmulevich,I., Zhang,W., Bittner,M., and Dougherty,E., "Growing genetic regulatory networks from seed genes", *Bioinformatics*, 20(8):1241-1247, 2004.

[27] Friedman,N., Linial,M., Nachman,I., and Pe'er,D., "Using Bayesian networks to analyze expression data", *J. Comput. Biol.*, 7(3-4):601-620, 2000.

[28] Friedman,N., Murphy,K., and Russell,S., "Learning the structure of dynamic probabilistic networks", in Proceedings of the 14th Conference on the Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, pp. 139-147, 1998.

[29] Kim,S., Imoto,S., and Miyano,S. "Inferring gene networks from time series microarray data using dynamic Bayesian networks", *Brief Bioinform.*, 4(3):228-235, 2003.

[30] Perrin,B., Ralaivola,L., Mazurie,A., Bottani,S., Mallet,J., and D'Alche-Buc,F, "Gene networks inference using dynamic Bayesian networks", *Bioinformatics*, 19(Suppl.2):II138-II148, 2003.

[31] Zou,M., and Conzen,S., "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data", *Bioinformatics*, 21(1):71-79, 2005.

[32] Pena,J., Bjorkegren,J., and Tegner,J., "Learning dynamic Bayesian network models via cross-validation", *Pattern Recognition Letters*, 26(14):2295-2308, 2005.

[33] Chen,T., He,H., and Church,G., "Modeling gene expression with differential equations", *Proc. Pac. Symp. on Biocomputing* 4:29-40, 1999.

[34] Sakamoto,E., and Iba,H., "Evolutionary inference of a biological network as differential equations by genetic programming", *Genome Informatics*, 12:276-277, 2001.

[35] Ando,S., and Iba,H., "Estimation of gene regulatory network by genetic algorithm and pairwise correlation analysis", *CEC '03. The 2003 Congress on Evolutionary Computation* , 1:207-214, 2003.

[36] De Hoon,M., Imoto,S., and Miyano,S., "Inferring gene regulatory networks from time-ordered gene expression data using differential equations", *Lecture Notes in Computer Science*, 2534:267-274, 2002.

[37] De Hoon,M., Imoto,S., Kobayashi,K., Ogasawara,N., and Miyano,S., "Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations", *Proc. Pac. Symp. on Biocomputing*, 17-28, 2003.

[38] Chen,K., Wang,T., Tseng,H., Huang,C., and Kao,C., "A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae", *Bioinformatics*, 21(12):2883-2890, 2005.

[39] Li,X., Rao,S., Jiang,W., Li,C., Xiao,Y., Guo,Z., Zhang,Q., Wang,L., Du,L., Li,J., Li,L., Zhang,T., and Wang,Q., "Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling", *BMC Bioinformatics*, 7:26, 2006.

[40] Zhao,W., Serpedin,E., and Dougherty,E., "Inferring gene regulatory networks from time series data using the minimum description length principle", *Bioinformatics*, 22(17):2129-2135, 2006.

[41] Woolf,P., and Wang,Y., "A fuzzy logic approach to analyzing gene expression data", *Physiological Genomics*, 3(1):9-15, 2000.

[42] Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligirui,M., Bloomfield,C. and Lander,E., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286(5439):531-537, 1999.

[43] Slonim,D., Tamayo,P., Mesirov,J., Golub,T. and Lander,E., "Class prediction and discovery using gene expression data", *RECOMB*, 263-272, 2000.

[44] Steinhoff,C., Muller,T., Nuber,U. and Vingron,M., "Gaussian mixture density estimation applied to microarray data", *RECOMB*, 147, 2003.

[45] H. Vincent Poor, *An Introduction to Signal Detection and Estimation*. Springer Texts in Electrical Engineering, 1994.

[46] Chen,X., Leung,S., Yuen,S., Chu,K., Ji,J., Li,R., Chan,A., Law,S., Troyanskaya,O., Wong, J., So,S., Botstein,D., and Brown,P., "Variation in gene expression patterns in human gastric cancers", *Mol. Cell Biol.*, 14(8):3208-3215, 2003.

[47] Chen,X., Cheung,S., So,S., Fan,S., Barry,C., Higgins,J., Lai,K., Ji,J., Dudoit,S., Irene,O., Rijn,M., Botstein,D., and Brown,P., "Gene expression patterns in human liver cancers", *Mol. Cell Biol.*, 13(6):1929-1939, 2002.

[48] Dhanasekaran,S., Barrette,T, Ghosh,D., Shah,R., Varambally,S., Kurachi,K., Pienta,K., Rubin,M., and Chinnaiyan,A., "Delineation of prognostic biomarkers in prostate cancer", *Nature*, 412(6849):822-826, 2001.

[49] Wong,Y, Selvanayagam,Z., Wei,N., Porter,J., Vittal,R., Hu,R., Lin,Y., Liao,J., Shih,J., Cheung,T., Lo,K., Yim,S., Yip,S., Ngong,D., Siu,N., Chan,L., Chan,C., Kong,T., Kutlina,E., McKinnon,R., Denhardt,D., Chin,K., and Chung,T., "Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray", *Clinical Cancer Research*, 9(15):5486-5492, 2003

[50] Garber,M., Troyanskaya,O., Schluens,K., Petersen,S., Thaesler,Z., Gengelbach,M., Rijn,M., Rosen,G., Perou,C., Whyte,R., Altman,R., Brown,P., Botstein,D., and Petersen,I., "Diversity of gene expression in adenocarcinoma of the lung". *Proceedings of the national academy of science of USA*, 98(24):12784-12789, 2001.

[51] Alon,U., Barkai,N., Notterman,D., Gish,K., Mach,S. and Levine,J., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Sci. USA*, 96(12):6745-6750, 1999.

[52] Singh,D., Febbo,P., Ross,K., Jackson,D., Manola,J., Ladd,C., Tamayo,P., Renshaw,A., D'Amico,A., Richie,J., Lander,E., Loda,M., Kantoff,P., Golub,T., and Sellers,W., "Gene expression correlates of clinical prostate cancer behavior", *Cancer Cell*, 1(2):203-209, 2002.

[53] Gordon,G., Jensen,R., Hsiao,L., Gullans,S., Blumenstock,J., Ramaswamy,S., Richards,W., Sugarbaker,D., and Bueno,R., "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", *Cancer Research*, 62(17):4963-4967, 2002.

[54] Antoniadis,A., Lambert-Lacroix,S. and Leblanc,F., "Effective dimension reduction methods for tumor classification using gene expression data", *Bioinformatics*, 19(5):563-570, 2003.

[55] Jonsson,K., Kittler,J., Li,Y., and Matas,Y., "Support vector machines for face authentication", *Journal of Image and Vision Computing*, 20(5-6):369-375, 2002.

[56] Dong,X., and Zhaohui,W., "Speaker recognition using continuous density support vector machines", *Electronics Letters*, 37(17):1099-1101, 2001.

[57] Choisy,C., and Belaid,A., "Handwriting recognition using local methods for normaliztion and global methods for recognition", *In Proceedings of Sixth Int. Conference On Document Analysis and Recognition*, 23-27, 2001.

[58] Rifkin,R., Mukherjee,S., Tamayo,P., Ramaswamy,S., Yeang,C., Angelo,M., Reich,M., Poggio,T., Lander,E., Golub,T., and Mesirov,J., "An Analytical Method For Multi-class Molecular Cancer Classification", *SIAM Review*, 45(4):706-723, 2003.

[59] Liu,Q., Krishnapuram,B., Pratapa,P., Liao,X., Hartemink,A., and Carin,L., "Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra", *ASILOMAR Conference: Biological Aspects of Signal Processing*, November 2003.

[60] Fisher,R., "The Use of Multiple Measurements in Taxonomic Problems" *Annals of Eugenics* 7, 179-188, 1936.

[61] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Ensemble Dependence Model-based Cancer Classification using Gene Microarray Data", *IEEE International Workshop on Genomic Signal Processing and Statistics, GENSIPS*, 2005.

[62] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Ensemble dependence model for cancer classification and prediction of cancer and normal gene expression data", *Bioinformatics*, 21(14):3114-3121, 2005.

[63] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Dependence Model and Network for Biomarker Identification and Cancer Classification", *14th European Signal Processing Conference, EUSIPCO*, 2006.

[64] Walhout,A., and Vidal,M., "Protein interaction maps for model organisms", *Nat Rev, Mol Cell Biol.*, 2(1):55-62, 2001.

[65] Tibshirani,R., Hastie,T., Narasimhan,B., Soltys,S., Shi,G., Koong,A., and Le,Q., "Sample classification from protein mass spectrometry, by peak probability contrasts", *bioinformatics*, 20(17):3034-3044, 2004.

[66] Adam,B., Qu,Y., Davis,J., Ward,M., Clements,M., Cazares,L., Semmes,O., Schellhammer,P., Yasui,Y., Feng,Z., and Wright,G., "Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men" *Cancer Research*, 62(13):3609-3614, 2002.

[67] Wu,C., Apweiler,R., Bairoch,A., Natale,D., Barker,W., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M., Mazumder,R., O'Donovan,C., Redaschi,N., and Suzek,B., "The Universal Protein Resource (UniProt): an expanding universe of protein information", *Nucleic Acids Res.*, 1;34(Database issue):D187-191, 2006.

[68] Framson,P., and Sage,E., "SPARC and tumor growth: where the seed meets the soil", *J Cell Biochem.*, 92(4):679-690, 2004.

[69] Maeng,H., Song,S., Choi,D., Kim,K., Jeong,H., Sakaki,Y., and Furihata,C., "Osteonectin-expressing cells in human stomach cancer and their possible clinical significance.", *Cancer Lett.*, 184(1):117-121, 2002b.

[70] Inoue,H., Matsuyama,A., Mimori,K., Ueo,H., and Mori,M., "Prognostic score of gastric cancer determined by cDNA microarray", *Clin Cancer Res.*, 8(11):3475-3479, 2002.

[71] Bosserhoff,A., "Novel biomarkers in malignant melanoma", *Clin Chim Acta.* 367(1-2):28-35, 2006.

[72] Thomas,R., True,L., Bassuk,J., Lange,P., and Vessella,R., "Differential expression of osteonectin/SPARC during human prostate cancer progression", *Clin Cancer Res.*, 6(3):1140-1149, 2000.

[73] Hippo,Y., Taniguchi,H., Tsutsumi,S., Machida,N., Chong,J., Fukayama,M., Kodama,T., and Aburatani,H., "Global gene expression analysis of gastric cancer by oligonucleotide microarrays", *Cancer Res.*, 62(1):233-240, 2002.

[74] Maeng,H., Choi,D., Takeuchi,M., Yamamoto,M., Tominaga,M., Tsukamoto,T., Tatematsu,M., Ito,T., Sakaki,Y., and Furihata,C., "Appearance of osteonectin-

expressing fibroblastic cells in early rat stomach carcinogenesis and stomach tumors induced with N-methyl-N'-nitro-N-nitrosoguanidine", *Jpn J Cancer Res.*, 93(9):960-967, 2002a.

[75] Blaveri,E., Simko,J., Korkola,J., Brewer,J., Baehner,F., Mehta,K., Devries,S., Koppie,T., Pejavar,S., Carroll,P., and Waldman,F., "Bladder cancer outcome and subtype classification by gene expression", *Clin Cancer Res.*, 11(11):4044-4055, 2005.

[76] Lai,J., Yu,C., Moser,C., Aderca,I., Han,T., Garvey,T., Murphy,L., Garrity-Park,M., Shridhar,V., Adjei,A., and Roberts,L., "SULF1 Inhibits Tumor Growth and Potentiates the Effects of Histone Deacetylase Inhibitors in Hepatocellular Carcinoma", *Gastroenterology*, 130(7):2130-2144, 2006.

[77] Wakasugi,K., and Schimmel,P., "Two distinct cytokines released from a human aminoacyl-tRNA synthetase", *Science*, 284(5411):147-151, 1999.

[78] Tzima,E., and Schimmel,P., "Inhibition of tumor angiogenesis by a natural fragment of a tRNA synthetase", *Trends Biochem Sci.*, 31(1):7-10, 2006.

[79] Ivakhno,S., and Kornelyuk,A., "Cytokine-like activities of some aminoacyl-tRNA synthetases and auxiliary p43 cofactor of aminoacylation reaction and their role in oncogenesis", *Exp Oncol.*, 26(4):250-255, 2004.

[80] Kubo,Y., Sekiya,S., Ohigashi,M., Takenaka,C., Tamura,K., Nada,S., Nishi,T., Yamamoto,A., and Yamaguchi,A., "ABCA5 resides in lysosomes, and ABCA5 knockout mice develop lysosomal disease-like symptoms", *Mol Cell Biol.*, 25(10):4138-4149, 2005.

[81] Haeryfar,S., and Hoskin,D., "Thy-1: more than a mouse pan-T cell marker", *J Immunol.*, 173(6):3581-3588, 2004.

[82] Rege,T., and Hagood,J., "Thy-1 as a regulator of cell-cell and cell-matrix interactions in axon regeneration, apoptosis, adhesion, migration, cancer, and fibrosis", *FASEB Journal*, 20(8):1045-1054, 2006.

[83] Duxbury,M., Ashley,S., and Whang,E., "RNA interference: a mammalian SID-1 homologue enhances siRNA uptake and gene silencing efficacy in human cells", *Biochem Biophys Res Commun.*, 331(2):459-463, 2005.

[84] Qiu,P., Wang,Z.J., Liu,K.J.R., Hu,Z.Z., Wu,C., "Dependence Network Modeling for Biomarker Identification", *Bioinformatics*, 23(2):198-206, 2007.

[85] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Genomic Processing for Cancer Classification and Prediction", *IEEE signal processing magazine*, 24(1):100-110, Jan, 2007.

[86] Stoer,J., and Bulirsch,R., "Introduction to Numerical Analysis", *Springer*, 1991.

[87] Poor,V., "An Introduction to Signal Detection and Estimation", *Springer*, 1994.

[88] Cho,R., Campbell,M., Winzeler,E., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T., Gabrielian,A., Landsman,D., Lockhart,D., and Davis,R., "A genome-wide transcriptional analysis of the mitotic cell cycle", *Mol. Biol. Cell*, 2(1):6573, 1998.

[89] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Tracking the Herd: Resynchronization Analysis of Cell-Cycle Gene Expression Data in Saccharomyces Cerevisiae", *Engineering in Medicine and Biology Society, EMBS*, 4826-4829, 2005.

[90] Qiu,P., Wang,Z.J., and Liu,K.J.R., "Polynomial Model Approach for Resynchronization Analysis of Cell-cycle Gene Expression Data", *Bioinformatics*, 22(8):959-966, 2006.

[91] Kauffman,S., Peterson,C., Samuelsson,B., and Troein,C., "Random Boolean Network Models and the Yeast Transcriptional Network", *Proc. National Academy of Science of the USA*, 100(25):14796-14799, 2003.

[92] Schmitt,W., Raab,R., and Stephanopoulos,G., "Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data", *Genome Res.*, 14(8):1654-1663, 2004.

[93] Ji,L., and Tan,K., "Identifying time-lagged gene clusters using gene expression data", *Bioinformatics*, 21(4):509-516, 2005.

[94] Gupta,A., Maranas,C., and Albert,R., "Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites", *Bioinformatics*, 22(2):209-214, 2006.