Integrative Modeling and Analysis of High-throughput Biological Data

Li Chen

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In
Computer Engineering

Jason J. Xuan, Chair

Chang-Tien Lu

Christopher L. Wyatt

Scott F. Midkiff

Yue J. Wang

December 15th, 2010
Arlington, Virginia

Keywords: Microarray Data Analysis, Transcriptional Regulatory Network, Biomarker Identification, Support Vector Regression, Markov Random Field, Support Vector Machine

Integrative Modeling and Analysis of High-throughput Biological Data

Li Chen

ABSTRACT

Computational biology is an interdisplinary field that focuses on developing mathematical models and algorithms to interpret biological data so as to understand biological problems. With current high-throughput technology development, different types of biological data can be measured in a large scale, which calls for more sophisticated computational methods to analyze and interpret the data. In this dissertation research work, we propose novel methods to integrate, model and analyze multiple biological data, including microarray gene expression data, protein-DNA interaction data and protein-protein interaction data. These methods will help improve our understanding of biological systems.

First, we propose a knowledge-guided multi-scale independent component analysis (ICA) method for biomarker identification on time course microarray data. Guided by a knowledge gene pool related to a specific disease under study, the method can determine disease relevant biological components from ICA modes and then identify biologically meaningful markers related to the specific disease. We have applied the proposed method to yeast cell cycle microarray data and Rsf-1-induced ovarian cancer microarray data. The results show that our knowledge-guided ICA approach can extract biologically meaningful regulatory modes and outperform several baseline methods for biomarker identification.

Second, we propose a novel method for transcriptional regulatory network identification by integrating gene expression data and protein-DNA binding data. The approach is built upon a multi-level analysis strategy designed for suppressing false positive predictions. With this strategy, a regulatory module becomes increasingly significant as more relevant gene sets are formed at finer levels. At each level, a two-stage support vector regression (SVR) method is utilized to reduce false positive predictions by integrating binding motif information and gene expression data; a significance analysis procedure is followed to assess the significance of each regulatory module. The resulting performance on simulation data and yeast cell cycle data shows that the multi-level SVR approach outperforms other existing methods in the identification of both regulators and their target genes. We have further applied the proposed method to breast cancer cell line data to identify condition-specific regulatory modules associated with estrogen treatment. Experimental results show that our method can identify biologically meaningful regulatory modules related to estrogen signaling and action in breast cancer.

Third, we propose a bootstrapping Markov Random Filed (MRF)-based method for subnetwork identification on microarray data by incorporating protein-protein interaction data. Methodologically, an MRF-based network score is first derived by considering the dependency among genes to increase the chance of selecting hub genes. A modified simulated annealing search algorithm is then utilized to find the optimal/suboptimal subnetworks with maximal network score. A bootstrapping scheme is finally implemented to generate confident subnetworks. Experimentally, we have compared the proposed method with other existing methods, and the resulting performance on simulation data shows that the bootstrapping MRF-based method outperforms other methods in identifying ground truth subnetwork and hub genes. We have then applied our method to breast cancer data to identify significant subnetworks associated with drug resistance. The identified subnetworks not only show good reproducibility across different data sets, but indicate several pathways and biological functions potentially associated with the development of breast cancer and drug resistance. In addition, we propose to develop network-constrained support vector machines (SVM) for cancer classification and prediction, by taking into account the network structure to construct classification hyperplanes. The simulation study demonstrates the effectiveness of our proposed method. The study on the real microarray data sets shows that our network-constrained SVM, together with the bootstrapping MRF-based subnetwork identification approach, can achieve better classification performance compared with conventional biomarker selection approaches and SVMs.

We believe that the research presented in this dissertation not only provides novel and effective methods to model and analyze different types of biological data, the extensive experiments on several real microarray data sets and results also show the

potential to improve the understanding of biological mechanisms related to cancers by generating novel hypotheses for further study.

Acknowledgments

I would like to take this opportunity to acknowledge all the people who give me their kind help and support in completing this dissertation.

First and foremost I want to thank my doctoral advisor, Dr. Jason J. Xuan, for his help, guidance, support and encouragement throughout my dissertation study. He provided me with professional and insightful guidance on my research topics; stimulated me to develop independent thinking and research skills; encouraged me staying positive when I encountered difficulties during the research; and helped me greatly on my dissertation and paper writing. Besides, Dr. Xuan also gave me many invaluable advices in various aspects in the past four years. Without his support, I could not have done what I am able to.

I would like to thank Dr. Yue J. Wang as my steering committee member. I appreciate his great insights in problem solving and remarkable knowledge in machine learning and mathematics, which helped me a lot in building up my dissertation. I am also grateful to him as the director of CBIL for providing me with great environment to conduct research in the past four years.

I would like to express my sincere gratitude to the other committee members, Dr. Chang-Tien Lu, Dr. Christopher L. Wyatt and Dr. Scott F. Midkiff for their useful suggestions, directions and help. They have provided me with valuable comments on my preliminary examination, suggested improvements in my presentation and shared their insightful feedback in the dissertation.

I greatly appreciate and enjoy collaborating with the laboratories at Georgetown University, Children's National Medical Center and Johns Hopkins Medical Institute. I am very grateful to Dr. Huai Li, Dr. Robert Clark, Dr. Rebecca B. Riggins, Dr. Eric P. Hoffman, Dr. Ie-Ming Shih and Dr. Tian-Li Wang as encouraging mentors to facilitate this inter-disciplinary research. They helped me identifying the biological problems, gave me usefully suggestions and provided me biological interpretations and support from the viewpoint of biologists. They also allowed the use of various pioneering high-throughput data generated from their laboratories before they were published.

I thank all of my colleagues and lab mates in CBIL. We have experienced so many wonderful moments together. I have benefited greatly through the daily discussion and group meetings with them. We shared and discussed many ideas related to machine learning and biological problems. Many of them also sat in my preparation talks and offered useful suggestions about the presentation of my work. They made this period of time a unique experience in my life.

I would thank to my extended family for all the supports from them despite living thousands of miles away. I am indebted to my parents, Yongming Chen, and Qizhen Zhuang, for their love and support in my life. Their dedication to my education shaped my life and values, and their selfless support allowed me to accomplish goals in life. My elder sister, Jie Chen, is my best friend who has always been supporting and encouraging everything in my life. My in-laws, Xiaozhi Chen and Guanzhen Geng, have provided strong support on the progress of my work and helped to take care of my baby.

Finally, I would like to dedicate this dissertation to my family. My husband, Wei Chen, gave me great support and encouragement during my PhD study. I could not have finished this dissertation without his support. My daughter, Jane Chen, although she is only eight month old, has brought lots of love, joy and hope in my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC	Area under the ROC Curve
BiNGO	Biological Network Gene Ontology tool
BMRF	Bootstrapping Markov Random Field
CDF	Cumulative Density Function
cDNA	complementary DeoxyriboNucleic Acid
ChIP	Chromatin ImmunoPrecipitation
COGRIM	Clustering of Genes into Regulons using Integrated Modeling
CRLB	Cramér-Rao Lower Bound
CRMs	*cis*-Regulatory Modules
css	core similarity score
DE	Differentially Expressed
DNA	DeoxyriboNucleic Acid
EE	Equally Expressed
ER+	Estrogen Receptor positive
FDR	False Discovery Rate
FIM	Fisher Information Matrix
FPR	False Positive Rate
GAs	Genetic Algorithms
GG	Gamma-Gamma
GRAM	Genetic Regulatory Modules
GO	Gene Ontology
HPRD	Human Protein Reference Database
ICA	Independent Component Analysis

IPA	Ingenuity Pathway Analysis
KGP	Knowledge Gene Pool
LS-regression	Least Square regression
MAP	Maximum A Posterior
MAS	Affymetrix MicroArray Suite
ml-SVR	multi-level Support Vector Regression
mNCA	motif-directed Network Component Analysis
MM	MisMatch
MRF	Markov Random Field
mss	matrix similarity score
mRNA	messenger RiboNucleic Acid
NCA	Network Component Analysis
PCA	Principal Component Analysis
PDF	Probability Density Function
Plier	Probe Logarithmic Intensity ERror
PM	Perfect Match
PPI	Protein-Protein Interaction
PWM	Position Weight Matrix
RMA	Robust Multichip Average
ROC	Receiver Operating Characteristics
RSF-1	Remodeling and Spacing Factor 1
SAM	Significance Analysis of Microarray
SNR	Signal to Noise Ratio
SOM	Self-Organizing Maps
SVM	Support Vector Machine

SVR	Support Vector Regression
TF	Transcription Factor
TFA	Transcription Factor Activity
TFBS	Transcription Factor Binding Site
TPR	True Positive Rate
TRM	Transcriptional Regulatory Module
TSS	Transcription Start Site

1 Background and Introduction

1.1. Motivation

Computational biology is an interdisplinary field that applies techniques from computer science, engineering, applied mathematics and statistics to address biological problems [1]. The main focus lies on developing mathematical models and applying computational algorithms and statistical techniques to interpret biological data and understand underlying biological mechanisms such as cancer related pathways. State-of-the-art high-throughput technologies, such as SNP arrays, DNA microarrays and mass spectrometry, are capable of producing terabytes of data that are not easily handled by conventional analysis approaches. As a result, the demand for computational, statistical or machine learning techniques is stronger than before to process, analyze and understand biological data in a comprehensive way.

Biological systems consists of different multi-functional elements that interact selectively and often non-linearly in order to have coherent and complex behaviors [2]. Multiple data sources can reveal different aspects of biological system. Traditional machine learning approaches cannot provide ‘global picture’ by focusing on one type of data source. Therefore integrating multiple data sources is becoming more and more important. The current and future needs will in particular require sophisticated integration of extremely diverse sets of data, aiming to better understand the main features of biological processes. On the other hand, understanding different types of biological data sources and their relationship will further help us develop mathematical models to tackle biological problems. In this study, we will focus on three different data sources for integration analysis: microarray gene expression data, protein-DNA interaction data and protein-protein interaction data, and develop different integrated methods for cancer research.

1.2. Biological background and data sources

Proteins are the structural components of cells and tissues that perform many key functions in biological systems. The production of proteins is controlled by genes, which are coded in deoxyribonucleic acid (DNA) [3]. A gene consists of a specific DNA fragment, and can be interpreted as a construction for a protein. Protein production from genes is explained by the central dogma of molecular biology (Figure 1.1) that includes two principal stages, transcription and translation. First the gene is transcribed into messenger ribonucleic acid, abbreviated as mRNA. Second, the mRNA is translated into a protein. There is huge variation in abundance and efficiency of transcription and translation among different cell types. The distribution is responsible for the appearance and state of a cell. Therefore, a cell's role is decided by the proteins it produces, which in turn depend on its expressed genes. Corresponding to the procedure from transcription to translation, multiple types of biological data can be measured at different levels as shown in Figure 1.1 and here we will focus on three types: microarray gene expression data, protein-DNA interaction data and protein-protein interaction data in our study.

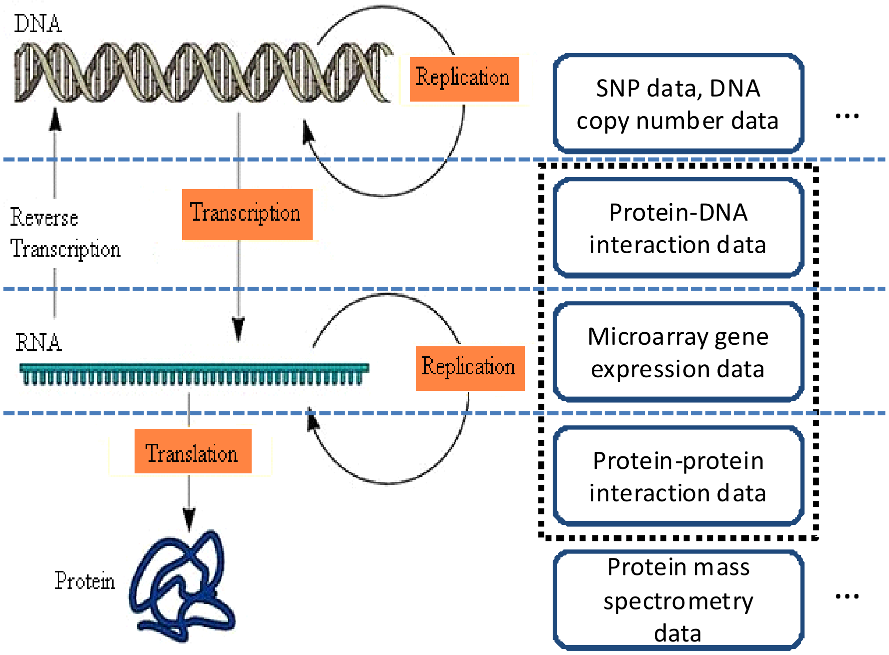

DNA
Replication
SNP data, DNA copy number data
...
Reverse Transcription
Transcription
Protein-DNA interaction data
RNA
Replication
Microarray gene expression data
Translation
Protein-protein interaction data
Protein
Protein mass spectrometry data
...

Figure 1.1. The central dogma of molecular biology from http://cats.med.uvm.edu/ and the corresponding different types of biological data.

1.2.1. Microarray gene expression data

Measuring changes in mRNA levels is one of possible methods to detect differences between cells. Scientists study different kinds and amounts of mRNA to learn which genes are expressed in a cell or changed among different types of cells. There are various methods for detecting and quantifying the amount of mRNA [4-6]. Traditional methods in molecular biology generally have the limits that the throughput is very limited and the "whole picture" of gene function is hard to obtain. A new technology, called DNA microarray, has been of interest among biologists. This technology can monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of gene.

DNA microarrays, or DNA chips are fabricated by high-speed robotics, for which probes with known identity are used to determine complementary binding, thus it allows massively parallel gene expression measurement and studies [3]. An experiment with a single DNA chip can generate thousands of gene expression levels simultaneously. The two most prevalent microarray technologies are cDNA microarrays and high-density oligonucleotide arrays. The differences between them are in the manner of placement of the DNA sequences on the array and in the length of these sequences during hybridization. Accordingly, the experimental approach and the data preprocessing differ as well.

cDNA microarray is a two-channel microarray where mRNA from two different biological samples (i.e. a sample of interest and a control sample) is reverse transcribed into cDNA, labeled with red and green fluorescent dyes, and distributed on the microarray. Then the cDNA competitively hybridizes to the corresponding DNA clones. Finally the remaining material is washed off and the amount of chemically bound cDNA is quantified by the intensity of the fluorescence in each spot measured by a laser scanner. This procedure was described in [6] and [7] and illustrated in Figure 1.2 [8]. Higher fluorescence indicates higher amounts of hybridized cDNA, which in turn indicates higher gene expression in the sample. The use of two samples in cDNA allows for measurement of relative gene expression across two sources of cDNA. Therefore it is less sensitive to the variable amount of spotted DNA, as well as other experimental variation in this way.

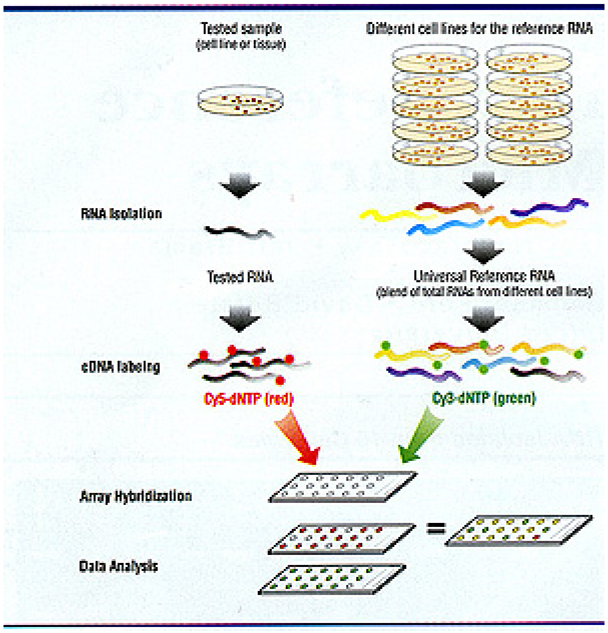


Figure 1.2. Two-color Microarray Experiment [8].

The oligonucleotide array, most widely used by the Affymetrix GenChip^{TM} , is a one-channel microarray. It is based on hybridization to small, high-density arrays containing tens of thousands of synthetic oligonucleotides [9]. Compared to cDNA, its main characteristics are 1) only one biological interest sample is fluorescently labeled and hybridized to the microarray. There is no competitive hybridization, and 2) the expression of each gene is measured by comparing hybridization of the sample mRNA to a set of probes, which is composed of 11-20 pairs of oligonucleotides and each of length 25 base pairs. The first type of probe in each pair is the perfect match (PM) which exactly corresponds to the gene sequence, whereas the second is the mismatch (MM), created by changing the middle (the 13th) base of the original sequence. The idea of this construction is to provide a control mechanism for random variation and cross-hybridization. For each gene, or probe set, the typical output consists of two vectors of intensities, one for PMs and one for MMs. To combine the PM and MM probe intensities for each probe set, many methods have been developed, such as MAS [10], Robust Multi-chip Average (RMA) [11] and Plier [12] with quantile normalization [13].

After the preprocessing and normalization, microarray gene expression data can be represented by a matrix where each row indicates one probe set and each column indicates one sample. The element in the matrix represents the gene expression level for a specific gene in a specific sample. Many studies have been investigated on microarray gene expression data, such as biomarker identification [14], gene/sample clustering [15], sample classification and prediction [16].

1.2.2. Protein-DNA binding data

Transcriptional regulation of gene expression is the process that regulatory protein binds to a DNA binding site located near the promoter to control when transcription occurs and how much RNA is created. The regulatory mechanism in eukaryotes is more complicated than in prokaryotes in that (1) transcriptional regulation tends to involve combinatorial interactions between several transcription factors, and (2) transcriptional regulation exhibit condition-specific characteristics so that it permits spatial (e.g. tissue-specific) and temporal (e.g. environment dependent) differences in gene expression. Therefore identifying condition-specific transcriptional regulation network between transcription factors and target genes becomes an important topic in computation biology.

To measure the protein-DNA binding data, ChIP-on-chip technology has been widely used. ChIP-on-chip, also known as genome-wide location analysis, is a technology for identifying genomic sites bounded by specific DNA binding proteins in living cells [17, 18]. It consists of two techniques which are Chromatin immunoprecipitation (ChIP) assay and DNA microarray technique. The principle underpinning ChIP assay is that DNA-bound proteins (including transcription factors) in living cells can be cross-linked to the chromatin on which they are situated. Figure 1.3 shows the major steps in ChIP-on-chip experiment, which include cross-linking, sonication, ChIP, reverse cross-linking and purify DNA and microarray [19].

Similar as gene expression microarrays, ChIP-on-chip experiments also have two color tiling array (e.g. NimbleGen) and single-channel tiling array (e.g. Affymetrix). The interpretation of data generated by a ChIP-on-chip experiment is similar to the interpretation of traditional gene expression microarrays in many respects, for instance, background correction and normalization. However, they differ in two important ways. First, the ChIP-on-chip signal may span several arrayed elements representing genomically adjacent DNA. Second, the measurements derived from ChIP-on-chip experiments arise as a mixture of two distributions: one is the signal distribution and another corresponds to the background and noise. Based on above characteristics, specific normalization methods [20, 21] and peak detection algorithms [19, 22, 23] have been proposed to analyze the ChIP-on-chip raw data for different type of arrays.

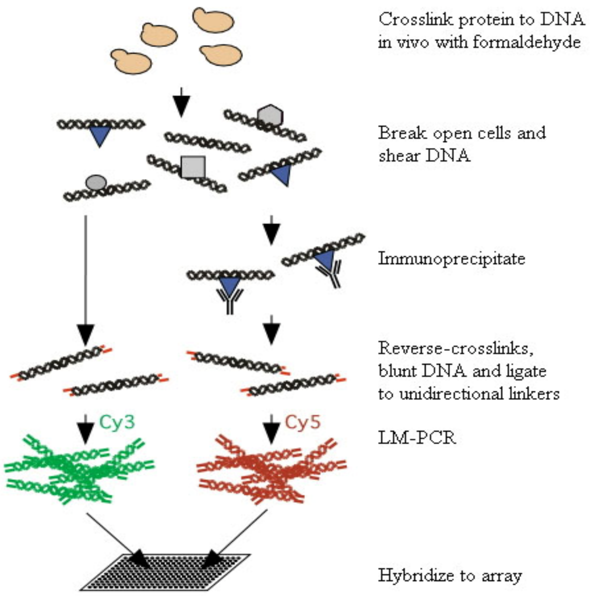


Figure 1.3. Procedure of ChIP-on-chip experiment [19].

Many studies and applications have been proposed based on ChIP-on-chip binding data, such as detecting target genes of transcription factors under specific conditions by integrating mRNA gene expression profiles [24], *de novo* motif discovery [25] and reconstruction of transcriptional regulatory network [26]. The major disadvantage is the requirement for highly specific antibodies for each protein to be tested, especially for higher eukaryotes studies, which usually results in limited scale of ChIP-on-chip binding data set. An alternative and practical way is to extract binding information from the promoter regions of focused genes based on sequence information.

The DNA sequence of a gene is composed of three functionally distinct regions: the regulatory (or promoter) region, RNA-coding and terminator. Determining and understanding the promoter structure is an important prerequisite to understanding gene regulation. The content of the regulatory region sequence determines which transcription factors will be recruited and bound to it. A transcription factor is a protein that binds to specific DNA sequences (binding sites) and thereby controls the transcription of genetic information from DNA to mRNA. The binding sites are short DNA sequences, comprising four to twenty nucleotides [27]. Most positions in the sequence are highly conserved (i.e., have low sequence variation) and are frequent in the regulatory regions of co-regulated genes bound by the transcription factor. For computational analysis a matrix representation of binding sites is normally used. The matrix defines the frequency of the four bases (Adenine, Thymine, Guanine, and Cytosine) at each position in a binding site, and usually it is called position weight matrix (PWM). The matrix can be obtained from an aligned set of (putative) binding sites, and it represents an average sequence of the entire set of binding sites. Given the PWM of a specific transcription factor and promoter sequence of focused genes, one can extract binding information using matching algorithm. There are many databases which provide all available transcription factors and their corresponding PWMs for user to search for specific binding information, such as TRANSFAC [28], JASPAR [29] etc. One of the major disadvantages of binding data extracted from sequence data is that the binding information is quite general and noisy. Therefore, sophisticated method is necessary to deal with the noise information.

1.2.3. Protein-protein interaction data

Proteins play an important role in many biological functions and they collaborate or interact with one another within a cell to perform some common purpose. The interactions between proteins have numerous different biological functions. For examples, signal transduction, protein complex and protein kinase. Therefore, protein-protein interactions (PPI) are of central importance for virtually every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches. There are many biochemical methods and physical methods to investigate the protein-protein interactions such as yeast-2-hybrid screens [30] or mass spectrometry techniques [31]. The interaction data are curated and can be downloaded from many databases [32, 33].

Protein-protein interaction network can be represented as a graph where the nodes represent the proteins and edges represent the interactions between proteins (See Figure 1.4). In most cases of protein-protein interactions, edges are undirected which show no cause/effect relationship between two proteins. In some particular networks, such as metabolic network or pathways, the interactions can be represented by directed edges which indicate cause/effect relationship between proteins. Due to the noises and inaccuracy in measuring the interaction data in different techniques, more reliable networks is constructed by combining different sources of data and applying multiple criteria. Real values between 0 and 1 could be assigned to the interactions in the PPI network that show the degree of confidence of the interactions.

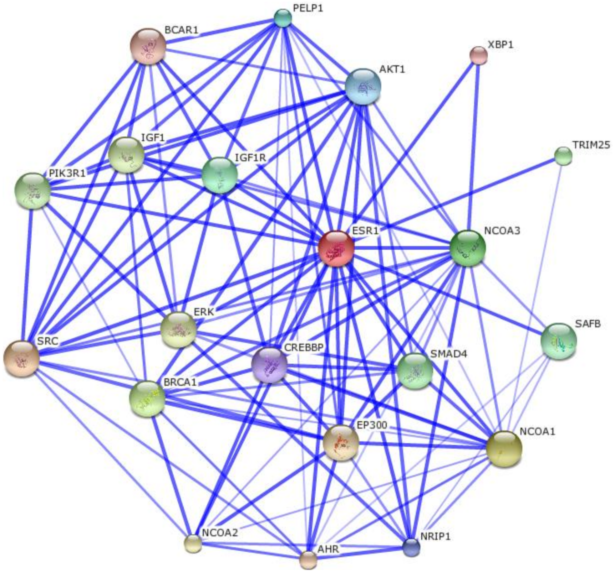


Figure 1.4. Confidence view of a PPI network, which is extracted from STRING database [33]. The network contains 21 proteins and 107 interactions. Stronger associations are represented by thicker lines.

Since PPI network can be represented by a graph, many graph theoretical techniques has been widely used to investigate the structural properties of networks. For examples, subgraph and centrality statistics analysis [34], identification of network motif [35], clustering on the PPI network for functional modules [36], prediction of functional category for unknown proteins [37] etc.

1.3. Objectives and statement of problem

In this dissertation, we develop several methods to help understand biological problems by integrating multiple data sources, including microarray gene expression data, protein-DNA interaction data (specifically for binding data) and protein-protein interaction data. We focus on data integration and mathematical modeling on multiple data sources and their applications to biological study.

1.3.1. Knowledge-guided biomarker identification on time course microarray data

Biomarker has a wide definition in biology and in genetics a biomarker is defined as a gene which causes disease or is associated with susceptibility to disease. Biomarker plays an important role for clinical diagnosis, drug discovery in disease study. As microarray technology makes it possible to measure the expression levels of thousands of genes simultaneously, biomarker identification has become one of the major goals of microarray data analysis. Microarray gene expression data can be roughly classified as static data or dynamic data. Static data measure gene expression levels of samples from different conditions, while dynamic data, also known as time course microarray data, represent gene expression levels of samples from successive times in a dynamic biological process. For static microarray data, the purpose of biomarker identification is to identify significant differentially expression genes between different conditions (e. g. cancer vs. normal). For dynamic microarray data, the purpose of biomarker identification becomes to find genes which are more informative to understand the underlying patterns of whole data or associate to a specific disease. Here we will focus on biomarker identification on dynamic time course microarray data.

In microarray data analysis, biomarker identification is challenging due to several reasons. First, many statistical methods fail at parameter estimation based on the insufficient data due to the ‘curse of dimensionality’. Typically a microarray data set contains tens to hundreds (n) of observations/samples but has tens of thousands (p) of genes as the variables/features. Therefore the convergence of any statistic estimator to the true value of a function defined on a high dimensional feature space is very slow when $p >> n$. Second, only a small portion of genes in microarray data are related to specific biological process of interest. With little knowledge about underlying biological mechanism, it is hard to extract biological related biomarkers. Finally, the significant noise in the microarray data requires developing robust and stable methods to deal with the noise impact.

Ideally, biomarkers should not only exhibit differential gene expressions between normal and disease samples, but more importantly, they should also reflect their biological role in the disease phenotype. Most existing methods applied to time-course microarray data have the limitation that extracted genes have statistical significance but little biologically relevant. Therefore, incorporation of prior knowledge is of great importance to improving the accuracy of computational methods identifying biologically relevant biomarkers for a specific disease. This leads to our **first research topic** to be addressed in the dissertation. We propose a novel method, namely knowledge-guided multi-scale ICA, to identify disease-specific biomarkers beyond partial prior knowledge.

1.3.2. Transcriptional regulatory network identification

The identification of gene regulatory modules is an important yet challenging problem in computational biology. Uncovering transcriptional regulatory networks helps us understand the complex cellular process. At the transcriptional level, a regulatory module is defined as a set of genes controlled by one or several transcription factors (TFs) in a condition-specific manner [38]. Therefore identification of regulatory network includes determination of active transcription factors and their target genes in given gene expression profiles under certain condition. At the same time, transcription factor activities inferred from regulatory network also reveal hidden activities of biological processes which could not be observed directly from gene expression levels.

Different mathematical models for transcriptional regulatory model have been developed, such as Boolean network [39], Bayesian network [38] and stochastic [40], to capture biological system behavior based on time course microarray data alone. On the other hand, since protein-DNA binding data provide more biologically and physically meaningful information, many methods have been proposed by integrating mRNA gene expression profiles and protein-DNA binding data to reconstruct condition-specific regulatory network [26].

While many computational methods have been proposed to identify regulatory modules, their initial success is largely compromised by a high rate of false positives in predicting gene module members, especially when applied to human cancer studies, due to high level of noises in the microarray gene expression data and protein-DNA binding data. New strategies are needed for reliable regulatory module identification, aiming to reduce the false positive rate by combating the noises in binding motif information and gene expression data. Our **second research topic** to be addressed in the dissertation is for reliable identification of regulatory modules, which will focus on several questions in the following:

- What are the active transcription factors in given mRNA gene expression profiles?
- What are the target genes for the active transcription factors?
- How significant and reliable are the active regulatory modules?

1.3.3. Subnetwork identification and network-based prediction

Traditional disease biomarkers are considered as individual genes that exhibit significantly differential expression between different phenotypes of samples. However, individual genes often result in poor generalizability with respect to prediction performance of cancer outcome, because of the deficiencies in experimental design, insufficient statistical power due to small sample size and even heterogeneity among patient samples. As an alternative, subnetwork identification has been proposed based on gene expression profiles and protein-protein interactions. Several methods [41, 42] have been developed and achieved promising results in that more reproducible and robust markers can be identified and also the resulting subnetworks could reveal the underlying molecular mechanisms involved in disease [43].

One of the limitations of existing methods is that genes in the PPI network were treated independently when the network score was designed and the network structure was not utilized in the analysis. It is believed that genes in a local subnetwork have similar functional annotation, therefore they should have similar differential expression pattern in order to form a significant subnetwork. Another limitation is that hub genes, which are more biologically relevant and have many interactions in PPI network, usually have little changes in expression compared with their downstream genes [44-46]. The resulting subnetwork may have less ability to reveal the underlying mechanism of disease by picking up downstream rather than hub genes. Therefore, it is urgent to develop mathematical models to characterize the protein network structure and consider the neighboring effect of a hub gene when designing the network score function. This forms **the first part of our third research topic** in the dissertation.

The identified biomarkers usually are evaluated by their prediction power on the unknown samples, given clinical outcomes. Traditional classification methods build the classifier only based on individual features/genes, ignoring the relationship/interactions among the genes. Recently many methods have been developed to identify significant gene sets or pathways involved in diseases or biological processes by incorporating some prior knowledge, with which to help understand the underlying biological mechanism. For example, gene set enrichment analysis and pathway enrichment analysis approaches [47-49] are proposed by using membership information in functional gene clusters or pathways.

Besides the membership information in prior knowledge, several algorithms were developed based on the interacting structure, since most prior knowledge could be represented by graph, such as protein-protein interactions, protein-gene interactions or regulatory pathways. However, the subnetworks were represented by network activities and they are still treated as individual features for classification in some methods [41], and their structure information is not explicitly shown in the decision function of classifiers. **The second part of our third research topic** is to develop new classifiers that consider the interactions among input features and select significant features based on their expression levels and characteristics in networks.

1.4. Summary of contributions

In the context of research topics discussed above, we summarize the main contributions of this dissertation in this section.

(1) We develop a knowledge-guided multi-scale independent component analysis method for biomarker identification on time course microarray data. Given limited knowledge genes that are related to a specific disease or biological process, we show an initial effort to integrate them with microarray data to discover more biologically meaningful markers, which is different from traditional statistical methods.

(2) We propose a multi-level support vector regression method for transcriptional regulatory network identification by integrating gene expression data and protein-

DNA binding information. Due to the high noise levels in binding information and gene expression data, we focus on how to reduce the high false positive rate that exists in most existing methods for the identification of regulatory modules, through a multi-level framework and statistical significance analysis. We demonstrate the effectiveness of the proposed method on simulation data and yeast cell cycle data. We also apply the method to breast cancer cell line data and discovered biologically meaningful results highly related to the development of breast cancer.

(3) We develop a bootstrapping Markov random field (MRF)-based method for subnetwork identification to indicate significantly differential subnetworks between two phenotypes, based on microarray gene expression data and protein-protein interaction network. The proposed method addresses the problem that biologically meaningful genes with little gene expression changes across different phenotypes are hard to detect using traditional statistical methods. We demonstrate the resulting subnetworks can include more important genes related to significant biological processes or pathways through the simulation study. We also apply the proposed method to several breast cancer patient data sets for survival analysis, and the experimental results demonstrate that the method can reveal many important subnetworks associated with cancer and/or drug resistance.

(4) We propose to develop network-constrained support vector machines for cancer prediction using microarray data. In particular, we formulate the relationship of gene-gene interactions in PPI networks as a Laplacian term, which is integrated into the support vector machine framework for network-constrained prediction. Using simulation studies, we show that the reproducibility of the proposed method is better than conventional prediction methods in terms of the prediction performance of independent test and the recovery of ground truth subnetwork. The prediction results on real breast cancer microarray data sets show that our method outperforms conventional methods and provides more biologically meaningful features for improved cancer prediction.

1.5. List of relevant publications

Peer-reviewed Journal Publications

[1] **L. Chen**, J. Xuan, C. Wang, I. M. Shih, Y. Wang, Z. Zhang, E. Hoffman, and R. Clarke, "Knowledge-guided multi-scale independent component analysis for biomarker identification," *BMC Bioinformatics,* vol. 9, p. 416, 2008.

[2] R. B. Riggins, J. P. Lan, Y. Zhu, U. Klimach, A. Zwart, L. R. Cavalli, B. R. Haddad, **L. Chen**, T. Gong, J. Xuan, S. P. Ethier, and R. Clarke, "ERRgamma mediates tamoxifen resistance in novel models of invasive lobular breast cancer," *Cancer Res,* vol. 68, pp. 8908-17, Nov 1 2008.

[3] C. Wang, J. Xuan, **L. Chen**, P. Zhao, Y. Wang, R. Clarke, and E. Hoffman, "Motif-directed network component analysis for regulatory network inference," *BMC Bioinformatics,* vol. 9 Suppl 1, p. S21, 2008.

[4] **L. Chen**, J. Xuan, R. B. Riggins, Y. Wang, E. Hoffman, and R. Clarke, "Identification of Condition-specific Regulatory Modules by Multi-level Motif and mRNA Expression Analysis," *Intl J. of Computational Biology and Drug Design,* vol. 2, pp. 1-20, 2009.

[5] **L. Chen**, J. Xuan, C. Wang, I. M. Shih, Y. Wang, Z. Zhang, and R. Clarke, "Biomarker identification by knowledge-driven multi-level ICA and motif analysis," *Intl J. Data Mining and Bioinformatics,* vol. 3, pp. 365-381, 2009.

[6] I. M. Shih, **L. Chen**, C. Wang, J. Gu, B. Davidson, L. Cope, R. J. Kurman, J. Xuan and T. L. Wang, "Distinct DNA methylation profiles in ovarian serous neoplasms and their implications in ovarian carcinogenesis." *Am J Obstet Gynecol*. 2010.

[7] **L. Chen**, J. Xuan, R. B. Riggins, Y. Wang, E. Hoffman, and R. Clarke, "Multi-level Support Vector Regression Analysis to Identify Conditional-Specific Regulatory Networks," *Bioinformatics,* 26(11):1416-22, 2010.

Manuscripts To Be Submitted

[8] **L. Chen**, J. Xuan, R. B. Riggins, Y. Wang, and R. Clarke, “Bootstrapping MRF-based subnetwork identification from microarray data and protein-protein interaction network”, to be submitted to *Bioinformatics*, 2010.

[9] **L. Chen**, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, “Identifying prognostic biomarkers by network-constrained support vector machines”, to be submitted to *BMC Genomics*, 2010.

Conference Publications

[10] **L. Chen**, J. Xuan, R. Clarke, and Y. Wang, "Biomarker Identification by Knowledge-Driven Multi-Scale Independent Component Analysis," in *Proc. Third IEEE-NIH Life Science Systems and Applications Workshop* Bethesda, MD, 2007. (Best student’s paper).

[11] **L. Chen**, C. Wang, I. M. Shih, T. L., Wang, Z. Zhang, Y. Wang, R. Clarke, E. Hoffman, and J. Xuan, "Biomarker Identification by Knowledge-Driven Multi-Level ICA and Motif Analysis," in *Proc. Intl Workshop on Machine Learning Methods in Biomedicine and Bioinformatics* Cincinnati, OH, 2007.

[12] C. Wang, J. Xuan, **L. Chen**, P. Zhao, Y. Wang, R. Clarke, and E. Hoffman, "Integrative Network Component Analysis for Regulatory Network Reconstruction," in *Proc. Fourth Intl Symposium on Bioinformatics Research and Applications* Atlanta, GA, 2008.

[13] C. Wang, J. Xuan, **L. Chen**, R. B. Riggins, E. Hoffman, and R. Clarke, "Reliability Analysis of Transcriptional Regulatory Networks," in *Proc. Intl Conf. on Bioinformatics, Computational Biology, Genomics and Chemoinformatics* Orlando, FL, 2008.

[14] T. Gong, J. Xuan, **L. Chen**, R. B. Riggins, Y. Wang, E. Hoffman, and R. Clarke, "Sparse Decomposition of Gene Expression Data to Infer Transcriptional Modules Guided by Motif Information," in *Proc. Fourth Intl Symposium on Bioinformatics Research and Applications* Atlanta, GA, 2008.

[15] **L. Chen**, J. Xuan, R. B. Riggins, Y. Wang, E. P. Hoffman, and R. Clarke, "Identification of Condition-specific Regulatory Modules by Multi-level Motif and mRNA Expression Analysis," in *The 2008 Intl Conference on Bioinformatics & Computational Biology* Las Vegas, Nevada, 2008.

[16] **L. Chen**, J. Xuan, Y. Wang, R. B. Riggins, and R. Clarke, "Network-constrained SVM for Classification," in *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications* San Diego, CA: IEEE Computer Society, 2008, pp. 60-65.

1.6. Organization of the dissertation

The main research topics of this dissertation are data integration and modeling of different biological data sources. Specifically, we conduct our study at different levels of prior biological knowledge, i.e., some known biologically relevant genes, the binding information from transcription factors to target genes and protein-protein interactions. We then model the computational problems based on the unique characteristics of different data sources. Among them, the following research topics - biomarker identification based on time course microarray data, transcriptional regulatory network identification, subnetwork identification and network-based prediction - will be studied in detail.

Chapter 2 addresses the first research topic: knowledge-guided biomarker identification on time course microarray data. We first review the existing methods on biomarker selection. Then we discuss how to use knowledge genes to identify biologically relevant component in independent component analysis model. Base on that, we propose a multi-scale independent component analysis strategy through clustering to improve the accuracy of biomarker selection. Motif enrichment analysis is designed to evaluate the biological relevance of identified biomarkers when ground-truth is not available. With the criteria designed, we compare the performance of the proposed method with those of comparable methods on two time course microarray data sets and show the improvement on biomarker identification with the proposed method.

Chapter 3 addresses the second research topic: condition-specific regulatory network identification by integrating gene expression data and binding information. Existing regulatory network identification methods with/without integrating binding information are first reviewed. A linear model for transcriptional regulatory network is then discussed and a two-step support vector regression algorithm is proposed to solve the linear model with noisy binding information in order to obtain reliable results. Finally, a multi-level regression strategy based on a clustering framework is developed to identify stable and consistent regulatory modules, followed by a statistical analysis to ensure the significance of identified modules. The proposed method is applied to both synthetic and real microarray data and compared with some existing benchmark methods. The performance in identifying transcription factors and their target genes is compared and the results demonstrate the effectiveness of the proposed method.

Chapter 4 addresses the third research topic: subnetwork identification and network-based prediction using microarray data and protein-protein interaction network. For subnetwork identification from microarray data, we first review the existing methods and discuss the limitation of these methods. Then we present a bootstrapping Markov random field (MRF)-based subnetwork identification method by integrating microarray gene expression data and protein-protein interaction network. A network score is derived based on the framework of MRF-maximum a posterior (MAP). Then simulated annealing search algorithm is implemented to ensure finding the optimal/suboptimal solution. Finally bootstrapping scheme is carried out for confidence assessment of selected nodes in the subnetworks. We evaluate the proposed method on simulated data and the comparison studies demonstrate that the method outperforms some benchmark methods for subnetwork identification. Finally we apply the proposed method to real microarray data and the results show that the proposed method can identify important subnetworks related to breast cancer and drug resistance. For network-based prediction, we propose a network-constrained support vector machines (SVMs) method for classification and prediction, by integrating microarray gene expression data and PPI networks. The structure of a network is represented by a Laplacian matrix of graph and incorporated into the support vector machine framework. The mathematical model is derived and then the solution of network-constrained SVMs is presented. Statistical significance analysis is designed for significant subnetwork identification. We conduct simulation experiments for the proposed method, and compare with conventional SVM. The results show that the proposed method improves feature selection capability and achieves better reproducibility on independent tests. The prediction performance on real microarray data sets outperforms conventional method and reveals more biologically meaningful features.

Chapter 5 summarizes the original contribution of this dissertation research work, discusses several problems and tasks for the future work and draws the conclusion of the dissertation.

2 Knowledge-guided Multi-scale Independent Component Analysis for Biomarker Identification

2.1. Introduction

Under their broadest definition, biomarkers include any biological or chemical indicator of a specific underlying process. In genetics, biomarkers are defined as a set of genes that are associated with a disease or are associated with the susceptibility to develop a specific disease. Microarray technology makes it possible to measure simultaneously the expression levels of thousands of genes, and identifying meaningful and useful biomarkers from these large data sets is a common goal. Specifically, investigators attempt to detect genes differentially expressed across different types of tissue samples or the samples obtained under different experimental conditions. Traditional biomarker identification methods have mainly been applied to statistical analysis of microarray data alone; T-test [50] and significance analysis of microarray (SAM) [14] are frequently used to detect differentially expressed genes between two phenotypes. Several new statistical methods have been developed to analyze time-course microarray data. Storey *et al.* proposed an algorithm (EDGE) to fit the time-course microarray data with natural cubic splines, followed by a goodness-of-fit test to detect differentially expressed genes [51]. Conesa *et al.* also proposed a two-step regression approach to sequentially identify differentially expressed genes from time-course microarray data under different conditions [52]. However, these and many related approaches do not incorporate knowledge of gene function, with respect to the phenotypes of interest, into their statistical models.

Ideally, biomarkers should not only exhibit differential gene expressions between normal and disease samples, but more importantly, they should also reflect their biological role in the disease phenotype. Most significance analysis methods applied to population (static) or time-course microarray data have the limitation that genes are analyzed independently and the interactions among them are ignored. Clustering methods, such as k-means clustering [53] and self-organizing maps (SOMs) [54], were introduced to group the genes with similar expression patterns. A shortcoming of the clustering methods is that they do not allow genes to be shared by multiple clusters. However, a single gene can be involved in multiple distinct biological processes [55]. One solution to this problem is to first infer gene regulatory networks [38, 56-59] that appear to control or regulate phenotypically relevant biological functions, and then to extract the most biologically and statistically relevant biomarkers.

The application of Independent Component Analysis (ICA) to microarray data has shown some utility in regulatory network inference [57, 60]. ICA is a statistically-principled linear decomposition method that models the observations as a linear combination of some latent (or hidden) variables [61]. From the perspective of a gene regulatory mechanism, any gene expression value can be regarded as a combinational effect of some regulatory inputs such as transcription factors, cellular functions, or responses to experiment conditions [57, 59]. As demonstrated in previous work [62] along with that of others [57, 59], novel applications of ICA to high-throughput data from microarray technology can help reveal dominant regulatory mechanisms.

It is not a trivial task to link the estimated latent variables from ICA to real biological functions. To identify biologically relevant biomarkers for a specific disease, the incorporation of prior knowledge is of great importance to improving the accuracy of computational methods [63]. However, complete prior knowledge is often difficult to obtain. Some prior knowledge, such as regulatory motif information (promoter responsive element sequence) is available and can be incorporated into microarray data analysis to assist in regulatory module identification [64, 65]. Recently, a new approach called motif-directed network component analysis (mNCA) is developed to infer transcription regulatory activities (TFAs). This approach incorporates a stability analysis procedure to overcome the problem of many false positives in motif information [66]. Since we can only use known motifs, a clear limitation of the mNCA method is that we cannot infer any new potential regulatory biomarkers beyond prior knowledge from the model.

We propose a novel method, namely knowledge-guided multi-scale ICA, to identify disease-specific biomarkers beyond partial prior knowledge. We assume that a latent variable estimated by ICA from the entire gene expression population represents the joint effect of several biological functions. Disease-specific biomarkers could be involved in several different biological functions by the ICA latent variables or linear regulatory modes. Therefore, we first cluster the whole gene population into multiple sub-populations in which only a few biological processes are involved. We then uncover the knowledge-relevant regulatory modes in each subpopulation based on the partial prior knowledge. Finally, disease-specific biomarkers are extracted according to the strength of their association with the extracted regulatory modes. A statistical test is applied to evaluate the significant enrichment of transcription factors for the extracted biomarkers based on motif information.

For algorithm validation, we applied our approach to two time-course microarray data sets to demonstrate its improved performance. The first data set is a yeast cell cycle microarray data set with 104 well known cell cycle-related genes; the second is a remodeling and spacing factor 1 (Rsf-1) induced microarray data set from a profiling study of ovarian cancer. The experimental results show that our approach can identify biologically meaningful disease-specific biomarkers related to ovarian cancer, as compared to other gene selection methods with or without prior knowledge.

Organization of this chapter is as follows. In Section 2.2, we discuss the whole procedure of the proposed method which includes knowledge-guided multi-scale independent component for biomarker identification and motif enrichment analysis for biological evaluation. In Section 2.3, we introduce several baseline methods for biomarker identification on time course microarray data set in order to compare with the proposed method, as well as the evaluation criteria in the condition that ground truth biomarkers are known and unknown. In Section 2.4, we show the experiments and results on two microarray data sets, followed by discussion and conclusion in Section 2.5 and 2.6, respectively.

2.2. Methods

If we apply ICA directly onto an entire gene expression population, the extracted regulatory modes will reflect the joint effect of several biological functions, some of which are related to the disease under study and some are not. To overcome this problem, we developed a divide-and-conquer strategy. We applied a knowledge-guided multi-scale ICA approach to extract disease-related regulatory modes reliably, and then we identify the biomarkers associated with the modes. The overall scheme is illustrated in Figure 2.1. Firstly, a knowledge gene pool (KGP) is constructed by collecting the genes that are known to be relevant to the specific disease from available databases and literatures. Secondly, the entire gene population is divided into sub-populations by a clustering method applied to the microarray data and, to identify regulatory modes, ICA is applied to each sub-population. The most relevant linear regulatory mode in each cluster is extracted using the gene metadata in the KGP and the associated biomarkers are ranked according to their weighted loading factors. Finally, motif enrichment analysis is conducted to evaluate the extracted biomarker candidates in terms of the enrichment of disease-related transcription factors.

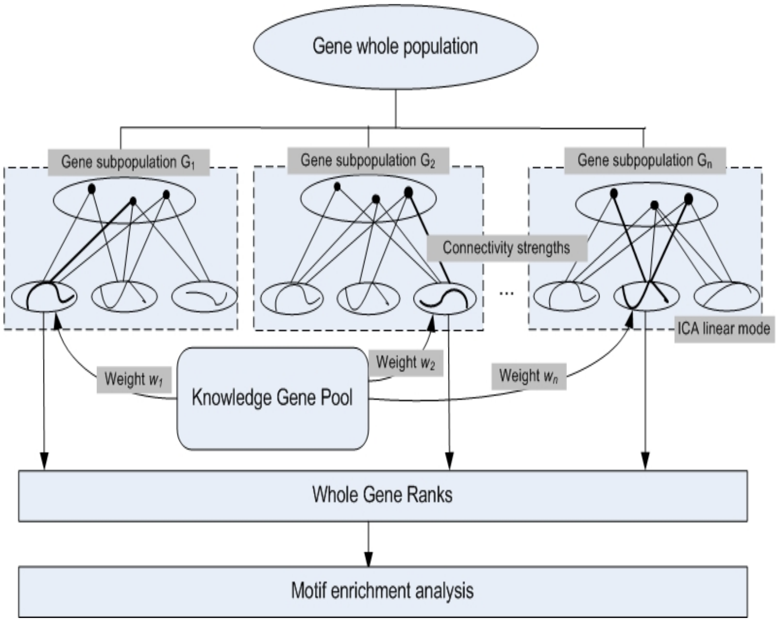


Figure 2.1. Flow chart of the proposed method - knowledge-guided multi-scale independent component analysis (ICA) - for biomarker identification.

2.2.1. Independent component analysis (ICA)

Consider a gene expression data matrix **X** = [x_{ji}], whose rows correspond to different microarray samples, and columns correspond to individual genes. ICA decomposition model can be mathematically formulated as (assuming noiselessness for simplicity):

$$\mathbf{X}_{N\times L} = A_{N\times M}\mathbf{S}_{M\times L}, \tag{2.1}$$

$$\mathbf{U}_{M\times L} = W_{M\times N}\mathbf{X}_{N\times L}, \tag{2.2}$$

where Equation (2.1) describes the linear combination model with mixing matrix ***A***, and Equation (2.2) the decomposition model with de-mixing matrix ***W***. ***S***, **X** and **U** are independent components, mixtures, and estimated independent components, respectively. *M* is the number of independent components, *N* the number of samples and *L* the number of genes.

In microarray data analysis, an ICA model could be interpreted as the expression value of an individual gene *i* under condition *j* ($\mathbf{x}_i(j)$) is the summation of different linear modes in ***A*** at condition *j* ($\mathbf{a}_k(j)$) weighted by independent loading factors s_{ik} in ***S*** [8], as shown below:

$$\mathbf{x}_i(j) = \sum_{k=1}^{M} s_{ik}\mathbf{a}_k(j), \quad i = 1,\ldots,L; j = 1,\ldots,N\,. \tag{2.3}$$

The linear modes in ***A*** might reflect distinct regulatory mechanisms involved in gene regulation, such as transcription factor (TF) activities. The FastICA algorithm [67] can be utilized to obtain ***A*** and ***S*** based on the assumption that the components are statistically independent and have non-normal distributions (typically super-Gaussian). This assumption is biologically plausible as most genes are not expected to change dramatically. Only the genes involved in distinct regulatory mechanisms will change, producing super-Gaussian distributions in microarray data.

Several methods have been developed to associate a set of genes with a specific linear mode [57, 59, 68]. These methods each assume that genes with the highest absolute loading values are the significant genes associated with linear mode $\mathbf{a}_k$. Here, genes are ranked by a modified criterion based on the same assumption as described in the next section.

2.2.2. Knowledge-guided multi-scale ICA

Since ICA is an unsupervised method, it is difficult to determine which linear modes are related to specific biological functions. To identify the biomarkers relevant to a specific biological function, prior knowledge could provide guidance for any computational method. In this approach, we will collect a KGP containing genes strongly associated with the disease and use these to guide the ICA approach for disease-relevant biomarker identification. Notice that the total connection strength of the knowledge genes associated with a disease-relevant linear mode would be larger, in principle, than that of irrelevant linear modes. Based on this observation, the most knowledge-relevant linear mode can be determined from the estimated ICA modes and the associated genes can then be extracted.

However, if we apply ICA to the entire molecular profile, the estimated linear modes will likely reflect the joint effect of several biological functions, even for the most knowledge-relevant mode, because many disease-irrelevant but differentially expressed genes co-exist in the data. Conversely, biomarkers should be involved in several different linear modes in relation to underlying biological processes. Therefore, it is reasonable to first separate the entire profile into sub-populations. We can then find the specific ICA linear modes from different subsets of genes rather than from the whole gene population; this approach is referred to as the "multi-scale ICA" approach here. Since these modes will be associated with different parts of the knowledge genes in the KGP, they are more suitable for biomarker identification. Clustering methods, such as k-means clustering and SOMs, can be used to form the subsets of genes, with the assumption that the genes involved in similar biological functions are more likely to exhibit similar expression patterns than genes involved in different biological functions.

Our method can be mathematically described as follows. Assume a whole gene population $\boldsymbol{G}$ in a microarray data $\mathbf{X}$ has been clustered into n subsets, $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, …, $\boldsymbol{G}_n$. For each subset $\boldsymbol{G}_i$ (i=1,...,n), we apply ICA to find the most knowledge-relevant linear mode $\mathbf{a}_j$ according to the total connection strength of the knowledge genes in this subset. Thus, the index j can be obtained by

$$j = \arg\max_{m}\left(\sum_{g \in K_i} \left|s_{gm}\right|\right) \qquad m = 1,\ldots,M_i \ , \tag{2.4}$$

where s_{gm} is the loading factor for gene g associated with linear mode $\mathbf{a}_m$, K_i the subset of knowledge genes in the i^{th} cluster, and M_i the number of independent components in the i^{th} cluster.

Then each gene g in this subset is assigned a score c, which is defined as follows:

$$c_g = w_i * \left|s_{gj}\right|, \quad g \in G_i, \quad w_i = \frac{|K_i|}{|K|} \quad , \tag{2.5}$$

where w_i is a weight to represent the significance of the linear mode in the i^{th} subset associated with the prior knowledge. Here we define w_i as the proportion of all knowledge genes in this subset with respect to the entire KGP (K). Once the knowledge-relevant linear modes in all subsets are determined, each gene will have a score assigned and we rank the genes in terms of their scores. The larger the score, the more strongly the gene is related to the biological process.

A key issue in this method is how to determine the optimal cluster number when forming the subsets of genes. Here, we determine the optimal cluster number by a cross-validation approach. Specifically, we assume the optimal cluster number is in some range, from 1 to an upper limit. For each cluster number, the knowledge genes are randomly stratified into a training gene set (as our partial prior knowledge gene set) and a test gene set by a ten-fold cross-validation approach. The method is applied with the partial prior knowledge genes to rank the whole gene population, and prediction accuracy is tested on the test gene set. The above procedure is repeated 10 times, once for each left out fold, and an average accuracy over the ten folds is reported. We select the number with the highest average accuracy as the optimal cluster number for clustering. The upper limit of cluster numbers should be cautiously determined by the number of knowledge genes and the number of genes in the full profile. If the number of clusters is too large, it will lose the ability to infer novel biomarkers. An extreme case is that each individual gene forms a cluster and then we can only obtain the correct ranks for known genes. Genes not in the KGP will be randomly ranked, which is not informative at all for biomarker identification. If the cluster number is too small, the estimated linear modes may be incorrect due to the presence of many irrelevant genes. In our experiments, we set the upper limit as 10 for the yeast cell cycle data set and 15 for the ovarian cancer microarray data set, respectively.

2.2.3. Knowledge gene pool (KGP)

Each KGP is a collection of those genes that are potentially most strongly related to a specific disease. Usually there are thousands of genes in microarray data and most of them are not relevant to a specific disease even though they exhibit changes in gene expression level. The knowledge gene pool is an important asset for data analysis since it helps reduce many false positives. However, in most cases, little prior knowledge can be obtained, and the available knowledge is usually neither complete nor sufficiently accurate to fully define the specific disease under study. Thus, the KGP is best used as a guide for biomarker identification. In our studies, the KGP is primarily constructed from the published biological literature or from databases such as Ingenuity Pathway Analysis (IPA; Ingenuity Systems: http://www.ingenuity.com) and the TRANSFAC 11.1 Professional Database [28].

2.2.4. Evaluation by motif enrichment analysis

For microarray data analysis, there is often no ground truth (i.e., true biomarkers known to be related to a specific biological process or disease under study) available for us to evaluate the performance of a biomarker identification method. However, we know that gene expression is often regulated by transcription factors (TFs), proteins that bind to promoter or enhancer sequence elements upstream of genes and either activate or inhibit gene expression. Here, with the motif information provided, we have designed a statistical test to evaluate the enrichment of transcription factors for a gene set identified. A gene-transcription factor matrix $\boldsymbol{M}$ is generated where each element in the matrix, m_{gf}, represents how well the upstream sequence of a gene g matches the motif that a transcription factor f binds to. For human genes, 2Kbp upstream regions from the transcription start sites (TSSs) of the genes are extracted from the UCSC genome databases [69]. Match™ [70] is then used to search the transcription factor binding site (TFBS) by its position-weighted matrices (PWMs) in a gene's upstream region, which outputs the scores of core similarity and matrix similarity for each matched motif. Since one TF may have multiple TFBSs, we use the summation of average scores of core similarity and matrix similarity to set the final value of m_{gf}.

Given a gene set S extracted by a computational method, a statistic to measure the enrichment of a specific transcription factor f is defined as

$$e_f = \sum_{g \in S} m_{gf} . \tag{2.6}$$

To calculate the statistical significance (p-value), we need to form a null distribution. The null hypothesis is that the gene set is randomly generated from the gene population and there is no significant enrichment of the transcription factor f. We randomly select gene sets with same size of S from the baseline gene population, and repeat B times to generate the corresponding null statistic enrichment score e_f^{0b}, for b = 1,…,B. The null hypothesis distribution is assumed to be symmetric in this study. The p-value can be obtained for each gene set by calculating the probability that a null gene set has a statistic more extreme than the observed statistic. Mathematically, the p-value can be calculated by:

$$p_S = \frac{\text{number of members in } \{b : e_f^{0b} > e_f , b = 1,\ldots, B\}}{B} . \tag{2.7}$$

2.3. Baseline experiments and evaluation method

To evaluate the performance of our proposed approach, EDGE algorithm [51] was first considered as a comparison method since it was specially designed to identify statistically significant genes from time-course microarray data. However, this comparison is insufficient due to that EDGE does not incorporate knowledge genes to provide guidance for biomarker identification. On the other hand, given partial prior knowledge genes, traditional supervised classification methods are not suitable to predict whether a gene is related to prior knowledge because there is no true negative gene available. Therefore, we design three baseline biomarker identification methods that incorporate partial prior knowledge for a fair comparison. The first baseline ICA method is designed to evaluate if our multi-scale strategy by clustering offers an improved performance for biomarker identification. Two correlation methods with or without clustering are then implemented to identify the genes exhibiting similar patterns with partial prior genes, compared to the ICA approach focusing on regulatory mode identification. Specifically, the first method is a baseline ICA method where ICA is applied to the entire expression profile and the partial prior knowledge is used to find the most knowledge-relevant linear mode by Equation (2.4). Genes are ranked according to their absolute connection strengths associated with this linear mode. The second method estimates the correlation with the partial prior knowledge genes without clustering (baseline correlation method-1). Genes are then ranked based on their absolute correlation coefficients between an individual gene expression profile and the average profile of partial prior knowledge genes. However, taking the average profile of all knowledge genes may reduce the sensitivity of detection, especially when the genes in KGP are not similar to each other. To overcome this problem, the third baseline method is a weighted correlation method based on a clustering approach (baseline correlation method-2). Similar to the multi-scale ICA method, the entire gene population is grouped into several sub-populations and a gene in each cluster is assigned a score. The score is the weighted absolute correlation coefficient between an individual gene expression profile and the average profile of partial prior knowledge genes in this cluster. The weight is then calculated using Equation (2.5) and genes are ranked according to their scores.

Given a ranked gene list and knowledge gene set, we can use the Receiver Operating Characteristic (ROC) curve [71] and the area under the curve (AUC) to measure the test accuracy for each biomarker identification method. ROC curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR). AUC is an important performance measure that provides an overall measure of accuracy for the test. Given a ranked gene list (g_1, g_2, …, g_n) with a total of n genes and the ground truth gene set G_k with k genes, true positive rate and false positive rate, when selecting top i genes G_i in the list, are calculated as follows:

$$TPR(i) = \frac{|G_i \cap G_k|}{k}, \tag{2.8}$$

$$FPR(i) = \frac{i - |G_i \cap G_k|}{n-k}. \tag{2.9}$$

Given the ground truth biomarkers, we can evaluate the performance of our proposed method through ROC and AUC analysis. However, it is also important to estimate the significance level (p-value) of a potential biomarker gene, especially with the false discovery rate (FDR) control when there is no ground truth. People are more interested in how many genes are statistically plausible as true biomarkers given a certain false discovery rate cut off. False discovery rate control is an effective statistical method used in multiple hypothesis testing to correct for multiple comparisons [72]. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors). In order to calculate FDR corrected p-value, we used gene score c_g to first calculate the p-value of each gene as the biomarker associated with the KGP. We assume that the gene score roughly follows an exponential distribution, which is consistent with a well-known biological constraint, i.e., only a few of genes will be involved in some specific biological processes. We then computed the p-value by estimating the probability of the observed gene score being more extreme than the ones following the estimated exponential distribution. Finally we used the Story procedure [73] to control the FDR and computed FDR corrected p-value for the biomarkers identified.

2.4. Experiments and results

We applied our knowledge-guided multi-scale ICA method to two gene expression profiling studies: (1) a yeast cell cycle microarray data set [74] and (2) an Rsf-1-induced microarray data set. The yeast cell cycle data set consists of the expression of 6178 Open Reading Frames (ORFs) during the cell replication cycle in the budding yeast (Saccharomyces cerevisiae). The data set consists of 77 samples corresponding to various experiment conditions. Approximately 800 genes have been identified as cycle-regulated genes; among these 104 genes have been well studied [74]. The goal of this experiment is to identify the cell cycle-regulated linear modes and then extract the corresponding genes associated with the cell cycle. We used the 104 genes as our training knowledge gene set and the remaining 704 genes as an independent test set for evaluation.

The Rsf-1-induced microarray data set was acquired and analyzed in our experiment. The dataset was generated using Affymetrix Human Genome U133 Plus 2.0 Arrays from an expression profiling study of ovarian cancer at the Johns Hopkins Medical Institutions. The study was designed to identify Rsf-1 regulated genes in ovarian cancer; Rsf-1 (also known as HBXAP) is a newly discovered gene frequently amplified in ovarian cancer [75]; the protein participates in chromatin remodeling which is essential for a variety of cellular functions including transcription, DNA replication, and DNA repair. The data set is composed of 7 samples with two biological conditions (Rsf-1-induced and not Rsf-1-induced) and four time points at 0 hour, 6 hours, 18 hours, and 30 hours. We used Affymetrix Probe Logarithmic Intensity Error (PLIER) algorithm with quantile normalization to preprocess the original intensity data for gene expression measurements [12]. After the preprocessing, we obtained expression measurements of 54,675 probe sets for each sample.

The EDGE algorithm was first applied to select statistically significant expressed genes from yeast cell cycle data and Rsf-1 induced ovarian cancer data, respectively. After ranking all genes in terms of their q-values estimated from EDGE, we calculated AUC values for yeast cell cycle-related genes and ovarian cancer-related genes, respectively (see below). As a result, both AUC values are relatively low (around 0.5), which indicates that the genes identified from pure data-driven methods (such as EDGE; without prior knowledge guidance) may not show strong biological relevance.

2.4.1. Yeast cell cycle data

To reduce computational complexity, k-means clustering was used to form the subsets of genes for both datasets. The number of independent components in the FastICA algorithm was set to five for this dataset, since our previous dimension estimation approach with a stability analysis procedure [76] showed that five independent components are sufficient to describe the gene expression data. We first conducted ten-fold cross-validation on the well studied 104 cell cycle-related genes. For each fold, the optimal cluster number is determined by a nested cross-validation procedure on the training gene set, as illustrated in Figure 2.2. The number of clusters ranges from 1 to 10. Notice that when the number is 1, no clustering is needed and the algorithm reduces to the baseline ICA method. Each ten-fold cross-validation is repeated 10-times with different randomly chosen stratified sets of knowledge genes. Since the k-means clustering method generates different results depending on its random initialization, we repeat the procedure ten times with different initializations to obtain more reliable results. The results reported here are the average results of the ten different initializations.

The resulting average AUC value of ten-fold cross-validation on 104 genes is 0.9206 with standard deviation of 0.0470. Figure 2.3 shows the histogram of determined optimal number of clusters during the ten-fold cross-validation procedure. From the figure we can see the most frequent number of clusters is five. Then we implemented three baseline methods for ten-fold cross-validation as comparisons. For baseline correlation method-2, we chose the optimal cluster number from the multi-scale ICA method for a fair comparison. The ROC curves of ten-fold cross validation for the two baseline correlation methods, the baseline ICA method, and our multi-scale ICA method are shown in Figure 2.4. The ROC curves show that the multi-scale ICA method outperforms the baseline correlation method-2, and that the baseline ICA approach is better than the baseline correlation method-1. Overall, the proposed multi-scale ICA method significantly outperforms all three baseline methods as estimated by the Kolmogorov-Smirnov (K-S) one-sided test (Table 2.1).

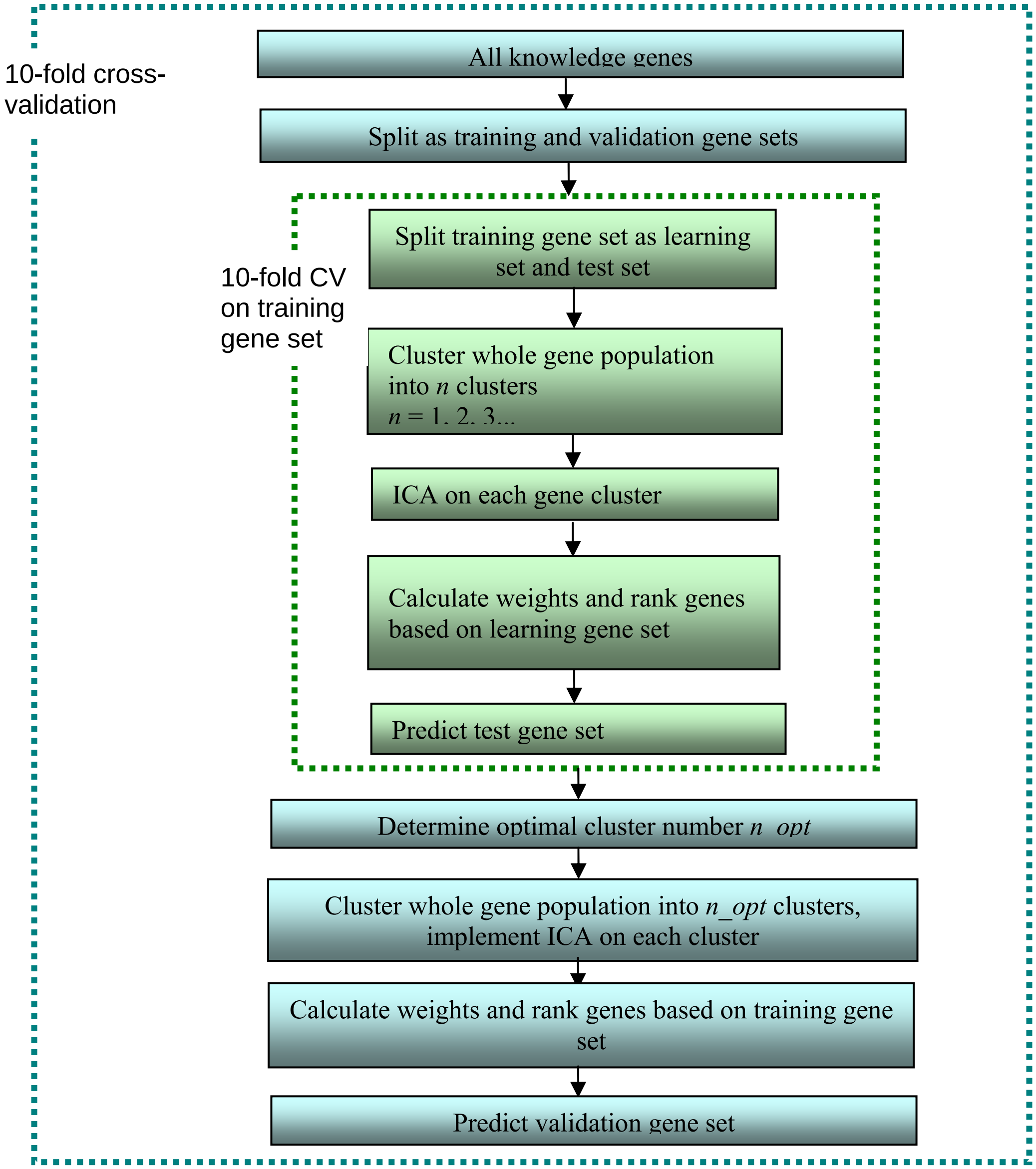

10-fold cross-validation
All knowledge genes
Split as training and validation gene sets
10-fold CV on training gene set
Split training gene set as learning set and test set
Cluster whole gene population into *n* clusters
n = 1, 2, 3...
ICA on each gene cluster
Calculate weights and rank genes based on learning gene set
Predict test gene set
Determine optimal cluster number *n_opt*
Cluster whole gene population into *n_opt* clusters, implement ICA on each cluster
Calculate weights and rank genes based on training gene set
Predict validation gene set

Figure 2.2. Procedure of ten-fold cross-validation. The optimal number of clusters is determined by a nested ten-fold cross-validation on training gene set.

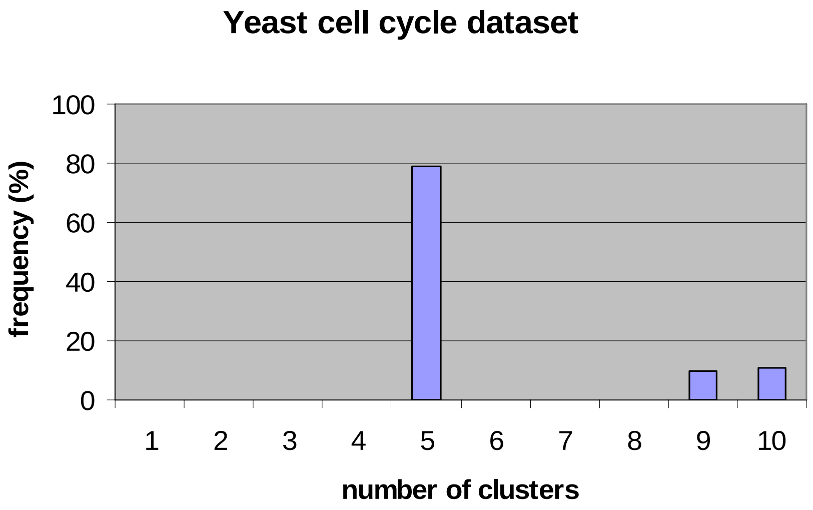


Figure 2.3. Histogram of determined optimal number of clusters in ten-fold cross- validation on yeast cell cycle data set.

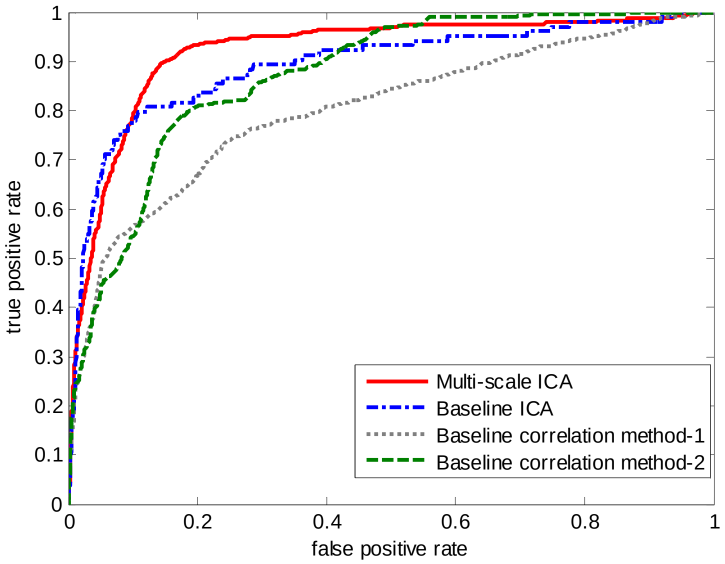


Figure 2.4. ROC curves of ten-fold cross-validation for four biomarker identification methods on training knowledge gene set of yeast cell cycle data set. Solid line represents the multi-scale ICA method; dash-dotted line represents the baseline ICA method; dotted line represents the correlation method-1; dash line represents the correlation method-2.

Table 2.1. P-values of Kolmogorov-Smirnov test for different methods on yeast cell cycle data using ten-fold cross-validation

Method 1	Method 2	P-values of K-S test
Optimal ICA	Baseline ICA	<1e-10
Optimal ICA	Correlation method 1	<1e-10
Optimal ICA	Correlation method 2	<1e-5

To further test the generalizability of our method, we conducted ten-fold cross-validation on the 104 genes using a subset of samples. The original data set includes 77 samples synchronized by three independent methods: α factor arrest, elutriation and arrest of a *cdc* 15 temperature-sensitive mutant [74] . We selected 63 samples from all the samples by excluding those samples under elutriation condition. The resulting average AUC value is 0.9157 with standard deviation of 0.0458. Also the most frequent optimal cluster number is five (with a frequency of 65%), which shows a great consistency when compared to the result using all the samples.

All 104 knowledge genes were then used as a training set in the algorithm to test 704 cell cycle-related genes for all four methods. During the training, we still used ten-fold cross-validation to determine the optimal number of clusters. Figure 2.5 shows the average AUC values and their standard deviations in ten-fold cross-validation across different number of clusters. From the figure we can see that the average AUC (standard deviation), starting at 0.892 (0.0006) for the full gene population, decreases a little at two and three clusters. The AUC increases gradually and reaches the peak of 0.9274 (0.0071) at five clusters, at which it remains constant. So the optimal number of clusters for multi-scale ICA approach is five. Then an independent evaluation was performed on the test gene set and the ROC curves for these four methods was calculated when the cluster number is five (Figure 2.6). The ICA-based methods significantly outperform the baseline correlation methods, and the multi-scale ICA is the best method when compared with the three baseline methods (Table 2.2).

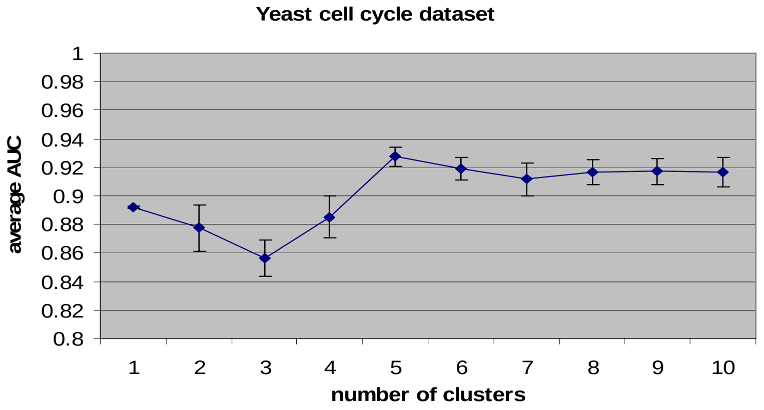


Figure 2.5. Average area under the curve (AUC) values using ten-fold cross-validation with different numbers of clusters on 104 knowledge genes. The knowledge-guided multi-scale ICA method is applied to yeast cell cycle data set for the identification of cell cycle-related genes.

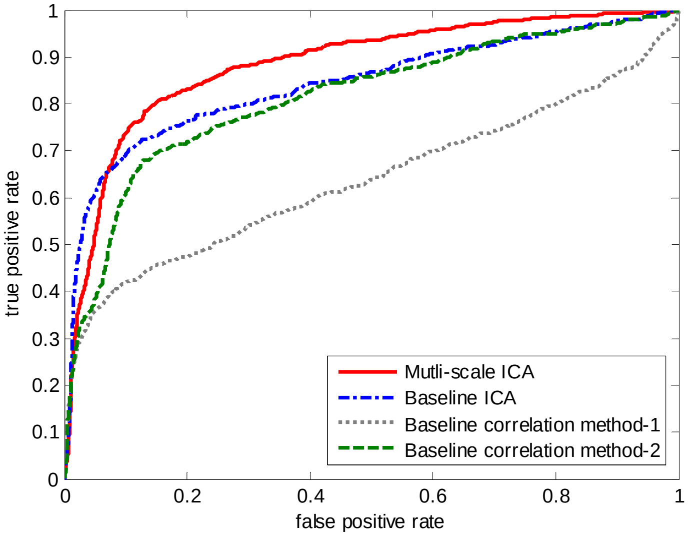


Figure 2.6. ROC curves of four biomarker identification methods on yeast cell cycle data set with an independent test gene set.

Table 2.2. P-values of Kolmogorov-Smirnov test for different methods on yeast cell cycle data using an independent test gene set.

Method 1	Method 2	P-value of K-S test
Optimal ICA	Baseline ICA	<1e-10
Optimal ICA	Correlation method 1	<1e-10
Optimal ICA	Correlation method 2	<1e-10

We examined in detail the extracted knowledge-relevant linear modes and the biological functions of their associated cell cycle-regulated genes. Figure 1.1 shows five knowledge-relevant linear modes and their weights as identified when the number of clusters is set at the optimum number of five (Figure 2.5). The top three linear modes have much higher weights than the lower two modes and their estimated TFAs clearly show periodic patterns related to cell cycle. We examined the biological functions of these well-known cell cycle-regulated genes associated with these three linear modes. The majorities of genes in linear mode L3 are associated with the M/G1 boundary or are known transcriptional targets of STE12/MCM1. Most of the genes in linear mode L1 are SCB/MCB regulated in late G1 and S phase. Finally, many genes in linear mode L2 are in S/G2 and G2/M phases. In summary, we can see that the linear modes L3, L1, and L2 correspond to different biological functions in cell cycle process.

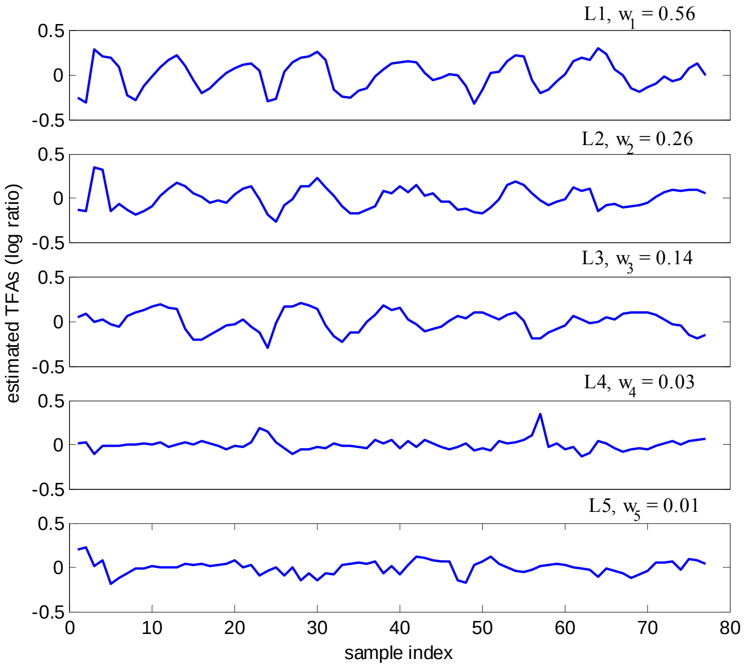


Figure 2.7. Five cell cycle-related linear modes in the proposed multi-scale ICA approach on yeast cell cycle data set. The weight is also listed in the figure for each linear mode.

Table 2.3. Top10 genes selected by the proposed multi-scale ICA method on yeast cell cycle data and their FDR corrected p-values < 1.0e-07.

Rank	ORF	Name	FDR corrected p-value	Short Description
1	YPL256C	CLN2	0.19e-10	CycLiN; G1 cyclin involved in regulation of the cell cycle
2	YOL007C	CSI2	0.19e-10	Chitin Synthesis Involved; protein of unknown function
3	YKR013W	PRY2	0.44e-10	Pathogen Related in Yeast; protein of unknown function
4	YDL003W	MCD1	0.15e-09	Mitotic Chromosome Determinant; expression is cell cycle regulated and peaks in S phase
5	YML027W	YOX1	0.15e-09	Homeodomain-containing transcriptional repressor
6	YBR088C	POL30	0.37e-09	POLymerase; proliferating cell nuclear antigen (PCNA)
7	YLR183C	TOS4	0.37e-09	Target of SBF; promoters of some genes involved in pheromone response and cell cycle;
8	YIL140W	AXL2	0.37e-09	AXiaL budding pattern; glycosylated by Pmt4p; potential Cdc28p substrate
9	YGR189C	CRH1	0.21e-08	Congo Red Hypersensitive; cell wall protein ; putative chitin transglycosidase
10	YER070W	RNR1	0.62e-08	RiboNucleotide Reductase; the RNR complex catalyzes the rate-limiting step in dNTP synthesis and is regulated by DNA replication and DNA damage checkpoint pathways via localization of the small subunits

The top 10 genes selected by multi-scale ICA method are listed in Table 2.3. Among them, four genes (CLN2, MCD1, POL30 and RNR1) are in the known training gene set. All other genes (CSI2, PRY2, YOX1, TOS4, AXL2 and CRH1) are the genes related to cell cycle beyond our training gene set, i.e., in the test gene set. The results show that our method is effective at finding novel biomarkers beyond knowledge, which is clearly an important feature of the proposed approach for novel biomarker identification beyond prior knowledge. In most of cases, especially for human disease, knowledge genes are limited and we need to infer the new ones from partial knowledge for biomarker discovery.

2.4.2. Rsf-1-induced gene expression data

Knowledge gene pool (KGP)

To construct the KGP, we started with the known gene Rsf-1 and its related genes, NF-kappa B (NFKB1) and SMARCA5 (also known as hSNF2H) as reported in [77], to search the databases. We used Ingenuity Pathway Analysis (IPA) to extract 95 genes that are thought to be directly related to NFKB1 and SMARCA5. Note that there is no network related to Rsf-1 in the current IPA database. We also included 43 genes from TRANSFAC 11.1 Professional Database [28], whose protein products are transcription factors biologically relevant to ovarian cancer as reported in literature. Hence, our KGP consists of 141 distinct Affymetrix probe set identifiers that represent the expression values for the 138 genes.

Multi-scale ICA results

We used '*tanh*' nonlinearity in the FastICA algorithm: other parameters were set at their default values. The number of the independent components is set to a maximum value of 6 due to the limitation of sample size. Ten-fold cross-validation was conducted on our partial prior knowledge genes, where the optimal cluster number was determined by a nested cross-validation approached for each fold as shown in Figure 2.2. The number of clusters was set from 1 to 15. We also repeated 10 times for ten-fold cross-validation and k-means clustering in order to generate more reliable results. The resulting average AUC is 0.7203 with standard deviation of 0.0804. Figure 2.8 shows the histogram of determined optimal cluster number in the ten-fold cross-validation procedure and we can see that the most frequent cluster number is 4. We compared the ROC curves for the two baseline correlation methods, the baseline ICA and the multi-scale ICA for ten-fold cross-validation (Figure 2.9). The results in Table 2.4 show that multi-scale ICA method performs significantly better than baseline ICA method and baseline correlation method-1 with p-value < 1e-10, while performing marginally better than baseline correlation method-2 (p-value = 0.0037). Since baseline correlation method-2 also calculates clustered average profiles of the prior knowledge genes, this result indicates that the multi-scale approach by clustering is an effective strategy to improve the performance for ovarian cancer-related biomarker identification. On the other hand, a major weakness in baseline correlation method-1 lies in that the average profile of all prior knowledge is used when their expression profiles are not similar to each other.

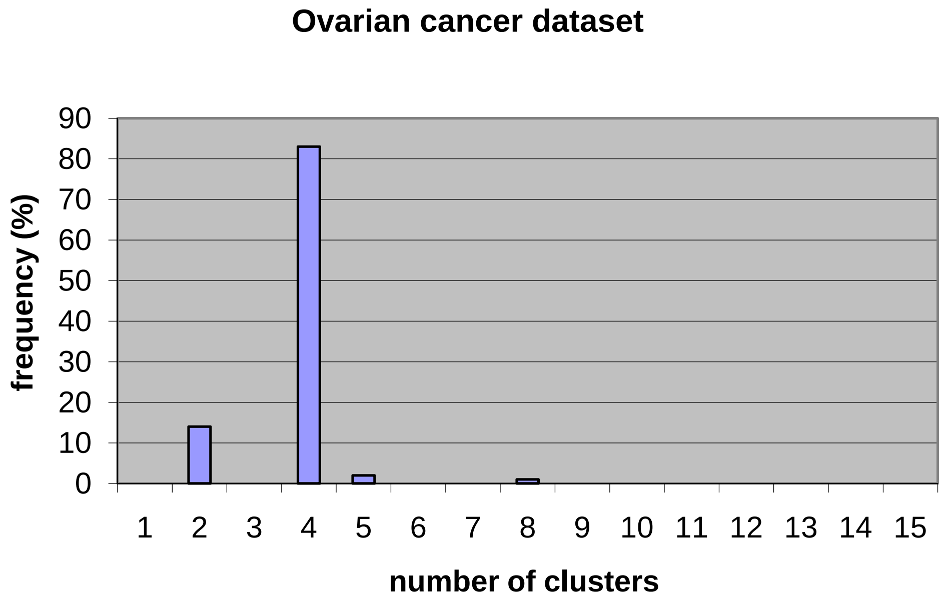


Figure 2.8. Histogram of determined optimal number of clusters in ten-fold cross- validation on ovarian cancer data set.

Table 2.4. P-values of Kolmogorov-Smirnov test for different methods on Rsf-1-induced ovarian cancer microarray data.

Method 1	Method 2	p-value of the K-S test
Optimal ICA	Baseline ICA	<1e-10
Optimal ICA	Correlation method 1	<1e-10
Optimal ICA	Correlation method 2	0.0037

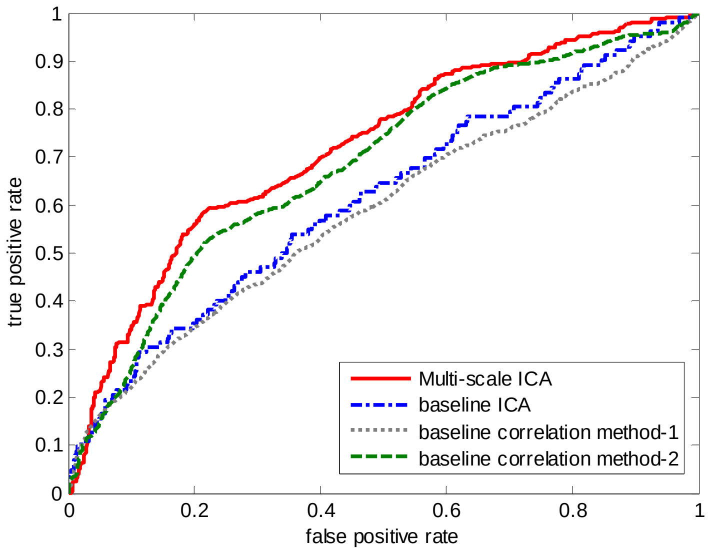


Figure 2.9. ROC curves of ten-fold cross-validation for four biomarker identification methods on knowledge gene set of ovarian cancer data set. Solid line represents the multi-scale ICA method; dash-dotted line represents the baseline ICA method; dotted line represents the correlation method-1; dash line represents the correlation method-2.

Evaluation by motif analysis

All knowledge genes were used as the training set in the algorithm to rank the whole gene population for all four methods. During the training, we still used ten-fold cross-validation to determine the optimal number of clusters in multi-scale ICA method. Figure 2.10 shows the average AUC values and their standard deviations obtained with different numbers of clusters for the ten-fold cross-validation; the average AUC (standard deviation), starting at 0.6146 (0.0004) for the whole gene population, increases to 0.7329 (0.0253) at two clusters and reaches the maximum value of 0.7343 (0.0210) at four clusters, and remains almost constant thereafter. Therefore, the optimal number of cluster for the multi-scale ICA approach was selected as four. Specifically, we examined estimated linear modes from ICA methods. Figure 2.11 shows the estimated knowledge-related TFAs using baseline ICA method and Figure 2.12 shows the estimated four knowledge-related TFAs and their weights using our multi-scale ICA method. We observe that one of the TFA patterns in Figure 2.12 (L3) is similar with that in Figure 2.11, which indicates that multi-scale ICA method can estimate more TFAs for knowledge-related genes than baseline ICA method. Four different linear modes and their weights in Figure 2.12 also indicate that the expression patterns of the genes in KGP are not similar to each other, which seems to be the major reason behind that baseline correlation method-1 (using the average profile of all prior knowledge) underperforms other methods.

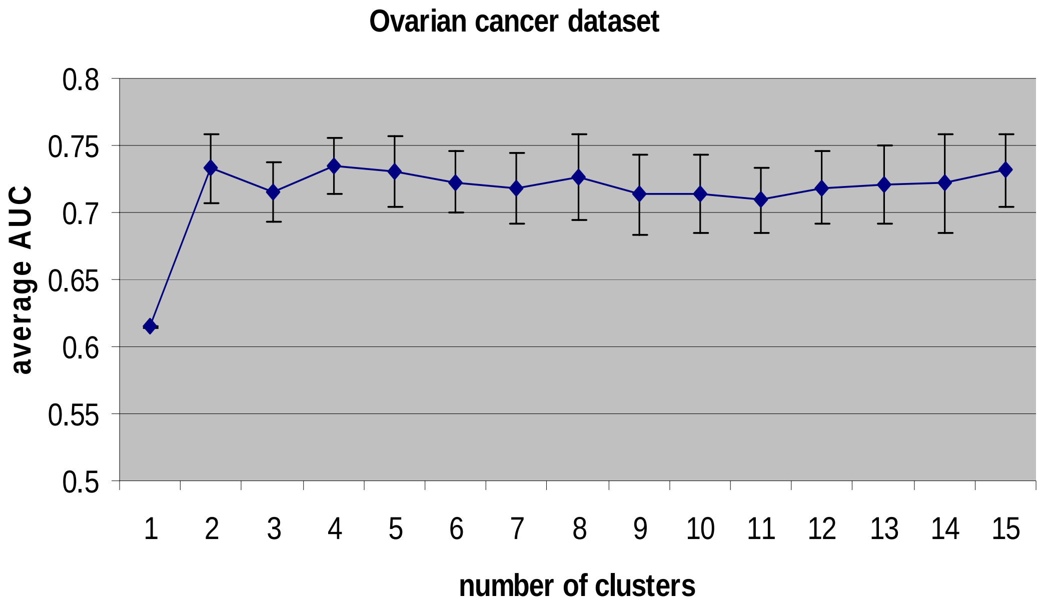


Figure 2.10. Average AUC values using ten-fold cross-validation across different numbers of clusters. The knowledge-guided multi-scale ICA method is applied to Rsf-1-induced ovarian cancer microarray data for the identification of disease-specific biomarkers.

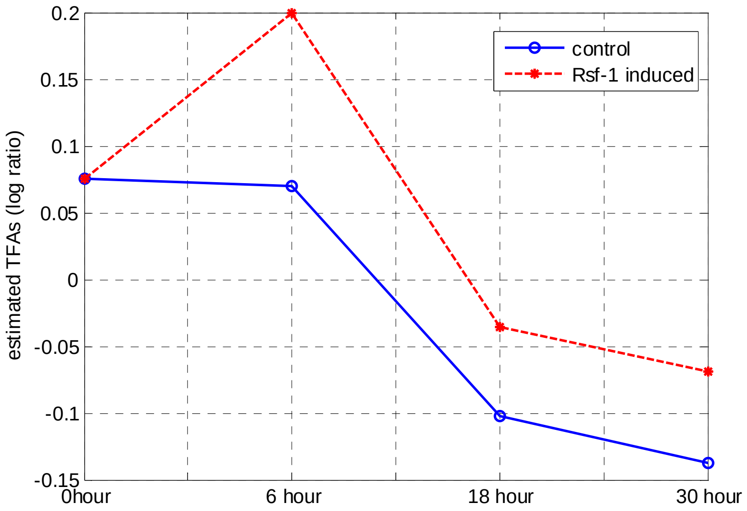
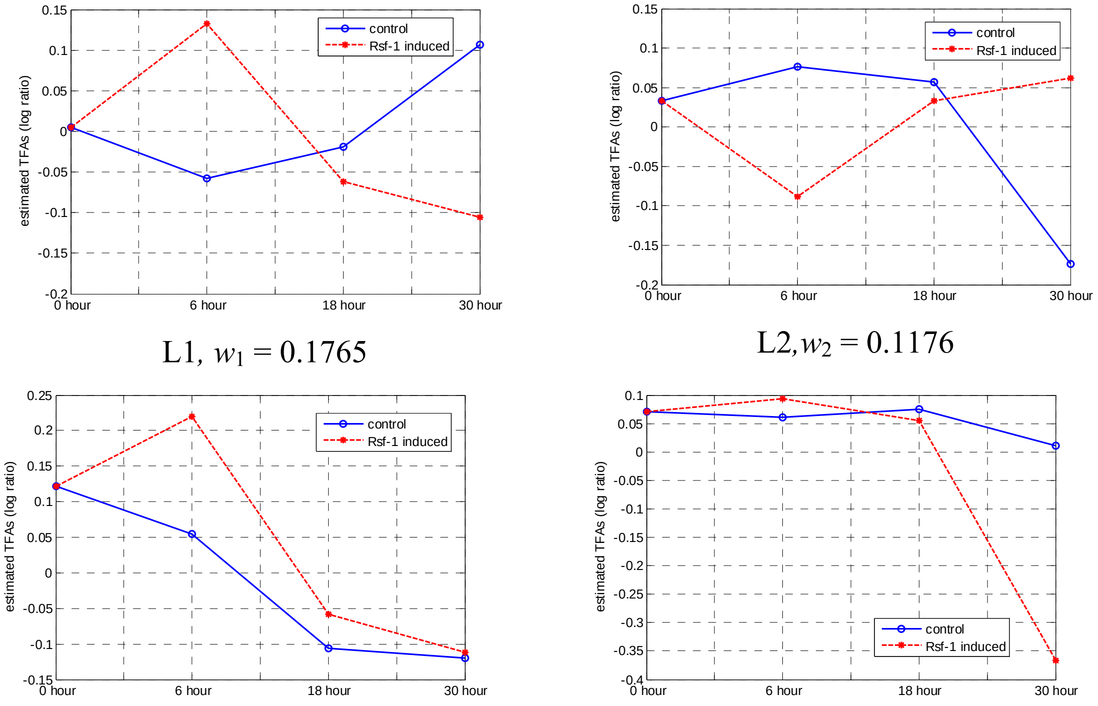


Figure 2.11. Estimated knowledge-related TFAs using baseline ICA method. X-axis represents the time and Y-axis represents the estimated TFAs.



L1, $w_1 = 0.1765$

L2, $w_2 = 0.1176$

L3, $w_3 = 0.3235$

L4, $w_4 = 0.3824$

Figure 2.12. Estimated four knowledge-related TFAs using the proposed multi-scale ICA method. X-axis represents the time and Y-axis represents the estimated TFAs.

For the final ranked gene lists, we performed motif enrichment analysis to evaluate the performance of each of the four different methods for biomarker identification. Specifically, among 43 ovarian cancer-related TFs extracted from TRANSFAC 11.1 Professional Database [28], 14 TFs have their PWMs available and we generated the gene-TF matrix $\boldsymbol{M}$ for them. For each TF, a PWM was chosen from the vertebrate non-redundant profiles. Table 2.5 lists their TRANSFAC PWM entry IDs and the corresponding TF descriptions. To increase the statistical power, we conducted multiple tests by selecting different gene sets with different sizes for different gene selection methods. The number of genes in each gene set ranges from 100 to 1,000 and the average p-values for 14 TFs are reported. Figure 2.13 shows the average p-values of TFs enrichment for different gene sets selected by different methods. Both ICA methods outperform the baseline correlation methods in terms of finding more enriched ovarian cancer-related TFs binding sites. Moreover, our multi-scale ICA method is slightly better than baseline ICA method for motif enrichment. It is worth noting that although both multi-scale ICA and baseline ICA methods can extract ovarian cancer-related biomarkers with significant motif enrichment, multi-scale ICA method can help reveal more biomarkers related to ovarian cancer. For this experiment, it is also expected to have similar TF enrichment from both methods, since one common linear mode is revealed by both methods (i.e., the mode in Figure 2.11 is very similar with the L3 mode in Figure 2.12). From the pattern of this common mode, we postulate that this is a major mode related to *RSF-1*-induced ovarian cancer. Therefore, the genes extracted from this mode will show a similar significance level in TF enrichment (as shown in Figure 2.13). However, the multi-scale ICA approach can extract other linear modes related to ovarian cancer (see Figure 2.12). Apparently, the biomarkers related to these other modes cannot be identified with the baseline ICA approach. This can be supported by the ROC curves in Figure 2.9, showing an improved performance of using multi-scale ICA approach compared to that of using baseline ICA approach.

Table 2.5. Ovarian cancer-related TFs and their TRANSFAC entry IDs & descriptions.

Index	TF Name	PWM Access No.	Consensus Binding Site	Factor Description
1	AP-2	M00189	MKCCCSCNGGCG	Activator protein 2
2	AP-2alpha	M00469	GCCNNNRGS	Activating enhancer binding protein 2 alpha
3	AP-2alphaA	M01045	ANNGCCTNAGGSNNT	Activating protein 2, AP-2A, Ker-1
4	AP-2gamma	M00470	GCCYNNGGS	Activator protein 2gamma, ERF-1
5	AP-2rep	M00933	CCCCGCCCCN	Specificity protein1, stimulating protein 1
6	BRCA1	M01082	KTNNGTTG	Breast cancer type 1 susceptibility protein
7	E2F	M00516	TTTSGCGCGMNR	EIIF protein, activator of myc, important for p107 promoter activity
8	Elk-1	M00007	NAAACMGGAAGTNCVH	Elk1, member of ETS oncogene family
9	NF-kappaB	M00774	NNNNKGGRAANTCCCN	Nuclear factor kappa B, p50
10	Sp1	M00933	CCCCGCCCCN	Specificity protein1, stimulating protein 1
11	TGIF	M00418	AGCTGTCANNA	5’-TG-3’ interacting factor, TG-interacting factor, TGFB-induced factor
12	c-Rel	M00053	SGGRNTTTCC	Nuclear factor kappa B c-Rel, p68
13	P53	M00272	NGRCWTGYCY	Tumor protein p53, TRP53
14	ER	M00191	NNARGNCANNNTGACCYNN	Estrogen receptor

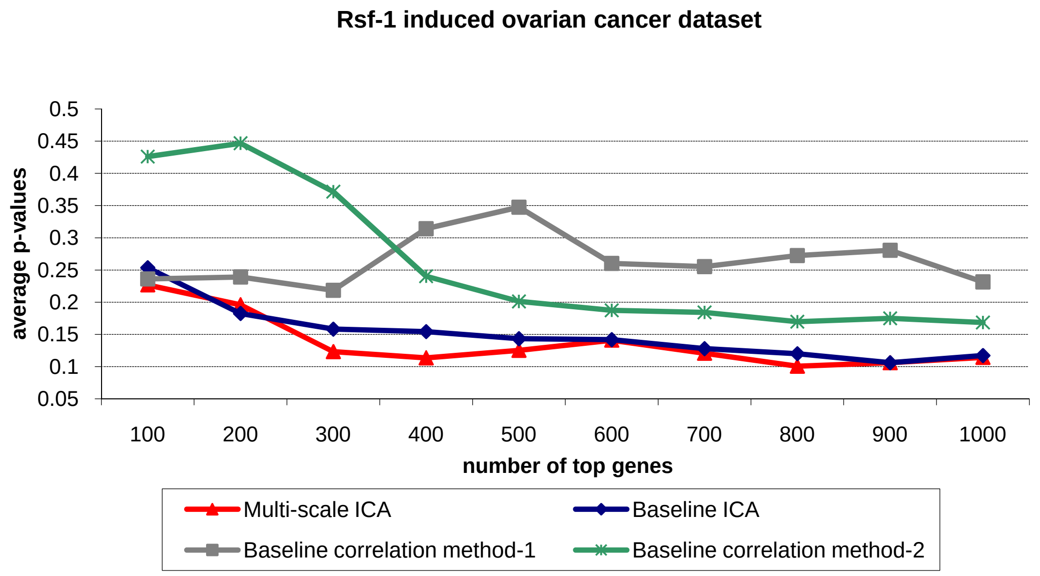


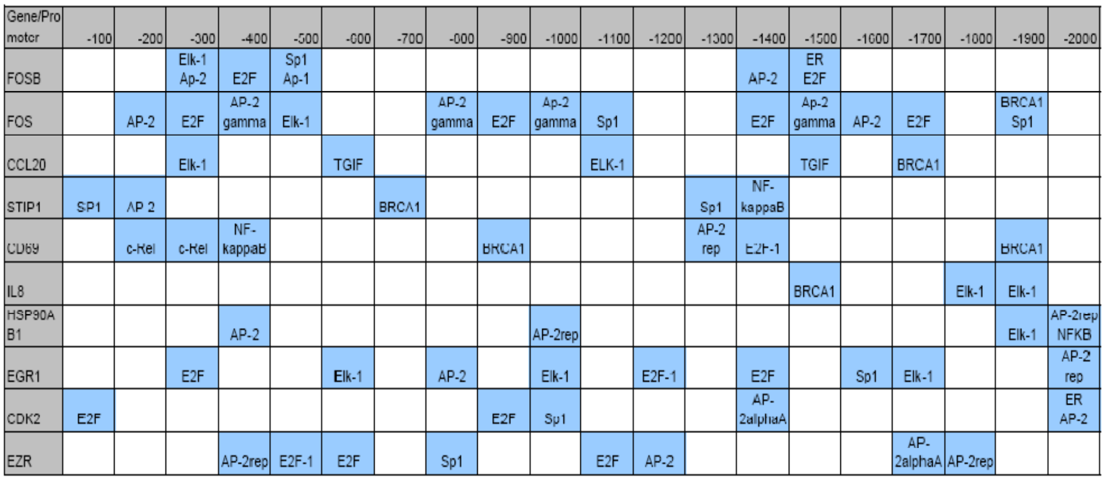
Figure 2.13. Average p-value of TF enrichment for different gene sets associated with different methods on Rsf-1-induced ovarian cancer microarray data set.

Discussion with biological interpretation

To enable a more detailed analysis, the top 10 genes extracted by optimal multi-scale ICA method are listed in Table 2.6 and the putative TFs in their promoter regions are shown in Figure 2.14. Since none of the genes are in the KGP, they were entered into an Ingenuity Pathways Analysis (IPA) where we found that all of these genes can be incorporated into a single hypothetical network (Figure 2.15). The major functions of this network are involved in gene expression, cancer development, and cellular motility. Five genes, FOSB, FOS, EGR1, IL8 and CDK2, are in the cancer module with p-values ranging from 1.84E-7 to 6.5E-3. FOSB and FOS belong to the Fos family that hetero-dimerizes with Jun proteins to form the AP-1 transcription factor complex [78]. AP-1 transcription factors control rapid responses of mammalian cells to stimuli that are associated with proliferation, differentiation and transformation [79]. IL-8 is a member of the C-X-C family of chemokines, and overexpression of IL-8 is observed in subsets of human ovarian cancer cells [80]. Previous studies have shown that the expression of interleukin-8 (IL-8) is directly correlated with the progression of human ovarian carcinomas implanted into the peritoneal cavity of nude mice [81]. The early growth response 1 (EGR1) is a transcription factor that acts as both tumor suppressor and tumor promoter depending on the cellular context. In the experiments of multiple pituitary and ovarian defects in Krox-24 (NGFI-A, Egr-1)-targeted mice, EGR1 was implicated as a novel key regulator of anterior pituitary physiology and that it may play important roles in specific cell lineages [82]. CDK2 is known to be involved in cell cycle regulation and the overexpression of CDK2 is associated with malignancy in ovarian tumors [83].

Table 2.6. Top 10 genes selected by the proposed multi-scale ICA on Rsf-1-induced microarray data and their FDR corrected p-values < 0.006.

Rank	Probe Set ID	Gene Symbol	FDR corrected p-value	Gene Full Name
1	202768_at	FOSB	0.0058	FBJ murine osteosarcoma viral oncogene homolog B
2	209189_at	FOS	0.0058	v-fos FBJ murine osteosarcoma viral oncogene homolog
3	205476_at	CCL20	0.0058	chemokine (C-C motif) ligand 20
4	212009_s_at	STIP1	0.0058	stress-induced-phosphoprotein 1
5	209795_at	CD69	0.0058	CD69 molecule
6	211506_s_at	IL8	0.0058	interleukin 8
7	1557910_at	HSP90AB1	0.0058	heat shock protein 90kDa alpha (cytosolic), class B member 1
8	227404_s_at	EGR1	0.0058	Early growth response 1
9	211804_s_at	CDK2	0.0058	cyclin-dependent kinase 2
10	208621_s_at	VIL2	0.0058	villin 2



Gene/Promotor	-100	-200	-300	-400	-500	-600	-700	-800	-900	-1000	-1100	-1200	-1300	-1400	-1500	-1600	-1700	-1800	-1900	-2000
FOSB			Elk-1 Ap-2	E2F	Sp1 Ap-1									AP-2	ER E2F					
FOS		AP-2	E2F	AP-2 gamma	Elk-1			AP-2 gamma	E2F	Ap-2 gamma	Sp1			E2F	Ap-2 gamma	AP-2	E2F		BRCA1 Sp1	
CCL20			Elk-1			TGIF					ELK-1				TGIF		BRCA1			
STIP1	SP1	AP 2					BRCA1						Sp1	NF-kappaB						
CD69		c-Rel	c-Rel	NF-kappaB					BRCA1				AP-2 rep	E2F-1					BRCA1	
IL8															BRCA1			Elk-1	Elk-1	
HSP90AB1				AP-2						AP-2rep									Elk-1	AP-2rep NFKB
EGR1			E2F			Elk-1		AP-2		Elk-1		E2F-1		E2F		Sp1	Elk-1			AP-2 rep
CDK2	E2F								E2F	Sp1				AP-2alphaA						ER AP-2
EZR				AP-2rep	E2F-1	E2F		Sp1			E2F	AP-2					AP-2alphaA	AP-2rep		

Figure 2.14. TFs and their locations in 2Kbp promoter region for top 10 genes selected by our approach. The promoter region is represented from -2,000bp to 0bp from TSS and each block in the figure represents a 100bp region.

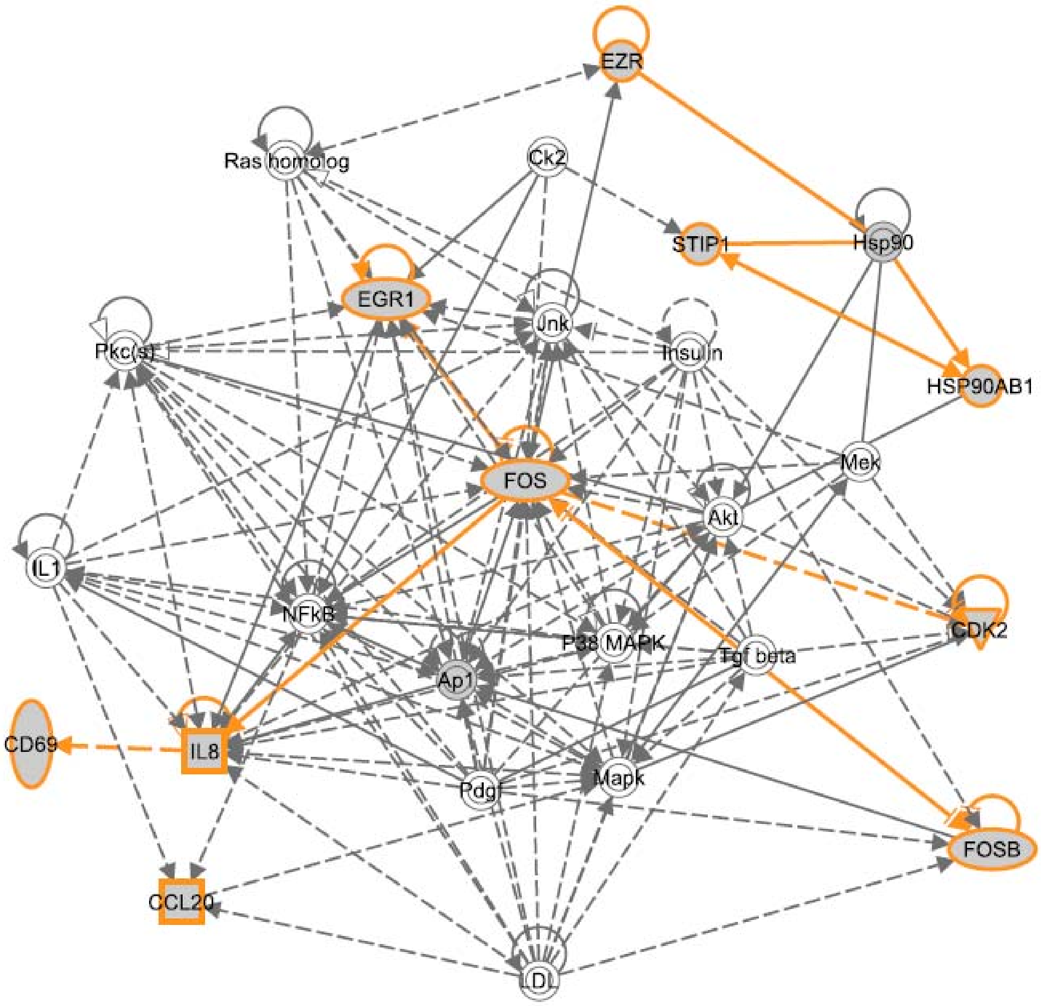


Figure 2.15. The network obtained from IPA with all of top 10 genes in Table 6. Five genes, FOSB, FOS, EGR1, IL8 and CDK2, are highly related to cancer module.

2.5. Discussion and conclusion

Biomarker identification is an important goal in many microarray data analyses. We propose a novel method, knowledge-guided multi-scale ICA, to find relevant biomarkers associated with specific biological functions. We aimed to infer knowledge-relevant regulatory signals and then identify corresponding biomarkers through a multi-scale strategy. A knowledge gene pool is constructed from multiple knowledge sources to help identify disease-specific gene clusters. By applying ICA to multi-scale gene clusters, an examination of the revealed regulatory modes can uncover knowledge of the underlying biological regulatory mechanisms. In addition, we have designed a statistical test procedure to measure the transcription factor enrichment of a selected gene set based on motif information. The approach was successfully applied to two gene expression profile data sets to identify biomarkers: yeast cell cycle microarray data and Rsf-1-induced microarray data. The experimental results show that our method can extract apparently biologically meaningful and condition-related biomarkers. The performance of the proposed method significantly outperforms several baseline methods for biomarker identification. More importantly, the proposed method has notable potential to discover novel biomarkers beyond any partial prior knowledge.

3 Multi-level Support Vector Regression Analysis to Identify Condition-specific Regulatory Networks

3.1. Introduction

Identifying regulatory modules is one of the key steps to understanding the molecular mechanisms of biological processes, especially important for defining the deregulated pathways in cancer. At the transcriptional level, a regulatory module is defined as a set of genes controlled by one or several transcription factors (TFs) in a condition-specific manner [38]. Transcription factors can either activate or inhibit gene expression, usually by binding to short, highly conserved, DNA sequences in the promoter (or upstream) region, i.e., transcription factor binding site (TFBS) or binding motif. In higher eukaryotes, TFBSs are often organized in clusters called *cis*-regulatory modules (CRMs). Many computational methods have been developed to facilitate the identification of CRMs from either gene expression data or DNA sequence data. Expression-based methods [38, 84, 85] take advantage of gene expression data but lack of sequence binding constraints. Sequence-based module discovery algorithms, such as CisModule [25], CREME [86] and ModuleSearch [87], analyze the promoter regions of a set of co-regulated genes to identify overrepresented motif combinations. A major limitation of sequence-based methods is that they do not consider the condition-specific nature of regulatory modules, i.e., they ignore the relationship between binding affinities and gene expression levels.

A living cell is a dynamic system in which gene activities and interactions exhibit temporal patterns and spatial compartmentalization [88]. Recently, several studies have shown that binding of transcription factors not only depends on their affinity for the binding sites, but binding also occurs in a condition-specific manner in response to various environmental changes [89, 90]. Thus, a transcription factor may play different regulatory roles to its downstream target genes or may even have different downstream targets under different conditions [89]. Motivated by this understanding, many computational algorithms were proposed to discover condition-specific regulatory modules by integrating condition-specific gene expression profiles and motif information. Regression models are widely used to combine these two types of information [91-95] . For example, a least square regression (LS-regression) method described by [95] identifies significant regulators by combining mRNA expression level and ChIP-on-chip binding data to minimize a fitting error. GRAM [26] is another regression method based on an iterative search to identify significant regulators and target genes. Bayesian models have also been used for regulatory module identification. A thermodynamic model [90] was proposed to predict expression patterns from regulatory sequence data in *Drosophila* segmentation. COGRIM [96] is a Bayesian hierarchical model with Gibbs Sampling implementation that integrates gene expression data, ChIP binding data, and transcription factor motif information to identify regulatory modules.

While these methods have achieved some degree of success, a high false-positive prediction rate is still a major problem mainly due to the noises in motif information and gene expression data. To reduce the false-positive rate, we propose a novel method, namely multi-level regulatory module identification through support vector regression (ml-SVR), to help find significant and stable regulatory modules. The ml-SVR method is particularly effective because of several novel adaptations: 1) a two-stage support vector regression (SVR) method is used to integrate binding motif information and gene expression data, aiming to improve the noise-tolerance capability; 2) a significance analysis procedure is applied to identify statistically significant regulatory modules; 3) a multi-level analysis strategy is developed to reduce the false-positive rate for reliable regulatory module identification; and 4) a weighted voting scheme is implemented for target gene identification, taking into account the entire multi-level analysis.

We have applied the ml-SVR method to simulation data and yeast cell cycle data to assess its performance for gene module identification, in comparison with existing methods. The comparison results clearly demonstrate that the proposed ml-SVR method notably outperforms other methods. We then applied our method to two breast cancer microarray data sets to identify condition-specific regulatory modules, respectively, in response to different estrogen conditions. The experimental results show that our method can successfully identify biologically meaningful modules associated with estrogen signaling and action in breast cancer.

Organization of this chapter is the following. In Section 3.2, we formulate a linear regression model for transcriptional regulatory network and introduce multi-level two stage support vector regression method for regulatory network identification. Statistically evaluation of the identified regulatory network is discussed and the convergence of two-stage SVR is proved in this section. In Section 3.3, we evaluate the proposed method on simulation data, yeast cell cycle data. Then we conduct the experiments on two conditions of breast cancer cell line data. Detailed experimental results is shown and discussed. In Section 3.4, we discuss the potential application and further study and we draw the conclusion in Section 3.5.

3.2. Methods

The multi-level support vector regression (ml-SVR) method is aimed to identify significant condition-specific regulatory modules by integrating mRNA gene expression data and binding motif information. Figure 3.1 illustrates the flow chart of the ml-SVR approach, shown as an iterative procedure in a nutshell. This multi-level analysis procedure, as conducted in a coarse-to-fine way, ensures that a condition-specific regulatory module becomes ever more significant as more relevant gene sets are formed at finer levels. At each level, support vector regression (SVR) is used to integrate binding motif information and gene expression data. Specifically, a two-stage SVR method is implemented to refine the estimation of transcription factor activity (TFA) and binding strength. Significance analysis of regulatory modules is achieved by evaluating the regression fitting errors compared to a baseline without motif information; an F-statistic is calculated from a permutation test to assess the significance (p-value) of a regulatory module. Finally, with the multi-level analysis, significant gene modules can be determined and their target genes identified by a voting scheme running through all levels. In the following subsections, we provide a detailed description of each component in the ml-SVR approach.

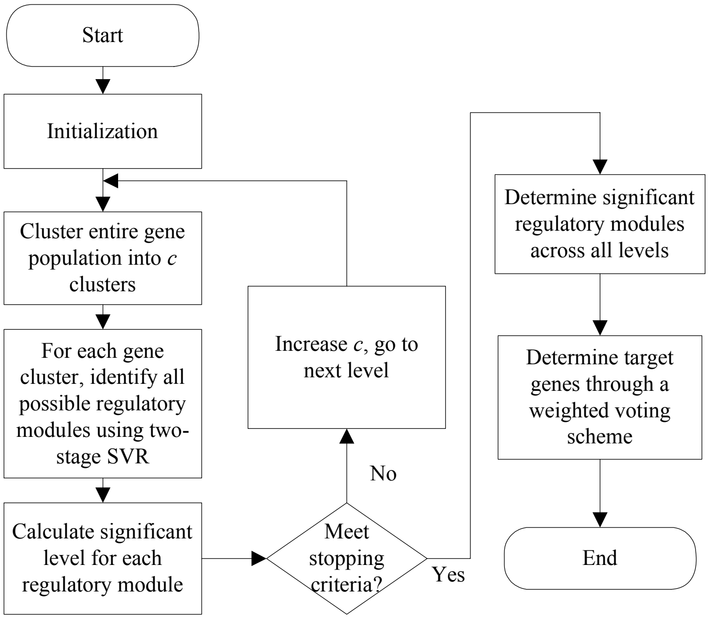

Start
Initialization
Cluster entire gene population into *c* clusters
For each gene cluster, identify all possible regulatory modules using two-stage SVR
Calculate significant level for each regulatory module
Meet stopping criteria?
No
Increase *c*, go to next level
Yes
Determine significant regulatory modules across all levels
Determine target genes through a weighted voting scheme
End

Figure 3.1. Flow chart of the multi-level support vector regression (ml-SVR) approach.

3.2.1. Sequence analysis for motif information

ChIP-on-chip, also known as genome-wide location analysis, is a technique that can isolate and identify DNA sequences occupied by specific DNA binding proteins [97]. However, it is not a trivial task to measure the binding strengths for all transcription factors (TFs) from ChIP-on-chip experiments due to the limited antibodies available, especially for higher eukaryotes studies. An alternative and practical way is to extract binding motif information from the promoter regions of focused genes. Motif information is usually represented by a position weight matrix (PWM) that contains log-odds weights for computing a match score between a binding site and an input DNA sequence. Many algorithms have been developed to either *de novo* discover motifs given multiple input sequences [25, 98] or search the known motifs in a given sequence based on their PWMs [70, 99]. Among them, Match^{TM} [70] takes DNA sequences as input, searches for potential TF binding sites using a library of PWMs, and outputs a list of potential sites in the sequence. The search algorithm uses two score values: the matrix similarity score (*mss*) and the core similarity score (*css*). These two scores measure the quality of a match between the sequence and the matrix, ranging from 0 to 1.0, where 0 denotes no match and 1.0 an exact match. The core of each matrix is defined as the first five most conserved consecutive positions of a matrix.

We assume that the binding strength for a specific transcription factor to its target gene is proportional to the similarity score of its binding site and the number of occurrences of the binding site in the gene promoter region. Here, all human promoter DNA sequences were obtained from the UCSC Genome database [69] (upstream 2,000 bp from the transcription start site (TSS)). With all vertebrate PWMs provided by the TRANSFAC 11.1 Professional Database [28], Match$^{\mathrm{TM}}$ algorithm is used to generate a gene-motif binding strength matrix **X** = [x_{gm}] with the cut offs that minimize the false-positive rate. The rows in the matrix **X** correspond to different genes, and the columns correspond to different binding sites (or motifs). Each element x_{gm} represents the binding strength at motif *m* in the promoter region of a gene *g*, which is calculated mathematically as follows:

$$x_{gm} = \sum_{i=1}^{N} \frac{1}{2}(mss_{gmi} + css_{gmi}), \tag{3.1}$$

where *N* is the number of occurrences of motif *m* in the promoter region of gene *g*; mss_{gmi} and css_{gmi} are the matrix similarity score and core similarity score for motif *m* and gene *g* in the i^{th} hit, respectively.

It is worth noting that although in this study, we opt to define the initial motif binding strength based on motif score and number of occurrences, there are other factors that are related to motif binding strength; phylogenetic conservation and distance to transcript start site (TSS) are two important factors that likely have influence on the binding strength. Phylogenetic conservation is an important evidence for motif information that considers the number of orthologous upstream regions in other genomes containing a particular binding site. Using phylogenetic conservation score could reduce false positive hits but also could lose the sensitivity dramatically, especially for high eukaryotes. The reason is that phylogenetic conservation score is based on the orthologous and conserved upstream regions from whole genome alignments cross different species. It would be difficult if many genomes are far away from the studied one. For example, a previous study has shown that the conserved block coverage for human and mouse is only 23.30% [100]. Cautions are also needed when considering the distance to TSS for binding strength in high eukaryotes, especially human genome. There is evidence showing that the binding strength is not as simple as a linear relationship with the distance to TSS. For example, a recent genome-wide ChIP-on-chip study shows that only 4% of estrogen receptor binding sites are mapped to 1,000bp promoter-proximal regions [101]; many sites are located in regions of 5,000bp to 10,000bp away from TSS. Nevertheless, it is important to incorporate the conservation and distance to TSS into a comprehensive definition of motif binding strength if the information is available and more critically, reliable for a particular study.

3.2.2. Two-stage support vector regression to infer regulatory modules

Suppose that there are *G* genes and *T* gene expression profiles, we represent microarray gene expression data as a matrix $\mathbf{Y}_{G\times T} = [y_{gt}], g = 1,\cdots,G; t = 1,\cdots,T$, where each element y_{gt} is the log-ratio of the expression level of gene *g* in sample *t* to that of the control sample. We also assume that there are *M* motifs on this gene set and the corresponding gene-motif binding matrix is $\mathbf{X}_{G\times M} = [x_{gm}], g = 1,\cdots,G; m = 1,\cdots,M$, where x_{gm} is the binding strength on motif *m* in the promoter region of gene *g*. The relationship between gene expression level and binding strength can be mathematically described by a linear model as follows:

$$\mathbf{Y}_{G\times T} = \mathbf{X}_{G\times M}\mathbf{A}_{M\times T} + \mathbf{N} , \qquad (3.2)$$

where $\mathbf{A}_{M\times T} = [a_{mt}], m = 1,\cdots,M; t = 1,\cdots,T$ is the TF activity matrix and **N** the noise matrix. Biologically, the model represents the log-ratio of gene expression levels expressed as a linear combination of log-ratios of transcription factor activities (TFAs) (denoted as a_{mt}) weighted by their binding strengths (i.e., x_{gm}) [63].

If **X** and **Y** are known, the solution to the linear model (Equation (3.2)) can then be easily obtained by a simple regression [102]. However, since both motif information and gene expression data are noisy, a simple regression will inevitably introduce a large number of false positive predictions. To alleviate this problem, we propose a two-stage support vector regression (SVR) method to specifically address the noises in motif information and gene expression data. SVR has been shown to have good robust properties against noise through the regularization term in its cost function [103]; the regularization term is intended to keep the estimated TF activity (in matrix **A**) as smooth as possible so as to combat the noise in gene expression data (**Y**). The ε-insensitive loss function is used in SVR to ensure the existence of the global minimum and a high tolerance to noise, which is defined by

$$L_{\varepsilon}(y_{gt}) = \begin{cases} 0, & \text{if } \left| y_{gt} - \hat{y}_{gt} \right| < \varepsilon \\ \left| y_{gt} - \hat{y}_{gt} \right| - \varepsilon, & \text{otherwise} \end{cases} , \tag{3.3}$$

where $\hat{y}_{gt}$ is the estimated value of expression log-ratio y_{gt}.

The goal of SVR is to find a function f that minimizes the loss function while keeping as flat as possible. By introducing slack variables ξ_g and ξ_g^* for soft margin, we can formulate the optimization problem as follows [104]:

$$\text{Minimize } \frac{1}{2}\|\mathbf{a}\|^2 + C\sum_{g \in G}(\xi_g + \xi_g^*),$$

$$\text{subject to} \begin{cases} y_g - < \mathbf{a}, \mathbf{x}_g > \le \varepsilon + \xi_g \\ < \mathbf{a}, \mathbf{x}_g > - y_g \le \varepsilon + \xi_g^* \\ \xi_g, \xi_g^* \ge 0 \end{cases} . \tag{3.4}$$

The constant C > 0 determines the tradeoff between the flatness of f and the amount up to which deviations larger than ε are tolerated.

By further introducing non-negative Lagrangian multipliers α_g and α_g^*, we can formulate the above optimization problem to be the following one of maximizing the dual Lagrangian function with respect to α_g and α_g^* [104]:

Maximize

$$-\frac{1}{2}\sum_{g_i, g_j \in G}(\alpha_{g_i} - \alpha_{g_i}^*)(\alpha_{g_j} - \alpha_{g_j}^*) < x_{g_i}, x_{g_j} > - \varepsilon\sum_{g \in G}(\alpha_g + \alpha_g^*) + \sum_{g \in G} y_g(\alpha_g - \alpha_g^*),$$

$$\text{subject to } \sum_{g \in G}(\alpha_g - \alpha_g^*) = 0, \quad \alpha_g, \alpha_g^* \in [0, C]. \tag{3.5}$$

By solving the above optimization problem, we can finally obtain the solution to the regression problem as follows [104]:

$$\mathbf{a} = \sum_{g \in G} (\alpha_g - \alpha_g^*) \mathbf{x}_g \text{ , } \hat{y}_g = f(\mathbf{x}_g) = \sum_{g \in G} (\alpha_g - \alpha_g^*) \langle \mathbf{x}_g, \mathbf{x}_g \rangle . \quad (3.6)$$

To combat the noise in motif information, we use a similar strategy as in the two-stage approach proposed by Yu et al. [94] to update the binding strength matrix **X** based on **Y** and the estimated **A.** In this way, we can reduce the number of false binding motifs, which are initially present in the binding strength matrix **X** but with no support from gene expression data (**Y**) and estimated TF activity (**A**).

The two-stage SVR method is implemented as an iterative procedure, which updates matrices **A** and **X** alternately until converged. In the implementation, we normalize (or standardize) the gene expression data to 0 mean and 1 standard deviation. We also standardize the estimated TF activity at each iteration step of our algorithm. The final algorithm of our two-stage SVR approach can be summarized as follows:

(1) Estimate **A** using **X** and **Y**. For each column vector $\mathbf{y}_t$ in matrix **Y**, regress $\mathbf{y}_t$ against **X** based on $y_{gt} = f(\mathbf{x}_g) = \sum_{m=1}^{M} x_{gm} a_{mt}$; calculate regression coefficient a_{mt} using ε-insensitive SVR.

(2) Update **X** using **A** and **Y**. For each row vector $\mathbf{y}_g$ in matrix **Y**, regress $\mathbf{y}_g$ against **A** based on $y_{gt} = f(\mathbf{a}_t) = \sum_{m=1}^{M} x_{gm}^{'} a_{mt}$; calculate regression coefficient $x_{gm}^{'}$ using ε-insensitive SVR; update **X** by $\mathbf{X} = \mathbf{X} + \eta(\mathbf{X}' - \mathbf{X})$, where η is a parameter in the range of (0, 1). (Note that η is set to 0.2 in our experiments.)

(3) Repeat Step (1) and Step (2) until converged. The convergence criterion is defined as the average correlation coefficient of TF activities between two successive iterations is larger than a predefined threshold r_0. (Note that r_0 is set as 0.9 in our experiments.)

The two-stage SVR can be proved to be converged and we will give the proof in Section 3.2.5.

3.2.3. Significance analysis of regulatory modules

A significance analysis procedure is designed to test if a selected motif set is statistically associated with the regulation of a given gene set, aiming to identify active regulators for that set. The null and alternative hypotheses *(H_0* and H_1, respectively) are given as follows:

H_0: The motif set is not actively involved in regulating a given gene set;
H_1: The motif set is actively involved in regulating a given gene set.

We use a summary statistic to represent the fitting results as described below:

$$\begin{aligned} &F = \frac{RSS_0 - RSS_1}{RSS_1} \\ &RSS_0 = \sum_{g,t} (y_{gt} - \overline{y}_t)^2, \ \overline{y}_t = \frac{1}{|G|} \sum_{g \in G} y_{gt} \quad , \\ &RSS_1 = \sum_{g,t} (y_{gt} - \hat{y}_{gt})^2, \ \hat{y}_{gt} = \sum_m a_{mt} x_{gm} \end{aligned} \tag{3.7}$$

where RSS_0 is the residual sum of squares without motif information, and RSS_1 is the residual sum of squares with motif information. The above equation is proportional to the typical F-statistic used to compare two models [105]. To calculate the p-value, we use the permutation method described below to form the null distribution. For a given motif set, we randomly select a gene set G_0 with the same size of G from the entire gene population, and then repeat B times to generate the corresponding null statistic score F^{0b}, for b = 1,2,…, B (B = 1,000 in our experiments). The p-value can be obtained for each gene set by calculating the probability that a null gene set has a statistic more extreme than the observed statistic. Mathematically, the p-value is calculated by the following equation:

$$p = \Pr_{H_0}(F^{0b} > F) = \frac{\#\{b : F^{0b} > F, b = 1, \cdots, B\}}{B} . \tag{3.8}$$

3.2.4. Multi-level analysis for regulatory module identification

Assuming that most genes involved in a regulatory module are co-expressed under a given condition, we can use a clustering method to form the gene set for regression analysis. However, simple gene clustering based on gene expression data alone often results in many false-positives for gene module identification. In addition, motif information is noisy and incomplete due to the current status of limited biological knowledge. Thus, false-positives would be included based on a fixed gene set and available motif information. To reduce the false-positives, we developed a multi-level analysis strategy to search for regulatory modules showing significance consistently from coarse level to fine levels. With this strategy, a condition-specific regulatory module and its enriched motifs will appear increasingly significant in finer levels, as the irrelevant genes are gradually eliminated. Figure 3.2 shows an illustration of the multi-level strategy. Technically, a multi-level gene clustering procedure, such as self-organizing map (SOM) clustering [54], is used to form the gene clusters to gradually reduce the irrelevant genes for multi-level analysis. The multi-level analysis strategy, incorporating the two-stage SVR approach described previously, is the backbone of the ml-SVR approach proposed here for reliable regulatory module identification. The final ml-SVR procedure is illustrated in Figure 3.2, which can also be summarized as follows:

(1) Set cluster number $c = 1$ and cluster level $l = 1$. For all possible enriched motif sets, calculate their p-values on current gene set G through the two-stage SVR analysis and significance analysis described in Sections 3.2.2 and 3.2.3.

(2) Increment c by 1 and l by 1. Cluster the gene population into c clusters, denoted as $\{G_1^l, G_2^l, \cdots, G_c^l\}$.

(3) For each gene cluster, calculate p-values for all possible enriched motif sets by the two-stage SVR analysis and significance analysis (Sections 3.2.2 and 3.2.3).

(4) Repeat Steps (2) and (3) until the following stopping criterion is met, that is, the number of genes is less than a threshold t_0 for all gene clusters.

(5) Let us use p_M^{lc} to denote the p-value of a candidate motif set M for cluster G_c^l at level l. Output the significantly enriched motif sets if they satisfy $\min_c(p_M^{lc}) < p_0^l, \forall l$, where p_0^l is the threshold of p-value at each level l. Assign the final weighted average p-value as $p_M = \sum_l {}^{l}\!/_{\Delta} \min(p_M^{lc}),\ \ \Delta = 1 + 2 + \cdots + L,$ L is the total number of levels.

(6) Use a voting scheme to determine the gene members of a regulatory module with the enriched motif set M; the voting scheme is described as follows:

a. Initialize a gene weight vector **w** as 0;

b. Update **w** by the following equation:

$$\forall l, c, \quad \mathbf{w}_{G_c^l} = \mathbf{w}_{G_c^l} + \sum_{m \in M} \mathbf{X}_{G_c^l m}, \quad if \ p_M^{lc} < p_0^l \ .$$

(7) Finally, the genes whose weights are greater than a threshold w_0 are chosen as the members of a corresponding regulatory module. In our implementation, we set w_0 as the mean of **w** plus one standard deviation.

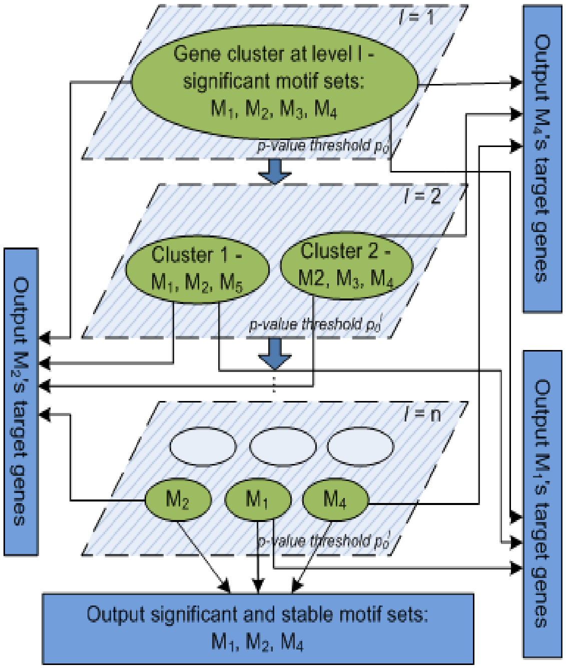

l = 1
Gene cluster at level l - significant motif sets: M_1, M_2, M_3, M_4
p-value threshold p_0
l = 2
Cluster 1 - M_1, M_2, M_5
Cluster 2 - M2, M_3, M_4
p-value threshold p_0^l
l = n
M_2
M_1
M_4
p-value threshold p_0^l
Output M_4's target genes
Output M_2's target genes
Output M_1's target genes
Output significant and stable motif sets: M_1, M_2, M_4

Figure 3.2. An illustration of the multi-level strategy. We can see from the figure that as the level moves from up to down, i.e., in a coarse-to-fine way, significant modules (e.g., M1, M2 and M4) will stably show up in fine levels, while non-significant modules (e.g., M3) will not show up at some fine levels.

We use the binding strength to represent how strong a gene is involved in a regulatory module. The larger the binding strength, the more likely a gene is the true target gene. Therefore, once the binding strengths are estimated, we regard genes with larger binding strengths (**w**) as the target genes. We assume that the binding strength (**w**) of target genes regulated by a transcription factor roughly follows an exponential distribution, which is consistent with a well-known biological constraint, i.e., a transcription factor only regulates a few target genes. The binding strength of target genes is likely located on the tail of the exponential distribution. In Figure 3.3, we show an example of the empirical distribution of **w** for one of the active transcription factors, AP-1, under estrogen-induced condition (see Section 3.3.4 for experimental description). For an exponential distribution, when the cut-off threshold is set as $\boldsymbol{w_0}$ = mean plus one standard deviation, the probability of $\mathbf{w} > \boldsymbol{w_0}$ is roughly 13.5%. In our experiments, this cut-off threshold, w_0 = mean plus one standard deviation, seems to give us a reasonable number of target genes for further study and biological validation. Should this threshold still results in too many target genes, it is advised to set a more stringent threshold such as $\boldsymbol{w_0}$ = mean plus two standard deviations; accordingly, the probability of $\mathbf{w} > \boldsymbol{w_0}$ is cut down to roughly 5%.

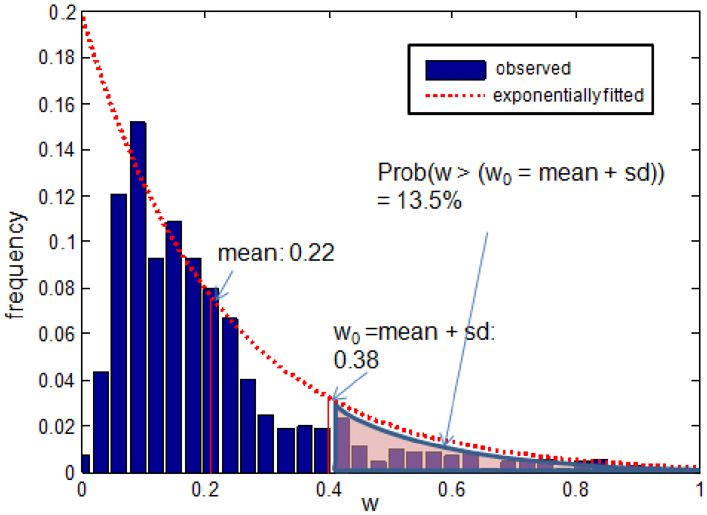


Figure 3.3. An example of the distribution of binding strength - binding strengths of AP-1's target genes in Cluster B under estrogen-induced condition. The distribution is fitted by an exponential function with the mean and threshold w0 (= mean plus one standard deviation) indicated. Note that the probability of w > w0 is roughly 13.5% for the exponential distribution.

3.2.5. Convergence of two-stage SVR

Suppose that there are G genes and T gene expression profiles, we represent microarray gene expression data as a matrix $\mathbf{Y}_{G\times T} = [y_{gt}],\ g = 1,\cdots,G;\ t = 1,\cdots,T$, where each element y_{gt} is the log-ratio of the expression level of gene g in sample t to that of the control sample. We also assume that there are M motifs on this gene set and the corresponding gene-motif binding matrix is $\mathbf{X}_{G\times M}=[x_{gm}], g=1,\cdots,G; m=1,\cdots,M$, where x_{gm} is the binding strength on motif m in the promoter region of gene g. The relationship between gene expression level and binding strength can be mathematically described by a linear model as follows:

$$\mathbf{Y}_{G\times T}=\mathbf{X}_{G\times M}\mathbf{A}_{M\times T}+\mathbf{N} \ , \tag{3.9}$$

where $\mathbf{A}_{M\times T}=[a_{mt}], m=1,\cdots,M; t=1,\cdots,T$ is the TF activity matrix and $\mathbf{N}$ the noise matrix. Biologically, the model represents the log-ratio of gene expression levels expressed as a linear combination of log-ratios of transcription factor activities (TFAs) (denoted as a_{mt}) weighted by their binding strengths (i.e., x_{gm}) [63].

The two-stage SVR method is implemented as an iterative procedure, which updates matrices **A** and **X** alternately until converged. We first write the matrix **A** as follows:

$$\mathbf{A}=\begin{bmatrix}\mathbf{a}_{c,1} & \mathbf{a}_{c,2} & \cdots & \mathbf{a}_{c,T}\end{bmatrix}, \tag{3.10}$$

where $\mathbf{a}_{c,i}$ is the i^{th} column of **A**. Now define a column vector $\mathbf{a}_c$ by stacking the columns of **A** as follows:

$$\mathbf{a}_c=\begin{bmatrix}\mathbf{a}_{c,1}\\ \mathbf{a}_{c,2}\\ \vdots\\ \mathbf{a}_{c,T}\end{bmatrix}.$$

Similarly we can write the matrix **X** as follows:

$$\mathbf{x}_r=\begin{bmatrix}\mathbf{x}_{r,1}\\ \mathbf{x}_{r,2}\\ \vdots\\ \mathbf{x}_{r,G}\end{bmatrix},$$

where $\mathbf{x}_{r,i}$ is the i^{th} row vector of **X**.

Based on the objective function of SVR, the optimal **X** and **A** can be found by solving:

$$\min_{\mathbf{a}_c,\mathbf{x}_r}\frac{1}{2}\|\mathbf{a}_c\|^2+\frac{1}{2}\|\mathbf{x}_r\|^2+cL_\varepsilon(\mathbf{Y}), \tag{3.11}$$

where L_ε is the ε-insensitive loss function, which is used in SVR to ensure the existence of the global minimum and a high tolerance to noise. It is defined by

$$L_\varepsilon(\mathbf{Y}) = \sum_{g,t} L_\varepsilon(y_{gt})$$

$$L_\varepsilon(y_{gt}) = \begin{cases} 0, & \text{if } \left|y_{gt} - \hat{y}_{gt}\right| < \varepsilon \\ \left|y_{gt} - \hat{y}_{gt}\right| - \varepsilon, & \text{otherwise} \end{cases}, \tag{3.12}$$

where $\hat{y}_{gt}$ is the estimated value of expression log-ratio y_{gt} from SVR.

The final algorithm of our two-stage SVR approach can be summarized as follows:

(1) Estimate $\mathbf{A}^{k-1}$ using $\mathbf{X}^{k-1}$ and $\mathbf{Y}$:

For each column vector $\mathbf{y}_t$ in matrix $\mathbf{Y}$, regress $\mathbf{y}_t$ against $\mathbf{X}^{k-1}$ based on $y_{gt} = f(\mathbf{x}_g^{k-1}) = \sum_{m=1}^{M} x_{gm}^{k-1} a_{mt}^{k-1}$; calculate regression coefficient a_{mt}^{k-1} using ε-insensitive SVR. The estimation error is:

$$e_1^{k-1} = \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}\left\|\mathbf{x}_r^{k-1}\right\|^2 + cL_\varepsilon(\mathbf{Y}). \tag{3.13}$$

(2) Update X^k using A^{k-1} and Y:

For each row vector $\mathbf{y}_g$ in matrix $\mathbf{Y}$, regress $\mathbf{y}_g$ against $\mathbf{A}^{k-1}$ based on $y_{gt} = f(\mathbf{a}_t^{k-1}) = \sum_{m=1}^{M} x'^{k}_{gm} a_{mt}^{k-1}$; calculate regression coefficient x'^{k}_{gm} using ε-insensitive SVR; update $\mathbf{X}$ by $\mathbf{X}^k = \mathbf{X}^{k-1} + \eta(\mathbf{X}'^k - \mathbf{X}^{k-1})$, where η is a parameter in the range of (0, 1).

Denote the estimation error based on $\mathbf{X}'^k$ as

$$e'^{k-1}_2 = \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}\left\|\mathbf{x}'^{k}_r\right\|^2 + cL_\varepsilon(\mathbf{Y}). \tag{3.14}$$

Since the estimation error can be guaranteed to be non-increasing, so $e'^{k-1}_2 \le e_1^{k-1}$. The estimation error for $\mathbf{X}^k$ is:

$$e_2^{k-1} = \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}\left\|\mathbf{x}_r^{k}\right\|^2 + cL_\varepsilon(\mathbf{Y}). \tag{3.15}$$

For equation (3.15), it can be derived that

$$\begin{aligned} e_2^{k-1} &= \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}\left\|\mathbf{x}_r^{k}\right\|^2 + cL_\varepsilon(\mathbf{Y}) = \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}\left\|(1-\eta)\mathbf{x}_r^{k-1} + \eta\mathbf{x}_r^{k}\right\|^2 + cL_\varepsilon(\mathbf{Y}) \\ &\le \frac{1}{2}\left\|\mathbf{a}_c^{k-1}\right\|^2 + \frac{1}{2}(1-\eta)\left\|\mathbf{x}_r^{k-1}\right\|^2 + \frac{1}{2}\eta\left\|\mathbf{x}_r^{k}\right\|^2 + cL_\varepsilon(\mathbf{Y}) \\ &= (1-\eta)e_1^{k-1} + \eta e'^{k-1}_2 \\ &\le e_1^{k-1} \end{aligned} \tag{3.16}$$

Repeat Step (1) and Step (2) until convergence. The estimation error is proved to be non-increasing in each step of the iterative optimization procedure. Therefore convergence to the optimal solution is guaranteed. It is also equivalent that the average correlation coefficient of TF activities between two successive iterations is non-increasing. During the experiment, the convergence criterion is defined as the average correlation coefficient of TF activities between two successive iterations is larger than a predefined threshold r_0.

3.3. Experiments and results

3.3.1. Reliability of knowledge information

Since the ml-SVR method relies on binding motif information as prior knowledge and usually it is quite noisy in real cases, we need to assess the influence of prior knowledge information on the recovered regulatory modules. To assess the reliability of knowledge information, we conducted a simulation experiment on ml-SVR for transcriptional network identification. Based on yeast ChIP-on-chip binding strength data, we randomly sampled 30 TFs and their target genes (p-value < 0.01) as the ground truth binding information. We generated gene expression profiles using a linear model Y = **XA,** where Y is the gene expression profile, **A** is the transcription factor activity and X is the binding strength. **A** is randomly sampled from a normal distribution. We purposely controlled the percentage of correct knowledge in the binding information and varied the percentage of knowledge from 0 to 100%. The ml-SVR algorithm was applied to identify transcription factors and target genes. We repeated the experiment 100 times at each percentage and reported the mean and standard deviation. Furthermore, to assess the influence of prior knowledge to the TRN identification, we conducted a baseline experiment using self-organizing maps (SOM) algorithm based on expression profiles alone. Comparing with many expression-based TRN identification methods [38, 84, 85], clustering methods have fewer assumption on the underlying TRN model, in particular, with no prior binding information needed; therefore it is selected to conduct the initial baseline experiment. However, the performance could be improved if we use more sophisticated and appropriate models (such as independent component models) for the baseline study. For this study, we grouped the entire gene population into 30 clusters for the identification of TRNs. The average of expression profiles in each cluster was calculated to represent the underlying transcription factor activity and the correlation coefficient between individual gene expression profile and the average profile was regarded as the binding strength of the underlying transcription factor to the gene. Since we could not determine which cluster represented which specific transcription factor without prior knowledge, we calculated the Pearson's correlation coefficients between cluster centers and ground truth TFAs to help establish their correct correspondence; specifically, we regarded one cluster as correctly recovered transcription factor module if the correlation coefficient was larger than a threshold, which was 0.7 (correlation p-value = 0.01) in our experiment. We also repeated the experiment 100 times and reported the mean of AUC values as an overall performance measure.

Figure 3.4 shows the area under the curve (AUC) values for TF identification at different percentages of correct prior knowledge for the ml-SVR method and the SOM clustering method. From the figure we can see that the performance of TF identification is improved when there is more correct knowledge information included in the prior binding data. The ml-SVR method is better than the SOM method when more than 20% of correct prior knowledge is included. Specifically, it can achieve more than 80% accuracy when at least 40% prior knowledge is correct. It is noticed that when the binding information contains little correct information (<10%), the performance of TF identification is even worse than a random guess, which indicates that insufficient information may lead to inaccurate results.

Figure 3.5 shows the AUC values for target gene identification at different percentages of correct prior knowledge for both ml-SVR and SOM methods, respectively. Similarly, the AUC value of TF target identification is gradually increasing with the percentage of correct prior knowledge. The ml-SVR method is better than the SOM clustering method when more than 10% of correct prior knowledge is included. When little correct prior knowledge (<10%) is included, the performance is a little better than that from random guess. When more than 40% prior knowledge is correct, it can give more than 75% accuracy in terms of AUC value. We notice that the improvement of AUC value for target identification is not dramatically improved as compared with TF identification. One possible reason is that one gene could be regulated by multiple transcription factors, but the ml-SVR method identified target genes for each individual TF, i.e., the joint effects were not considered for those genes.

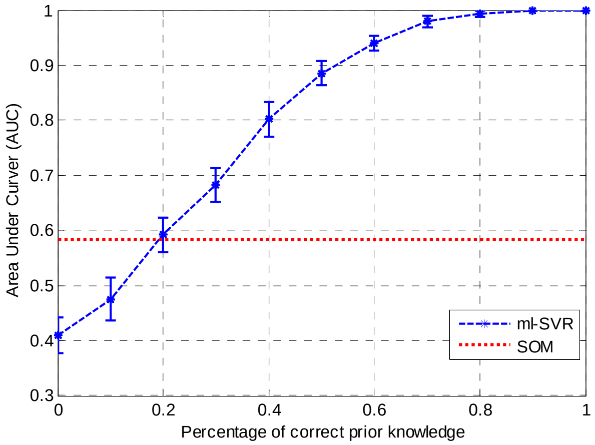


Figure 3.4. AUC values for TF identification at different percentage of correct prior knowledge.

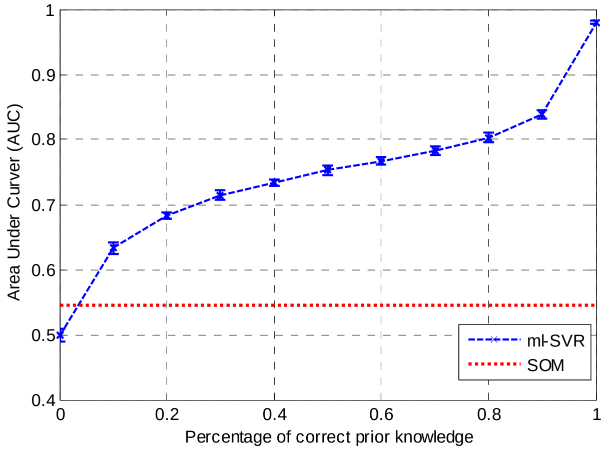


Figure 3.5. AUC values for TF targets identification at different percentage of correct prior knowledge.

3.3.2. Simulation data

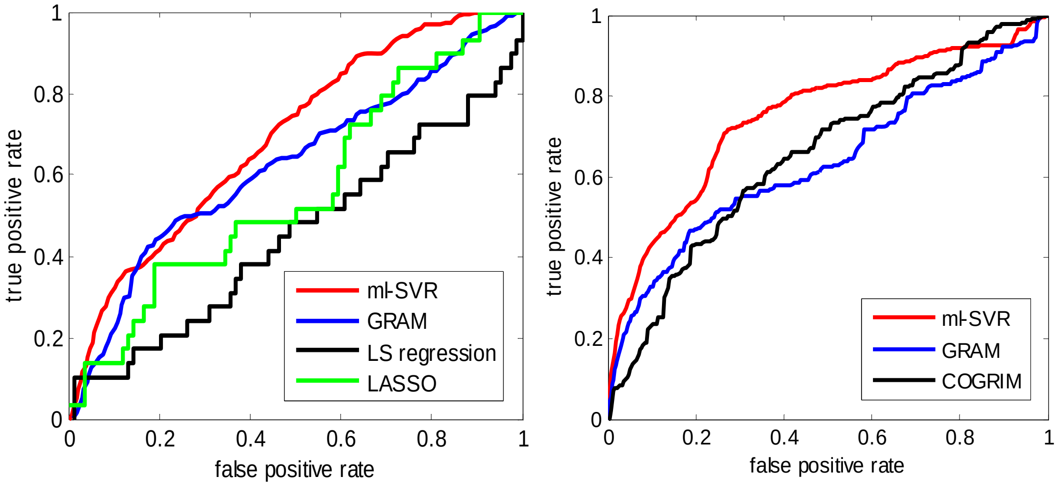
We first tested our method on a synthetic yeast microarray data set. The microarray data set was simulated using the network generator software SynTReN [106], where network topologies are generated from yeast regulatory networks using a neighbor addition strategy. The network consists of 29 transcription factors and 260 target genes. The mRNA expression profiles were generated for 260 genes at 50 different conditions based on the network. In our algorithm, we use a ChIP-on-chip experiment [89] as our binding information that includes the binding p-values of 113 regulators to all genes. The purpose of this study is to first identify true regulators and then their downstream target genes.

We transformed the binding p-values to binding strength by taking the negative logarithm values based on 10. We used the MATLAB SVM toolbox [16] to implement the ε-insensitive linear SVR and SOM clustering algorithm [54] to form multi-level gene clusters. The parameters in the algorithm were empirically determined for this experiment. Specifically, we set threshold $t_0 = 30$ (see Step (4) in the multi-level analysis procedure) and p-value threshold for each level $p_0^l = \min(1, 0.01 + 0.001 \times (L - l))$. Since we know the underlying transcriptional network from which the microarray data set was generated, we can use the Receiver Operating Characteristic (ROC) curve [71] and the area under the curve (AUC) to measure the test accuracy for the identification method. ROC curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR). AUC is an important performance measure that provides an overall measure of accuracy for the test. We rank the TFs in terms of their weighted average p-values across all levels. Since the clustering method (SOM) generated slightly different results depending on its random initializations, we repeated the entire procedure ten times with different initializations in search of more reliable results. The significant motif sets and their regulatory modules were determined according to their average values of ten different initializations. We also compared k-means [53] clustering results with SOM results after multiple initializations and found the overlap rate is greater than 90% in terms of clustering membership, which generates similar results for regulatory module identification. Therefore we only present the results based on SOM clustering method.

To evaluate our proposed ml-SVR approach, we compared its performance with similar methods including LS-regression [95], LASSO [107], GRAM [26] and COGRIM [96]. Among these four existing methods, only GRAM can simultaneously identify significant regulators and target genes. LS-regression and LASSO can only identify significant regulators with known target genes by assuming the binding information is known from ChIP-on-chip data. COGRIM is derived from a Bayesian hierarchical model, which assumes the transcription factors (TFs) and their activities are known so as to infer new target genes based on binding information. As a common practice [38] but faulty [63], mRNA expression level of each TF is often used to approximate the TF activity for COGRIM. Therefore, in this comparison study we compared ml-SVR with GRAM, LS-regression and LASSO for transcription factor identification, while we compared ml-SVR with GRAM and COGRIM for target gene identification.

Figure 3.6(a) shows the ROC curves of transcription factor identification for ml-SVR, GRAM, LS regression and LASSO, respectively. From the figure we can see that ml-SVR outperforms GRAM, LS-regression and LASSO methods in identifying significant transcription factors. The mean AUC value of ml-SVR is 0.6912 (with a standard deviation of 0.0196), which is greater than the AUC values of GRAM (0.6245), LS-regression (0.5530) and LASSO (0.5620). It should be noted that in this comparison experiment, the overall performances of all three methods are relatively low; this is indeed a relatively difficult case since some non-linear relationships between TFs and target genes were included by SynTReN in the simulation data. Nevertheless, the false positive rate (FPR) is much reduced by ml-SVR as compared to GRAM, LS-regression and LASSO. When the true positive rate (TPR) is fixed at 80%, the FPR for ml-SVR is 55.48% while 74.64% for GRAM, 94.05% for LS-regression and 71.42% for LASSO, showing a substantial improvement in FPR reduction.

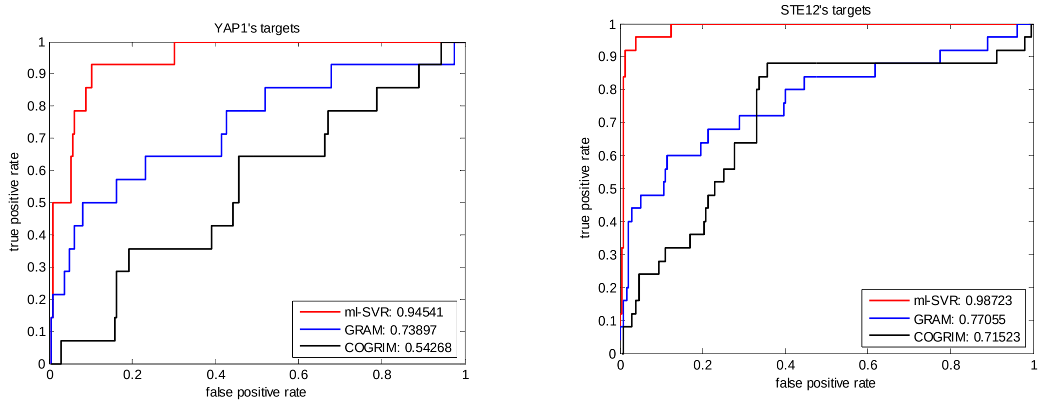
For the 29 known transcription factors, we compared the performance of target gene identification for ml-SVR, GRAM and COGRIM. Figure 3.6(b) shows the average of ROC curves of target gene identification for all TFs using ml-SVR, COGRIM and GRAM, respectively. The ml-SVR approach gave us the best performance with a mean AUC value of 0.7358 (and a standard deviation of 0.0090). The performances of COGRIM and GRAM are similar with the AUC values of 0.6434 and 0.6438, respectively, which are much lower than that of ml-SVR. Also seen from Figure 3.6(b), the FPR for ml-SVR is 42.12% given TPR = 80%, which shows a reduction of ~25% when compared to 68.79% for GRAM and 66.04% for COGRIM. This comparison result demonstrates the advantage of ml-SVR over other methods for identifying significant TFs and their target genes. For more ROC analysis results, please refer to Figure 3.7 to see the detailed performance of target gene identification for several individual transcription factors.



(a)

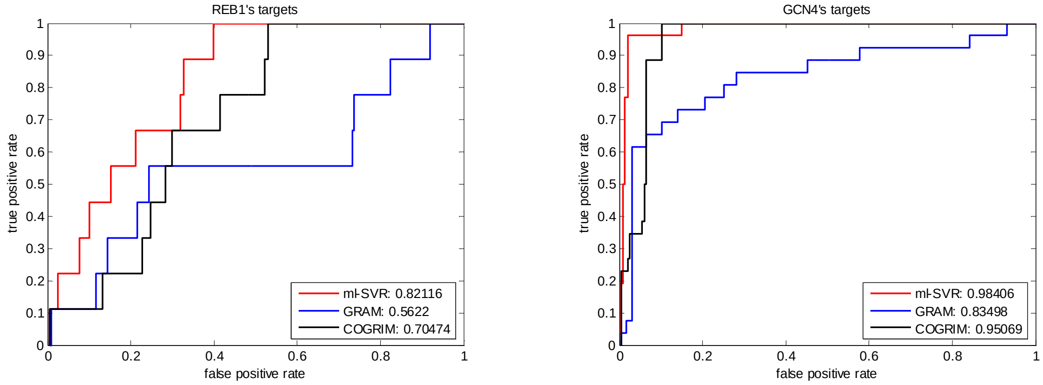
(b)

Figure 3.6. Comparison of Receiver Operator Characteristic (ROC) curves for ml-SVR and other methods on simulation data. (a) Transcription factor identification. (b) Target genes identification.



(a) YAP1

(b) STE12



(c) REB1

(d) GCN4

Figure 3.7. Comparison of Receiver Operator Characteristic (ROC) curves for ml-SVR, GRAM and COGRIM on target gene identification for some selected transcription factors. (a) YAP1. (b) STE12. (c) REB1. (d) GCN4. The values shown in the figure are the AUC values of ROC curves.

3.3.3. Yeast cell cycle data

We also applied the ml-SVR method to a yeast cell cycle microarray data set [108]. The yeast cell cycle data set comprises the expression of 6178 Open Reading Frames (ORFs) during the cell replication cycle in the budding yeast (Saccharomyces cerevisiae). The cell cycle process consists of four distinct phases: G1 phase, DNA synthesis (S) phase, G2 phase (also known as interphase) and mitosis (M) phase [108]. This microarray data set includes 77 samples collected with three different synchronization experimental conditions (alpha, cdc and elu). For the binding information, we used the ChIP-on-chip data from [89], which provides significance levels (p-values) of 113 transcription factors binding to their target genes. We took negative of logarithm (base 10) of p-values to convert the significance levels to binding strengths. After mapping these two data sets, we finally obtained 6,099 ORFs that have both expression measurements and binding information. Approximately 800 of these genes have been identified as cell cycle-regulated genes [108]. Among the 113 TFs, 19 regulators have been identified as cell cycle-related TFs. The goal of this study is to identify the cell cycle-related condition-specific transcription factors and their target genes.

To demonstrate the feasibility of applying ml-SVR to real microarray data, we compared the performances of ml-SVR, GRAM, LS-regression and LASSO for transcription factor (TF) identification using 19 known cell cycle-related regulators as the ground truth. The parameters in our algorithm are same as those in the simulation study.

Figure 3.8 shows the ROC curves of transcription factor identification by ml-SVR, GRAM, LS-regression and LASSO methods. The mean AUC value for ml-SVR is 0.9284 (with a standard deviation of 0.0127). The AUC values for GRAM, LS regression and LASSO methods are 0.6691, 0.8761 and 0.7704, respectively. The improvement of ml-SVR over GRAM is substantial in terms of false positive rate (FPR) reduction. Again, when the true positive rate (TPR) is fixed at 80%, the FPR for ml-SVR is 11.31% while it is 74.06% for GRAM, 18.68% for LS-regression and 52.18% for LASSO, showing a substantial improvement in FPR reduction. These results clearly show that ml-SVR outperforms the GRAM and LS-regression and LASSO methods for the identification of cell cycle-related transcription factors.

For target gene identification of all cell cycle-related TFs, since the ground truth target genes are not known for all TFs, we assessed their Gene Ontology (GO) functional enrichment as an alternative using software BiNGO [109]. The GO function enrichment score is defined as the negative logarithm of Benjamin-corrected p-value from an over-representative analysis in BiNGO. The average GO functional enrichment scores are 3.53 for ml-SVR, 3.41 for COGRIM and 2.86 for GRAM, which indicates that our method can identify more functionally coherent gene clusters associated with specific TFs.

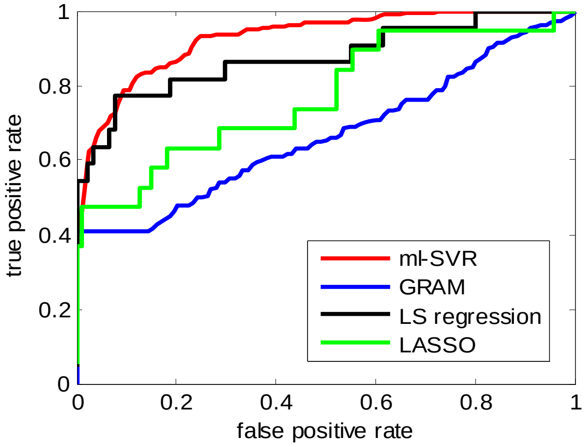


Figure 3.8. Receiver Operator Characteristic (ROC) curves of transcription factor identification for ml-SVR, GRAM and LS-regression on yeast cell cycle data.

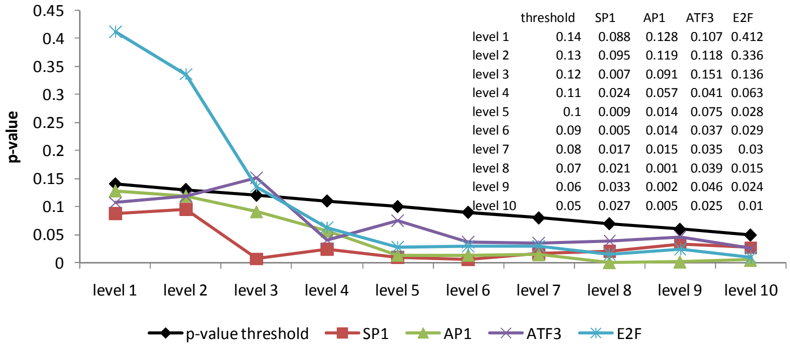
3.3.4. Breast cancer data

Estrogen-induced condition

A breast cancer cell line microarray data set [110] was used to identify condition-specific regulatory modules associated with estrogen signaling in breast cancer. Estrogen plays a significant role in breast cancer development and progression. The original profiling study was designed to examine how estrogen-induced gene expression patterns observed in vitro correlate with the expression patterns in breast tumors in vivo. Three estrogen-dependent breast cancer cell lines (MCF-7, T47D and BT-474) were treated with 17β-estradiol (E2) from 0 to 24 hours, and then profiled for gene expression using Affymetrix GeneChip Arrays. As reported in the paper [110], eight E2-induced gene clusters were formed and among them, the expression pattern in four clusters (i.e., Cluster A, B, C and D as denoted in [110]) clearly showed up-regulation along the time from early to late, which provides us an important starting point to study regulatory mechanisms related to estrogen signaling and action in breast cancer.

The ml-SVR approach was applied to identify significant regulatory networks. After mapping the expression data with binding motif strength data (Section 3.2.1), we obtained gene expression measurements with 39,407 probe sets and their corresponding binding strengths with 318 non-redundant motifs for the four up-regulated clusters. As a pre-processing step, we took the average of expression levels across all samples as the control value for each gene to calculate the log-ratio data. Parameters in the ml-SVR algorithm were empirically determined. Specifically, we set threshold $t_0 = 50$ (see Step (4) of the multi-level analysis) and the p-value threshold for each level $p_0^l = \min(1, 0.05 + 0.01 \times (L - l))$. We repeated the entire procedure ten times with different clustering initializations for a more reliable result. The significant motif sets and their regulatory modules were selected according to their average values of the ten different initializations.

The identified significant motifs in each cluster are shown in Table 3.1, along with their average p-values across all levels, number of regulated probe sets in the module and the description of the corresponding transcription factors. The significant motifs are defined by average p-values ≤ 0.05. Figure 3.9 shows an example of using the multi-level strategy to determine that SP1 and AP1 are significant TFs while ATF3 and E2F are not significant.



	threshold	SP1	AP1	ATF3	E2F
level 1	0.14	0.088	0.128	0.107	0.412
level 2	0.13	0.095	0.119	0.118	0.336
level 3	0.12	0.007	0.091	0.151	0.136
level 4	0.11	0.024	0.057	0.041	0.063
level 5	0.1	0.009	0.014	0.075	0.028
level 6	0.09	0.005	0.014	0.037	0.029
level 7	0.08	0.017	0.015	0.035	0.03
level 8	0.07	0.021	0.001	0.039	0.015
level 9	0.06	0.033	0.002	0.046	0.024
level 10	0.05	0.027	0.005	0.025	0.01

Figure 3.9. An example of the multi-level strategy for significant and stable motif discovery on the gene set of Cluster B in the estrogen-induced condition. SP1 and AP1 are significant since their p-values are less than thresholds across all levels, while ATF3 and E2F are not called significant since their p-values are larger than the thresholds at some levels. The p-values of transcription factors at each level are also shown in a table in the figure.

Table 3.1. Significant binding motifs in each cluster in breast cancer estrogen-induced cell line data set.

Cluster	Rank	Name	P-value	No. of probe IDs	Description
A	1	V$EGR1	0.008	43	Egr-1/Krox-24/NGFI-A immediate-early gene product
	2	V$AP1	0.009	79	Activating protein 1; AP1; Fos/Jun
	3	V$SP1	0.016	68	Specificity protein 1; Stimulating Protein1
	4	V$TCF11	0.020	60	TCF11/KCR-F1/Nrf homodimers
	5	V$MYCMAX	0.022	55	c-Myc:Max heterodimer
	6	V$E2F	0.036	64	E2F transcription factor
B	1	V$KROX	0.014	58	Krox-24/NGFI-A immediate-early gene product
	2	V$EGR	0.019	56	Early growth response protein
	3	V$LEF1	0.027	87	Lymphocyte enhancer binding factor 1
	4	V$SP1	0.033	102	Specificity protein 1; Stimulating Protein1
	5	V$TCF11	0.035	81	TCF11/KCR-F1/Nrf homodimers
	6	V$CREB	0.038	104	Cyclic AMP Responsive Element Binding factor
	7	V$AP1	0.049	120	Activating protein 1; AP1; Fos/Jun
C	1	V$AP1	0.033	66	Activating protein 1; AP1; Fos/Jun
	2	V$GATA	0.044	46	GATA binding protein
	3	V$PBX1	0.048	35	Homeobox protein PBX1
D	1	V$MYCMAX	0.018	49	c-Myc:Max heterodimer
	2	V$NFY	0.019	63	Nuclear factor Y (Y-box binding factor)
	3	V$USF	0.023	49	Upstream Stimulatory Factor 1
	4	V$P53	0.032	70	Tumor protein p53
	5	V$OCT1	0.038	57	Octamer-binding factor
	6	V$SP1	0.039	58	Specificity protein 1; Stimulating Protein1
	7	V$CREB	0.040	78	Cyclic AMP Responsive Element Binding factor

Among all listed motifs and their corresponding TFs, we found that several TFs are tightly related to estrogen signaling as reported in previous studies [111], to name just a few here, AP-1, SP-1 and CREB. From the table, we can also see that the significantly enriched motifs are different in each cluster, reflecting the condition-specific nature of transcriptional regulation. Since the target genes in Clusters A and B are up-regulated within 4 hours, we assigned the significantly enriched motifs in these two clusters to the early up-regulation condition. The target genes in Clusters C and D showed sustained induction by E2 at 8 hours and 12 to 24 hours, respectively. We assigned the significantly enriched motifs in Clusters C and D to the late up-regulation condition.

Figure 3.10 shows the significantly enriched motifs in two different conditions, i.e., early and late conditions. We can see that AP-1, SP-1, MYCMAX and CREB are significantly enriched in both early and late conditions, suggesting their important roles in estrogen signaling and action. AP-1 and SP-1 are known to form transcription factor complexes with estrogen receptor (ER) to regulate genes with the appropriate binding site(s); the transcription factor CREB is phosphorylated after the MAPK signaling pathway has been activated by 17β-estradiol and the phophorylation of CREB leads to the expression of genes that contain CRE binding motifs [111]. EGR, TCF11, E2F, KROX and LEF1 are only significantly enriched in the early up-regulation condition. Since many of their transcriptional functions are not known, we annotated their target genes biological function through GO analysis; their significant GO terms are related to 'ribosome biogenesis, 'RNA metabolism' and 'protein folding' (p-value < 0.01). This may suggest some potential functions of these binding transcription factors. For example, a change in the ability to fold proteins adequately induces the unfolded protein response, which we have previously implicated in antiestrogen resistance [112, 113]. Similarly, NFY, USF, P53, OCT1, GATA and PBX1 are only significantly enriched in the late up-regulation condition. Significant GO terms of their target genes include 'cell cycle', 'cell proliferation', 'mitosis' and 'DNA replication' (p-value < 0.01). Among them, previous studies [114] have shown that nuclear transcription factor Y (NFY) and p53 are related to cell cycle arrest; Octamer transcription factor-1 (Oct-1) is a member of the POU family of transcription factors and is involved in the transcriptional regulation of a variety of gene expression related to cell cycle regulation, development and hormonal signals [115];

Upstream stimulatory factor 1 (USF) is a transcription coactivator that plays a role in regulation of cell proliferation and associated with breast neoplasms [116]; Pre-B-cell leukemia homeobox 1 (PBX1) is a transcription activator that promotes transcription factor activity and cell growth, which may play an important role in Wnt receptor signaling [117].

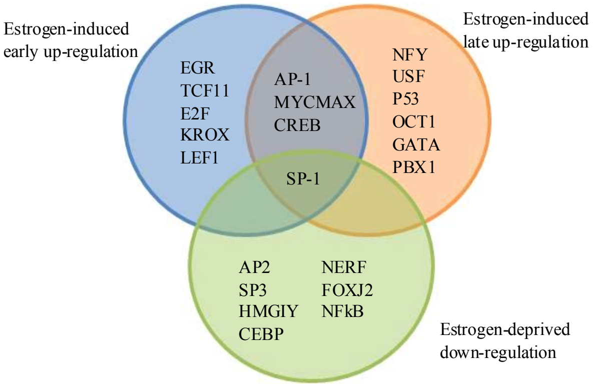

Estrogen-induced early up-regulation
EGR
TCF11
E2F
KROX
LEF1
AP-1
MYCMAX
CREB
Estrogen-induced late up-regulation
NFY
USF
P53
OCT1
GATA
PBX1
SP-1
AP2
SP3
HMGIY
CEBP
NERF
FOXJ2
NFkB
Estrogen-deprived down-regulation

Figure 3.10. Venn diagram of significantly enriched motifs in estrogen-induced and estrogen-deprived conditions.

Estrogen-deprived condition

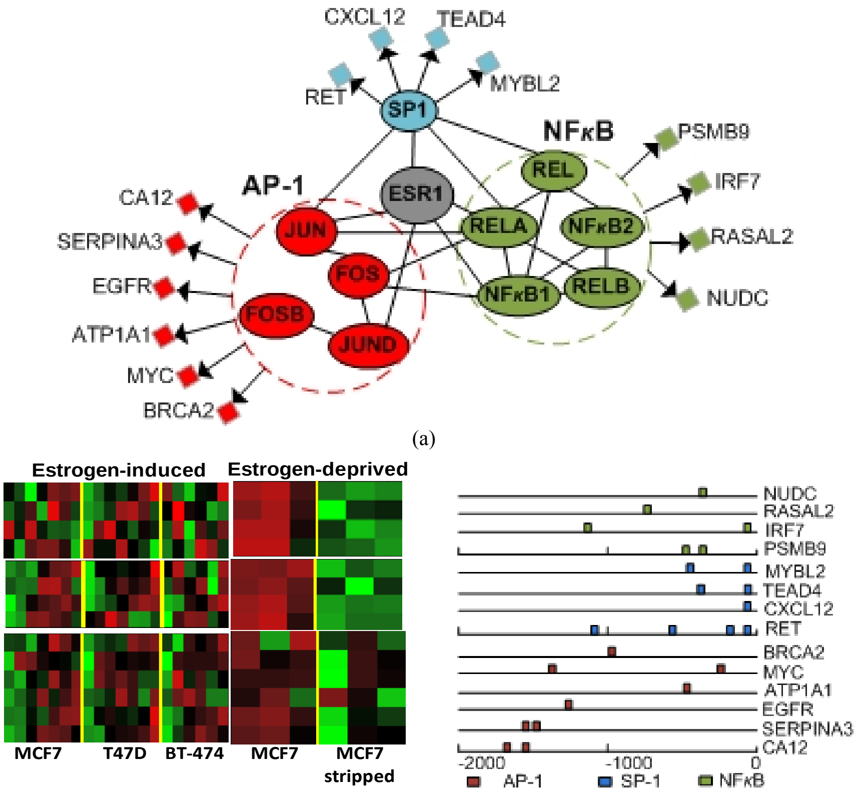
We previously derived a series of breast cancer variants that closely reflect clinical phenotypes of endocrine sensitive and resistant tumors [118, 119]. We selected two cell lines for this study: MCF-7 and MCF-7-stripped. MCF-7-stripped denotes estrogen-deprived MCF-7 human breast cancer cells, which were grown in the absence of estrogen for 96 hours. Three independent total RNA samples were extracted for each cell line (MCF-7 and MCF-7-stripped) and the samples were arrayed using Affymetrix GeneChip HG-U133A. Raw data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/; accession number: GSE 20700). We analyzed the enriched motifs and their targets for the genes significantly down-regulated in MCF-7-stripped cells as compared to MCF-7 cells. Down-regulated genes are identified by SAM analysis [14] with FDR < 0.05. Again, we applied the ml-SVR approach for this study to identify significant regulatory networks. After mapping the expression data with motif binding strength data, we obtained gene expression measurements with 1,120 probe sets and their corresponding binding strengths with 318 non-redundant motifs for the down-regulated gene set. We used the ml-SVR approach to identify the regulatory modules that are specifically enriched in estrogen deprivation condition. The parameters in the ml-SVR algorithm were empirically determined, which are same as the ones previously specified in estrogen induced condition.

Table 3.2 shows the identified significant motifs, their average p-values across all levels, number of probe sets in the module and the description of the corresponding transcription factors. As in the previous section, significant motifs were selected when their average p-values ≤ 0.05. From these motifs and their corresponding transcription factors, we found several transcription factors that have known associations with breast cancer, such as SP-1 and NFκB [111]. For the their target genes, the significant GO terms functions are related to cell cycle, intracellular membrane-bound, DNA replication etc (p-value < 0.01).

Table 3.2. Significant binding motifs in breast cancer estrogen-deprived cell line data set.

RANK	Name	P-value	No. of probe IDs	Description
1	V$AP2	0.008	113	General transcription factor IIIA
2	V$SP3	0.015	145	Stimulating Protein 3
3	V$HMGIY	0.018	102	High mobility group AT-hook protein 1
4	V$CEBP	0.019	146	CCAAT Enhancer Binding Protein
5	V$SP1	0.032	115	Specificity protein 1; Stimulating Protein1
6	V$NERF	0.041	131	E74-like factor 2 (ets domain transcription factor)
7	V$FOXJ2	0.046	125	Fork head homologous X
8	V$NFKB	0.048	86	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2

In Figure 3.10, we show a Venn diagram of significantly enriched motifs in both estrogen-induced and estrogen-deprived conditions. We can see that SP-1 is significantly enriched in both conditions, while AP-1 is only enriched in the estrogen-induced condition and NFκB is only enriched in the estrogen-deprived condition. A number of publications have reported that elevated AP-1 and NFκB activities are each associated with tamoxifen-resistant breast cancer [120-123]. We depicted these three transcription regulatory modules with their target genes in Figure 3.11(a). The interactions among the transcription factors ESR1, AP-1, SP-1and NFκB are extracted from Human Protein Reference Database [32]. Figure 3.11 (b) shows the binding sites for AP-1, SP-1 and NFκB in the promoter regions of their target genes, and their expression patterns in both conditions. In the next section, we provide a detailed description of the SP-1 network to establish its function role in estrogen signaling and action.



(a)

(b)

Figure 3.11. (a) Identified transcription regulatory modules of AP-1, SP-1 and NFκB in breast cancer study. (b) Gene expression patterns of target genes of AP-1, SP-1 and NFκB in estrogen-induced and estrogen-deprived conditions (left), and the binding sites of AP-1, SP-1 and NFκB on the promoter regions of their target genes (right).

SP-1 motifs are significantly enriched under both estrogen-induced and estrogen-deprived conditions, but the role of this transcription factor in estrogen and anti-estrogen signaling is less clear. Kim et al. [124] have reported that in breast cancer cells, E2 and anti-estrogens can both stimulate transcription on G/C-rich promoters via ER/SP-1 complexes. Table 3.3 shows the SP-1 target genes common to the estrogen-induced and estrogen-deprived conditions. Among these genes, some of them have been confirmed to be regulated by SP-1 in previous studies, and may have direct relevance to breast cancer, estrogen signaling, and antiestrogen resistance. For instance, it has been shown that the transcription factor SP1 can bind to the promoter of CXCL12 [125], and that estrogen-stimulated proliferation of ER+ T47D breast cancer cells can be blocked by a specific antagonist of the receptor for CXCL12 [126]. MYBL2 (B-MYB) is a ubiquitous protein required for mammalian cell growth, and a study by Sala, et al. [127] showed that B-MYB functions as a coactivator of SP1, binding to the 120bp B-MYB promoter fragment. Moreover, it has recently been shown that MYBL2 mRNA expression is significantly increased in breast cancer cells resistant to the tamoxifen analogue Toremifene [128]. Finally the RET proto-oncogene, more commonly associated with multiple endocrine neoplasia and medullary thyroid carcinoma, is also known to be transcriptionally regulated by SP1 [129]. Boulay et al. have reported that RET is induced by estrogens; RET signaling enhances the proliferative effect(s) of estrogen in ER+ MCF7 and T47D breast cancer cells, and RET is co-expressed with ER in primary breast tumors [130]. We have also observed RET mRNA overexpression in tamoxifen-resistant SUM44 breast cancer cells [121]. These results demonstrate that our method can successfully identify relevant transcription factor targets that play key, functional roles in estrogen signaling and action in breast cancer.

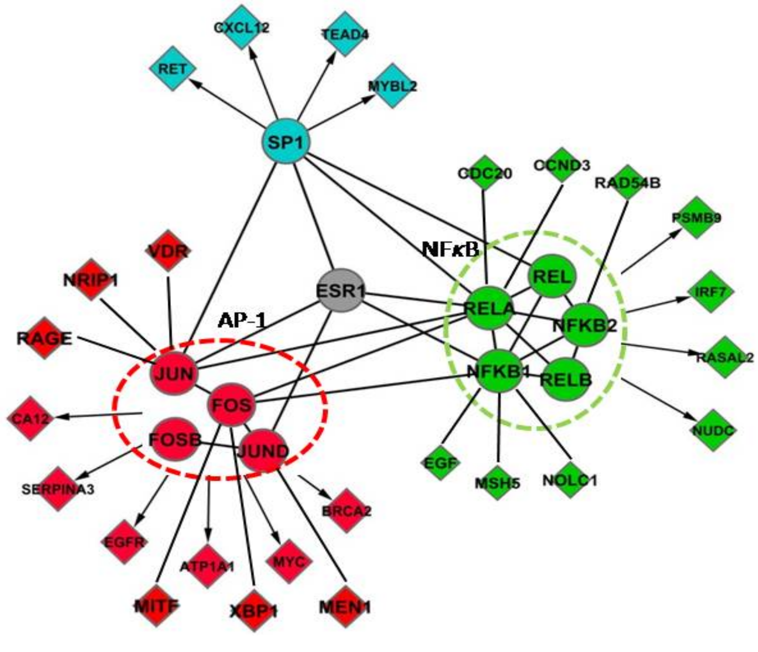


Figure 3.12. Identified transcription regulatory modules of AP-1, SP-1 and NFκB in the breast cancer study.

Table 3.3. Overlapped target genes of SP1 in estrogen-induced and estrogen-deprived conditions.

Probe Set ID	Gene Symbol	Gene Name
204507_S_AT	PPP3R1	PROTEIN PHOSPHATASE 3 (FORMERLY 2B), REGULATORY SUBUNIT B, 19KDA, ALPHA ISOFORM (CALCINEURIN B, TYPE I)
211421_S_AT	RET	RET PROTO-ONCOGENE (MULTIPLE ENDOCRINE NEOPLASIA AND MEDULLARY THYROID CARCINOMA 1, HIRSCHSPRUNG DISEASE)
201349_AT	SLC9A3R1	SOLUTE CARRIER FAMILY 9 (SODIUM/HYDROGEN EXCHANGER), MEMBER 3 REGULATOR 1
201819_AT	SCARB1	SCAVENGER RECEPTOR CLASS B, MEMBER 1
203666_AT	CXCL12	CHEMOKINE (C-X-C MOTIF) LIGAND 12 (STROMAL CELL-DERIVED FACTOR 1)
204498_S_AT	ADCY9	ADENYLATE CYCLASE 9
209687_AT	CXCL12	CHEMOKINE (C-X-C MOTIF) LIGAND 12 (STROMAL CELL-DERIVED FACTOR 1)
218657_AT	RAPGEFL1	RAP GUANINE NUCLEOTIDE EXCHANGE FACTOR (GEF)-LIKE 1
218670_AT	PUS1	PSEUDOURIDYLATE SYNTHASE 1
218944_AT	PYCRL	PYRROLINE-5-CARBOXYLATE REDUCTASE-LIKE
219990_AT	E2F8	E2F TRANSCRIPTION FACTOR 8
221223_X_AT	CISH	CYTOKINE INDUCIBLE SH2-CONTAINING PROTEIN
221987_S_AT	TSR1	TSR1, 20S RRNA ACCUMULATION, HOMOLOG (YEAST)
41037_AT	TEAD4	TEA DOMAIN FAMILY MEMBER 4
201000_AT	AARS	ALANYL-TRNA SYNTHETASE
201678_S_AT	C3ORF37	CHROMOSOME 3 OPEN READING FRAME 37
201710_AT	MYBL2	V-MYB MYELOBLASTOSIS VIRAL ONCOGENE HOMOLOG (AVIAN)-LIKE 2
202309_AT	MTHFD1	METHYLENETETRAHYDROFOLATE DEHYDROGENASE (NADP+ DEPENDENT) 1, METHENYLTETRAHYDROFOLATE CYCLOHYDROLASE, FORMYLTETRAHYDROFOLATE SYNTHETASE
203422_AT	POLD1	POLYMERASE (DNA DIRECTED), DELTA 1, CATALYTIC SUBUNIT 125KDA
205086_S_AT	NCAPH2	KLEISIN BETA
205240_AT	GPSM2	G-PROTEIN SIGNALLING MODULATOR 2 (AGS3-LIKE, C. ELEGANS)
208808_S_AT	HMGB2	HIGH-MOBILITY GROUP BOX 2
216299_S_AT	XRCC3	X-RAY REPAIR COMPLEMENTING DEFECTIVE REPAIR IN CHINESE HAMSTER CELLS 3

Figure 3.12 shows a detailed regulatory network for AP-1, SP-1 and NFκB. Different colors indicate different regulatory modules. Transcription factors are represented by ellipses and target genes are represented by diamonds. The AP-1 transcription factor family includes JUN, FOS, FOSB and JUND. The NFκB transcription factor family includes NFκB1, NFκB2, REL, RELA and RELB. The interactions among these transcription factors are extracted from Human Protein Reference Database (HPRD) [32], which are indicated by undirected edges. The regulations from transcription factors to target genes are indicated by directed edges in the figure. Besides directly regulated genes, in each regulatory module we also include target genes that have protein-protein interactions with the transcription factors (extracted

from HPRD), indicated by undirected edges in the module. The AP-1 module is significantly enriched in the estrogen-induced up-regulated condition, while the NFκB module shows significantly in the estrogen-deprived down-regulated condition. The SP-1 module is significantly enriched in both conditions. We show selected target genes for each module as supported by published studies from literature survey. For AP-1 target genes, some are already known to be regulated by AP-1. For instance, FOS protein is present in a complex that binds a negative regulator of MYC [131], which plays an important role in cellular proliferation and the genesis of diverse tumors. CA12, APT1A1, MYC and EGRF have all been identified as estrogen-induced/AP-1-dependent genes in [132], and each has a high frequency of AP-1 binding sites in their promoters. Kordula, et al. [133] have shown that SERPINA3 (ACT) contains an AP-1 binding site that contributes to the full responsiveness of the ACT gene to cytokines. For NF*κ*B target genes, an NF*κ*B site binds the p50 and p65 heterodimer and is required for the induction of IRF7 [134]. *In vitro* binding studies [135] have also confirmed the capacity of the NF*κ*B site to bind p50/p65 and p52/p65 heterodimers for upstream of PSMB9. NUDC and RASAL2 were identified as the candidate components of NFκB signaling pathway through physical and functional study in [136].

3.4. Discussion

Identification of transcription regulatory modules has become increasingly important to understand the molecular mechanisms associated with cancer. Previous methods [91-95] focused on how to model the relationship of transcription factor binding and gene expression levels, assuming either active transcription factors or target genes are known. However, it is a challenging problem in many cancer studies due to significant noise in data sources: inaccurate motif binding information, noisy gene expression data, and incomplete knowledge of the biological problem under study. The ml-SVR method is intended to address these problems and simultaneously identify significant transcription factors and their target genes through a multi-level strategy. SVR is utilized because its performance for combining binding motif information and gene expression data is robust in the presence of noise; note that it can also be potentially extended to model the nonlinear relationship between binding information and expression data through kernel functions. Clustering is used to group genes in multiple levels, in a coarse-to-fine way, to avoid hard split of the genes, which may be undesirable considering the noises.

There are several issues for further investigation. The method described here assumes that co-expressed genes should be co-regulated to some degree; hence, genes are clustered based on their expression profiles alone. Recently, Gong et al. [137] proposed to cluster genes based on their gene expression data and binding motif information together, which may provide more accurate gene clusters for analysis. Another important issue that needs to be addressed is how to determine an appropriate motif set for SVR fitting. In our experiment, we only focused on each individual transcription factor and their modules. However, finding the cooperative transcription factors is also important for many biological studies. Due to the large number of motifs under study (typically in a range of 50 to 500), it is not feasible to consider all possible motif combinations when the order of the motif set increases. In our recent work [138], we developed a stepwise forward greedy search strategy, using a modified loss function to find the co-operative motifs in a given gene set.

3.5. Conclusion

We have proposed a multi-level two-step SVR method to identify significant condition-specific regulatory networks. Binding motif information and gene expression data are integrated by support vector regression followed by significance analysis to find the active motif sets. A multi-level analysis strategy is further developed to help reduce false positives for reliable regulatory module identification. The simulation study and the experiment on yeast cell cycle data demonstrated the effectiveness of our method in identifying transcription factor and target genes.

4 Bootstrapping MRF-based Subnetwork Identification and Network-based Prediction from Microarray Data and Protein-protein Interaction Network

4.1. Introduction

As microarray technology makes it possible to measure the expression levels of thousands of genes simultaneously, biomarker identification has become one of the major goals of microarray data analysis, with which to detect differentially expressed genes across different types of tissue samples or samples obtained under different experimental conditions in this high-dimensional gene space. Given clinical outcomes, the problem can be formulated as a prediction problem that is designed to find informative genes to build a classification model with accurate prediction performance when future unknown samples are interrogated.

Many classification methods [71, 139, 140] have been developed, with filter- [14, 141, 142] or wrapper-based [143, 144] feature selection methods incorporated. These methods have demonstrated their initial effectiveness in finding differentially expressed genes. However, these computational methods are primarily designed based on microarray data alone, and the selected features/genes usually have little biological functional coherence related to a specific disease or cancer under study.

Recently many methods have been developed to identify significant gene sets or pathways involved in diseases or biological processes by incorporating some prior knowledge, with which to help understand the underlying biological mechanism. For example, gene set enrichment analysis or pathways enrichment analysis approaches [47-49] are proposed by using the membership information in functional gene clusters or pathways.

Besides the membership information in prior knowledge, most of the prior knowledge contains network or structure information, which could be represented by graphs, such as protein-protein interactions, protein-gene interactions or regulatory pathways. Based on the structure of PPI network, identification of active subnetworks, which are related to the underlying mechanisms governing the observed changes in gene expression [41], is becoming more and more important in systems biology. Traditional individual gene-based biomarker identification is extended to subnetwork marker identification, aiming to reveal more biologically relevant information by incorporating prior knowledge such as PPI network or pathway information. Consequently, it is of great importance to develop network-based classification and prediction schemes to achieve better reproducibility among different platform data sets.

4.1.1. Existing methods and their limitations

Many methods have been developed to search subnetworks from PPI data with significant changes in gene expression over different conditions [41, 42, 145, 146]. Among them, Chuang et al. [41] proposed a protein-protein network-based approach to identify markers of metastasis using gene expression profiles. The markers are not encoded as individual genes or proteins, but as subnetworks of interacting proteins within a larger human protein-protein interaction network. The resulting subnetworks not only provide models of the molecular mechanisms underlying metastasis, but also include several key proteins that cannot be easily detected through their differential expression alone. Ideker et al. [42] converted the p-value of gene expression levels between two phenotypes into z-score and aggregated z-scores in a subnetwork to form the network score. A search algorithm is then implemented to find the subnetwork with maximum network score. Although these methods have achieved some successes in identifying biologically relevant subnetworks, one of the limitations of the methods is that genes in a PPI network were treated independently when the network score was calculated, i.e., the dependency among the genes in a subnetwork was ignored during the analysis. It is well known that genes in a local subnetwork have functional relevance; therefore they should form a significant subnetwork even though not all of them have significantly differential gene expression. Another limitation is that many hub genes, which are biologically important and have many interactions in PPI network, often show little change in expression compared with their downstream genes. Therefore by picking up downstream rather than hub genes the resulting subnetwork may not reveal the underlying mechanism of disease.

With the identified significant subnetworks, an important issue to be considered is how to construct features on the subnetworks in order to build a classifier for further prediction. Generally speaking, there are two ways for feature construction. One is to calculate the subnetwork activity by aggregating the expression profiles of genes within the subnetwork [41]. Through this way, each subnetwork activity is regarded as a feature and the classifier is built on these features. Although using network activity as feature may offer some benefit to reduce the noise in gene expression data, the structure of subnetwork is totally omitted when building the classifier. Another way is to use each individual gene expression profile in the subnetwork as the input feature, but the structure of subnetwork is explicitly modeled in the classifier formulation [147, 148]. For example, Li et al. [147] introduced a network-constrained regularization procedure for linear regression analysis. In this method, the entire network is represented as a graph and its corresponding Laplacian matrix. The network-constrained penalty function penalizes the L1-norm of the regression coefficients but encourages the smoothness of the coefficients on a network. The effectiveness of this method, in a general regression framework, has been demonstrated in identifying the relevant genes and subnetworks as they are related to the phenotypes under study. Nevertheless, since there is no subnetwork identification procedure conducted before the classification in [147], the method overemphasizes on a global network structure rather than a local network structure. Consequently, it has a severe limitation to deal with large networks due to a high computational cost paid in the construction of Laplacian matrix.

4.1.2. Our proposed methods

In this chapter, we propose to develop a novel subnetwork identification method for network analysis on different phenotypes of microarray data. In Particular, we develop a bootstrapping Markov Random Field (BMRF)-based subnetwork identification method, which follows an MRF-MAP (maximum a posterior) framework, by integrating gene expression data and protein-protein interaction network. Note that Markov random filed model has been applied for network-based analysis to predict protein function using PPI data and achieved some successes [37, 149] due to its flexibility to represent different types of dependency in network. A simulated annealing search algorithm is implemented to avoid local maximum and then a bootstrapping scheme is developed to help identify confident subnetworks. The simulation experiments demonstrate the effectiveness of the proposed method and the comparison results show that our method outperforms several benchmark methods in identifying subnetworks and hub genes. We have further applied our method onto two types of breast cancer microarray data sets; the experimental results show that the proposed method not only can achieve good prediction performance, but also helps identify several important subnetworks associated with the development of breast cancer and drug resistance.

We also develop a network-constrained support vector machine (netSVM) algorithm for classification and prediction. Laplacian matrix of network is explicitly incorporated in the objective function to impose the smoothness of estimated coefficients along the network. We evaluated the effectiveness of netSVM on simulation data and then applied the method to the real breast cancer data. The experimental results demonstrate that our method can help improve classification performance compared with conventional SVM, but more importantly, it can identify significant subnetworks that might be related to the underlying mechanism associated with clinical outcomes.

The organization of this chapter is as follows. In Section 4.2, we propose a novel subnetwork identification method based on a Markov random field (MRF) framework. We derive the mathematical model to calculate network score of a subnetwork and then present a search algorithm based on simulated annealing. A bootstrapping scheme is further implemented to generate confidence measures for subnetworks. In Section 4.3, a network-constrained support vector machine algorithm is proposed and the solution of the problem is derived. Statistical test for the significance of genes in the subnetworks is developed. The simulation study for subnetwork identification and netSVM is shown in Section 4.4. In Section 4.5 we present the subnetwork identification results on two types of breast cancer data with or without drug treatment. The prediction results by network-based SVM are shown in terms of cross validation and independent test. Finally, discussion and conclusion are presented in Sections 4.6 and 4.7, respectively.

4.2. Bootstrapping MRF-based subnetwork identification

4.2.1. Assumption and framework

We present a new algorithm to identify subnetworks from mRNA expression data and PPI network based on an MRF-MAP framework. The underlying assumption in our model is that the significant score of one gene in a subnetwork depends not only on its own gene expression profile, but also on the profiles of its neighbors in PPI network. Figure 4.1 shows the framework of BMRF-based subnetwork identification method, which takes PPI network and gene expression profiles as input, searches for subnetworks with large MRF-based network scores and outputs significant subnetworks after confidence assessment.

Different from the average activity score in [41], we will use an MRF-based framework to derive a new network score for subnetwork identification, taking into account the dependency among the genes in a subnetwork. The goal of subnetwork identification is to find a connected subnetwork or clique that maximizes the likelihood of posterior probability, $\boldsymbol{P}(\mathbf{f}|\mathbf{z})$, where $\mathbf{f}$ is a random variable vector representing the underlying discriminative scores and $\mathbf{z}$ is a vector representing the observed discriminative scores of a subnetwork; this goal can be achieved by maximizing a network score of the subnetwork as described below. A search algorithm based on simulated annealing will be implemented to identify the subnetwork for each candidate 'seed' gene using protein-protein interaction (PPI) data and gene expression data, to avoid the local maxima problem that is inevitably existed in the greedy algorithm adopted in [41]. Finally, a bootstrapping scheme is implemented on the data set to generate confident subnetworks through credibility test.

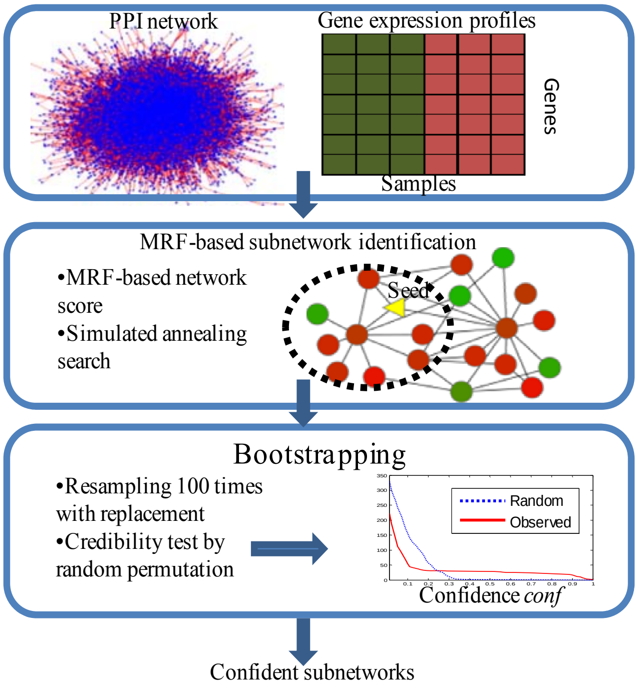


Figure 4.1. Framework of BMRF-based subnetwork identification from microarray gene expression profiles and PPI network.

4.2.2. Network score of a subnetwork

Let's first define a random variable vector $\mathbf{f} = \{f_1, \cdots, f_m\}$ to represent a set of discriminative scores of *m* genes (or proteins) between two phenotypes (e.g., 'Cancer' and 'Normal'). In the context of a PPI network, *S* represents the gene set of *m* genes in a network and N_i represents the connected neighbors of gene *i*. We define a pairwise clique C_2 on N_i and *S* as follows:

$$C_2 = \{\{i, i'\} \mid i' \in N_i, i \in S\}. \tag{4.1}$$

The random variable vector **f** is said to form a Markov random field on *S* with respect to N_i subject to the following conditions:

$$\begin{aligned} &P(\mathbf{f}) > 0, \ \forall \mathbf{f} \in \mathbf{F} \\ &P(f_i \mid f_{S-\{i\}}) = P(f_i \mid f_{N_i}) \end{aligned}. \tag{4.2}$$

The second criterion is the Markov property for a random field and states that the probability of a certain configuration at a site *i* is statistically independent of the configurations of all other $i \in S$ given the configurations at $i \in N_i$.

Specifying the joint probability $P(\mathbf{f})$ for a Markov random field is in general intractable. However, the equivalence between MRF and Gibbs distribution [150] provides us an alternative to specify $P(\mathbf{f})$ using Gibbs distribution. The possible configuration **f** of a set of random variable vector **F** obeys a Gibbs distribution if the joint distribution takes the form:

$$P(\mathbf{f}) = \frac{1}{Z} \times e^{-\frac{1}{T}U(\mathbf{f})}, \tag{4.3}$$

where Z is a normalizing constant given by:

$$Z = \sum_{\mathbf{f} \in \mathbf{F}} e^{-\frac{1}{T}U(\mathbf{f})}, \tag{4.4}$$

and U is given by:

$$U(\mathbf{f}) = \sum_{c \in C} V_c(\mathbf{f}). \tag{4.5}$$

U is an energy function that is determined by a sum of clique potentials $V_c(\mathbf{f})$ over all cliques. Clique potentials allow the modeling of knowledge (*a priori)* about the contextual interactions between genes at neighboring sites. Usually we assign 0 potentials to all cliques of size greater than 2. The energy $U(\mathbf{f})$ corresponds to the probability of that configuration. From Equation (4.3), we can see that lower energies correspond to more likely configurations. The parameter T is often referred to as 'temperature' and controls the sharpness of the distribution. Calculation of the partition function Z is intractable even for relatively small problems. However, calculation of Z is unnecessary since it is normalization constant.

Denote the observed discriminative scores of genes between two phenotypes as $\mathbf{z} = \{z_1, \cdots, z_m\}, z_i \in (-\infty, +\infty)$. Here, we define z_i as the z-score of its corresponding p-value p_i using $z_i = \Phi^{-1}(1 - p_i)$, where Φ^{-1} is the inverse normal cumulative density function (CDF). We assume here that the observed discriminative score is a result of the addition of independent zero mean Gaussian noise to the underlying discriminative score, i.e., $\mathbf{z} = \mathbf{f} + \mathbf{e}, \mathbf{e} \sim N(0,1)$. A possible estimate of the underlying discriminative score **f** is the MAP estimate $\hat{\mathbf{f}}$ that maximizes the likelihood of posterior probability (i.e., $\log P(\mathbf{f} \mid \mathbf{z})$); with the help of Bayes' rules and Gibbs distribution, it is equivalent to say that the MAP estimate $\hat{\mathbf{f}}$ minimizes the following posterior potential function:

$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}}(U(\mathbf{f}) + U(\mathbf{z} \mid \mathbf{f}))$. The first term in the posterior potential function is the prior potential given by:

$$U(\mathbf{f}) = \sum_{i \in S} V_1(f_i) + \sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'}) = \frac{-1}{m} \sum_{i \in S} f_i + \frac{\lambda}{k} \sum_{(i,i') \in E} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_{i'}}{\sqrt{d_{i'}}} \right)^2, \tag{4.6}$$

where d_i is the degree of gene i in the PPI network, k the number of interactions/edges and λ a trade-off parameter. As we can see, the first term in Equation (4.6) is the average discriminative score in a subnetwork; the second term in Equation (4.6) imposes the smoothness across the subnetwork while putting more weights on the genes with large degrees. Notice that the posterior potential function is normalized by the number of genes and the number of edges in the subnetwork, hence, independent of the subnetwork size.

The second term in the posterior potential function is the likelihood potential given by:

$$U(\mathbf{z} \mid \mathbf{f}) = \frac{\gamma}{m} \sum_{i \in S} (z_i - f_i)^2 / 2. \tag{4.7}$$

where γ is a trade-off parameter. The likelihood potential gives the average square of difference between observed and underlying discriminative scores, given the assumption of Gaussian distribution of noise signal with 0 mean and 1 standard deviation.

Thus, we can define the subnetwork score as the negative posterior potential function that takes into account the dependency among the genes of a subnetwork, which, in the form of estimated discriminative scores, can be defined as follows:

$$NetScore(G) = -U(\hat{\mathbf{f}} \mid \mathbf{Z}) = \frac{1}{m} \sum_{i \in S} \hat{f}_i - \frac{\lambda}{k} \sum_{(i,i') \in E} \left(\frac{\hat{f}_i}{\sqrt{d_i}} - \frac{\hat{f}_{i'}}{\sqrt{d_{i'}}} \right)^2 - \frac{\gamma}{m} \sum_{i \in S} (z_i - \hat{f}_i)^2 / 2. \tag{4.8}$$

4.2.3. Properties of MAP estimator for random variable f

First we define the Laplacian (**L**) of G with the uvth element to be:

$$\mathbf{L}(u,v) = \begin{cases} 1 & \textit{if } u = v \textit{ and } d_u \neq 0 \\ -1/\sqrt{d_u d_v} & \textit{if } u \textit{ and } v \textit{ are adjacent} \\ 0 & \textit{otherwise} \end{cases}. \tag{4.9}$$

With the Laplacian matrix **L**, the prior potential in Equation (4.6) can be written in a matrix format as:

$$U(\mathbf{f}) = \frac{-1}{m}\mathbf{f}^T\mathbf{1} + \frac{\lambda}{k}\mathbf{f}^T\mathbf{L}\mathbf{f}\,, \tag{4.10}$$

where **1** is a $m\times 1$ vector where each element is 1.

Based on Equation (4.10) and Gibbs distribution, Equation (4.3), we can derive the expectation of random variable **f** as follows:

$$E(\mathbf{f}) = \frac{k}{2\lambda m}\mathbf{L}^{-1}\mathbf{1}\,. \tag{4.11}$$

Through the first derivation of Equation (4.8), we can obtain the close form of the maximum-a-posterior (MAP) estimator of **f**:

$$\hat{\mathbf{f}} = \left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma\mathbf{I}}{m}\right)^{-1}\left(\frac{\mathbf{1}+\gamma\mathbf{z}}{m}\right). \tag{4.12}$$

From Equation (4.12) we can see that when parameter *γ* goes to infinity, the MAP estimator approximately equals to the observed **z** score. When *γ* approximately equals to 0, the MAP estimator is almost determined by network structure alone. Otherwise, it is a tradeoff between network structure and observation. Similarly, when parameter *λ* goes to infinity, the MAP estimator is close to 0. When *λ* approximately equals to 0, the MAP estimator is totally dependent on the observed **z** score. Otherwise, it is a tradeoff between network structure and observation.

It can be proved that this MAP estimator is an unbiased estimator through following derivation:

$$\begin{aligned} E(\hat{\mathbf{f}} - \mathbf{f}) &= E\left(\left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma\mathbf{I}}{m}\right)^{-1}\left(\frac{\mathbf{1}+\gamma\mathbf{z}}{m}\right) - \mathbf{f}\right) \\ &= E\left(\left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma\mathbf{I}}{m}\right)^{-1}\left(\frac{\mathbf{1}+\gamma(\mathbf{f}+\mathbf{e})}{m}\right) - \mathbf{f}\right) \\ &= \left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma\mathbf{I}}{m}\right)^{-1}\left(\frac{\mathbf{1}+\gamma E(\mathbf{f})}{m}\right) - E(\mathbf{f}) \\ &= \left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma\mathbf{I}}{m}\right)^{-1}\left(\frac{\mathbf{1}+\gamma\dfrac{k}{2\lambda m}\mathbf{L}^{-1}\mathbf{1}}{m}\right) - \frac{k}{2\lambda m}\mathbf{L}^{-1}\mathbf{1} = \mathbf{0} \end{aligned} \quad . \tag{4.13}$$

Therefore we can calculate its Cramér-Rao Lower Bound (CRLB) for an unbiased estimator based on the information inequality. The Information inequality [151] tells us that the variance of the estimator $\hat{\mathbf{f}}$ is bounded by the inverse of Fisher Information Matrix (FIM) $I(\mathbf{f})$, whose ij^{th} element is given by:

$$\left[I(\mathbf{f})\right]_{ij} = -E\left\{\frac{\partial^2 \ln\left[p(\mathbf{z};\mathbf{f})\right]}{\partial f_i \partial f_j}\right\}. \tag{4.14}$$

Based on Equation (4.14), we can derive the FIM of **f** as follows:

$$I(\mathbf{f}) = \frac{2\lambda\mathbf{L}}{k} + \frac{\gamma I}{m}. \tag{4.15}$$

Thus the variance of estimator $\hat{\mathbf{f}}$ is bounded by the following inequality:

$$Var(\hat{\mathbf{f}}(\mathbf{z})) \geq \frac{1}{I(\mathbf{f})_{nn}} = \left(\frac{2\lambda\mathbf{L}}{k} + \frac{\gamma I}{m}\right)^{-1}_{nn}. \tag{4.16}$$

Assume there are M observations $\{\mathbf{z}^j; j=1, 2, \ldots, M\}$, we can easily derive the CRLB for estimator $\hat{\mathbf{f}}$ as follows:

$$Var(\hat{\mathbf{f}}(\mathbf{z})) \geq \frac{1}{I(\mathbf{f})_{nn}} = \frac{1}{M}\left(\frac{2\lambda\mathbf{L}}{kM} + \frac{\gamma I}{m}\right)^{-1}_{nn}. \tag{4.17}$$

From Equation (4.17), we can see that the variance of estimation is decreasing when the number of observations increases. Given fixed number of observations, the variance of each node is related to the Laplacian matrix of given graph *G*. When all nodes are isolated or isotropic in the graph, the variances of the elements in vector **f** are same and equal to some constant. Otherwise, the variances of the elements in vector **f** are unequal and determined by the graph characteristics. Generally speaking, we observed in PPI network that the gene with large degree may have large variance, since its estimated discriminative score is determined and affected by more adjacent genes.

4.2.4. Search algorithm based on simulated annealing

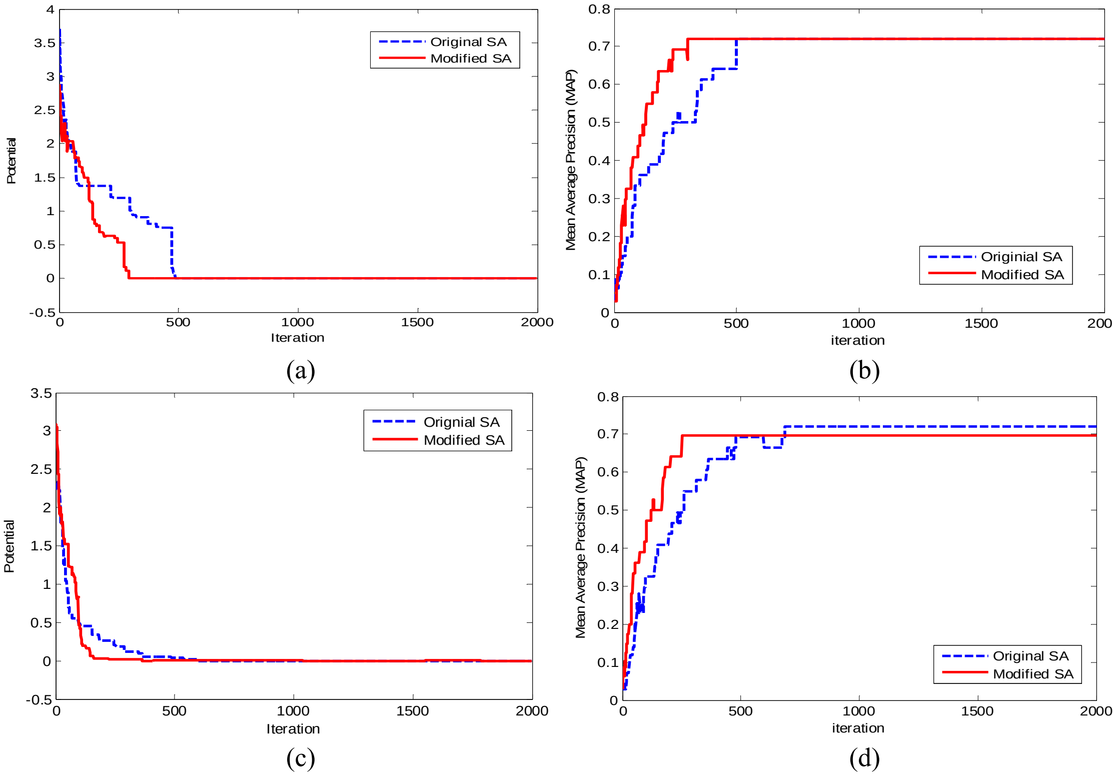
The network score in Equation (4.8) allows us to properly evaluate a given subnetwork, but finding the maximally scoring subnetwork in the full PPI network is an NP-hard problem. It is infeasible to perform exhaustive search since the searching space is exponentially increasing with the number of nodes in a graph. Greedy search algorithm is a fast solution but usually stuck in local maxima. Therefore heuristic methods are more suitable in our study because they can search very large spaces of candidate solutions and optimize a problem by iteratively improving a candidate solution with few or no assumptions. Among them, genetic algorithms [152], simulated annealing [42] tabu search [153, 154] and their related methods [155-157] are widely used for optimization problem. Genetic algorithms (GAs) mimic the process of natural evolution to find near-optimal solutions through inheritance, mutation, selection and crossover. Simulated annealing is a generalization of the Monte Carlo method for combinatorial optimization [158, 159]. It traverses the search space by generating neighboring solutions of a current solution and replaces the current solution by new solution with certain probability determined by the difference in quality and a temperature parameter. Tabu search is similar to simulated annealing but it generates multiple candidate solutions and selects the one with the lowest fitness. A tabu list is utilized to keep the partial solutions to avoid revisiting the existing solutions. Since crossover step is required in GAs, it often greatly increases the computational burden in checking the connectivity of subnetwork for each iteration. Furthermore previous studies [160, 161] have shown that GAs are not effective for problems involving a great number of variables of roughly equal influence compared with simulated annealing. Tabu search requires addition memory structure and defining the length of the tabu list is a tradeoff between the efficiency and computational complexity. Considering the flexibility of a search algorithm and its associated computational complexity, we implement our search algorithm based on simulated annealing to avoid the local maxima problem [42] for this study.

The computational complexity of simulated annealing (SA) depends on several parameters, such as diameter of the search graph, the candidate generator procedure, the acceptance probability function and the annealing schedule. It has been proven [162] that for any given finite problem, simulated annealing search can converge to the global optimal solution as the annealing schedule is extended. However, it is not practically useful since the iteration time required to ensure convergence will usually exceed the time required for exhaustive search of the solution space. Therefore we compromise the optimal solution in order to reduce the computational complexity; the sub-optimal solution is acceptable in our problem since the most differentially expressed genes may not be the ones biologically meaningful. In the implementation, we reduce the computational complexity through several aspects: 1) reducing the search space to a local search, 2) generating more heuristic candidate genes and 3) terminating the searching procedure when the objective cost function is small enough. Since the solution space increases exponentially with the number of nodes in the graph, we restrict the search space for each seed gene within two jumps since we focus on localized subnetworks derived from seed genes. In the context of PPI network, usually more than 1,000 genes are included within two jumps.

The second modification for the search algorithm is to generate more heuristic candidates rather than randomly picked after several iterations since the current state is expected to have much lower energy than a random state. We use a weighted sampling to randomly sample a new gene to generate new state. The weight is positively correlated with gene's z-score if adding one node, and negatively correlated with gene's z-score if deleting one node. Through this way, it is more possible to include a differentially expressed gene or delete a non-differentially expressed gene to increase the network score. Although it may get trapped in local optimum, this modified approach is expected to be better than greedy search in terms of optimal/suboptimal solution and better than (totally) random search in terms of convergence time. We conducted a comparison experiment for modified SA and original SA on subnetwork identification using simulation data (see the simulation study in Session 4.4 for further details). Figure 4.2 shows the trajectories of convergence rate and mean average precision for two algorithms on two simulation data sets, respectively. From the figure we can see that the modified SA converges much faster than the original SA does. The modified SA can also achieve a comparable performance in terms of MAP as the original SA can (see Figure 4.2(b) for an example). However, in some cases it may result in a slightly degraded performance (indicating being stuck in the local optimum) as compared to the original SA (Figure 4.2(d)), but it is still acceptable considering the saving in computational time.

The third modification is to terminate the searching procedure if the objective cost function could not be reduced in certain number of successive iterations. For example, the potential shown in Figure 4.2(a, c) for both SA algorithms are not decreasing after 1,000 iterations. Therefore we may terminate the procedure earlier to save the computational time.

With these modifications, we can save approximately one third of computational time, which is a significant improvement considering the running time ranges from several hours (on the simulation data set) to several days (on the breast cancer data set) using the modified SA algorithm.



(a) (b) (c) (d)

Figure 4.2. Comparison of convergent rate and performance between modified SA and original SA on two different simulation data sets when w = 40. (a) Trajectories of posterior potentials along iterations on first data set. (b) Trajectories of mean average precision of identified subnetworks along iterations on first data set. (c) Trajectories of posterior potentials along iterations on second data set. (d) Trajectories of mean average precision of identified subnetworks along iterations on second data set.

Therefore the modified search algorithm for each seed gene can be described as follows:

INPUT: A PPI network, seed gene g_0 and observed significance scores over the PPI network.

OUTPUT: A subnetwork G seeded by g_0 that maximizes the network score.

(1).***Initialization***: set i = 0, G = $\{g_0\}$, $T_i = T_0$ and $K_{\max} = K_0$; set the network score $NetScore_i$ as the significance score of g_0; note that T_i is the temperature parameter in simulated annealing and $K_{\max}$ is the maximum iteration step.

```
(2). **Iteration:** while $T_i$>0 and $i < K_{\max}$, perform the following steps:
        a. $i = i + 1$;
        b. Obtain all connected genes from current subnetwork $G$ and assign the
           outer layer of genes in $G$ except $g_0$, as the candidate gene set;
        c. Sampling a gene from candidate gene set based on their significance score
           and toggle its state ('insert' or 'remove') to form a new subnetwork $G_{\text{new}}$;
        d. Calculate the network score $NetScore_i$ for subnetwork $G_{\text{new}}$;
        e. IF $NetScore_i > NetScore_{i-1}$ , update current subnetwork as new subnetwork,
           i.e., $G = G_{\text{new}}$;
           ELSE update current subnetwork as new subnetwork $G = G_{\text{new}}$ with a
           probability of $p = e^{(NetScore_i - NetScore_{i-1})/T_i}$ ;
        f. Update $T_i = 0.9T_{i-1}$.
        g. IF $NetScore$ keeps constant within 500 successive iterations, stop (2).
(3). Output subnetwork $G$ that is of the maximum network score.
```

4.2.5. Bootstrapping for confidence measures of selected genes in subnetworks

Due to the heterogeneity and different noises in microarray data, the reproducibility of identified subnetworks across different data sets is usually low. In order to obtain more reliable subnetworks, we have implemented non-parametric bootstrap methods to select the most confident genes in the subnetwork identified [163]. The underlying intuition is that we should be more confident on genes frequently included in the identified subnetworks when we perturb the data. In the non-parametric bootstrap we generate such perturbations by re-sampling with replacement from the given dataset. The confidence score can be computed from the bootstrapping results as follows [164]:

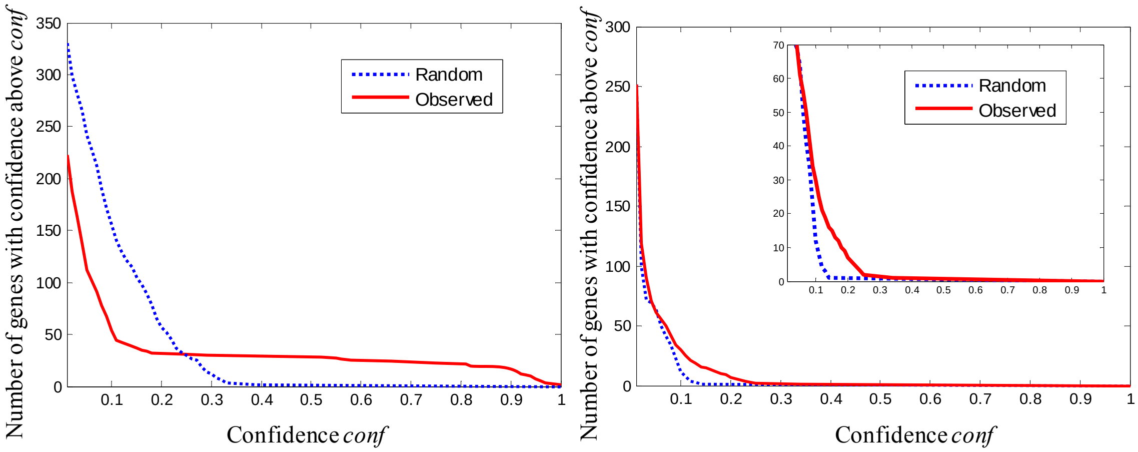
$$\text{conf}(gene_i) = \frac{1}{B}\sum_{b=1}^{B} f(gene_i^b) . \tag{4.18}$$

where $f(gene_i^b)$ is 1 if gene i is selected, and 0 otherwise; B is the number of bootstrap versions of the data. In this study, we generated 100 (i.e., B = 100) versions of the data set by sampling with replacement, and applied the MRF-based simulated annealing procedure to the data to obtain 100 subnetworks for each seed for confidence calculation. Furthermore, we tested the credibility of our confidence assessment by randomly permuting the phenotype labels of data samples [164]. Using 100 random permutations, we can obtain a baseline distribution of the confidence score. Then we can calculate the false discovery rate for a given confidence $conf_0$ in the observed data as follows:

$$\begin{aligned} \mathrm{FDR}(conf_0) &= \frac{\#\,\text{of expected false discoveries}}{\#\,\text{of discoveries}} \\ &= \frac{\#\text{ of genes with } conf_{\text{baseline}} \geq conf_0}{\#\text{ of genes with } conf_{observed} \geq conf_0}\ . \end{aligned} \tag{4.19}$$

The final subnetwork is composed by the genes that the corresponding false discovery rate is less than a predefined threshold, which is 0.05 in our experiments.

As an example, Figure 4.3(a) shows the confidence distribution of one of the subnetworks on a simulation data set (see Section 4.4 for further details). With a false discovery rate threshold of 0.05, we can obtain 25 genes in the subnetwork and the corresponding confidence score as 0.8. Figure 4.3(b) shows an example for the confidence measure of one of the subnetworks (seed gene: CDK25) on an in-house data set (see Section 4.5 for more details). Similarly, 16 genes are included in the subnetwork when false discovery rate is set as 0.05, with the corresponding confidence score as 0.15.



(a) (b)

Figure 4.3. Confidence distributions of one of the subnetworks for the observed data set (solid line) and the randomized data set (dotted line) on (a) simulation data set and (b) in-house data set. The *x*-axis represents the confidence threshold and the *y*-axis represents the number of genes with confidence equal or higher than the corresponding *x* -value. Inset in the right figure is the plot of tail distribution.

4.3. Network-based support vector machine for prediction

In most cases, only binary information of clinical outcomes is known, therefore a binary classification scheme is widely used to build a prediction model. In this section, we develop a network-constrained support vector machine (netSVM) in order to achieve better prediction performance and identify biologically meaningful genes by incorporating gene-gene interacting network. Specifically, we add a network constraint in the objective function of SVM to impose the smoothness of coefficient over the network. The network constraint is represented by the Laplacian matrix of gene-gene interacting network.

4.3.1. Support vector machine

Support vector machine (SVM) is a classification scheme that addresses the general case of nonlinear and non-separable classification tasks efficiently. The goal of an SVM is to find a hyperplane that maximizes the width of the margin between the classes and at the same time minimizes the empirical errors. Since the coefficients in weight vector correspond to real genes for linear SVM, we will focus on discussing the network-constrained SVM for linear case only in order to have a clear biological interpretation of those significant features (i.e., genes).

Given a training sample set ($\mathbf{x}_1$, y_1), …, ($\mathbf{x}_l$, y_l) with p features and l samples, where $\mathbf{x}_i \in R^p$ and $y_i \in \{-1,+1\}$, the SVM learning algorithm aims to find a linear function of the form $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x} + b$, with $\boldsymbol{\beta} \in R^p$ and $b \in R$ such that a data point $\mathbf{x}$ is assigned to a label +1 if $f(\mathbf{x}) > 0$, and a label -1 otherwise. The linear SVM classifier can be obtained by solving the following optimization problem:

$$\min_{\boldsymbol{\beta},b,\xi} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i(\boldsymbol{\beta} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \tag{4.20}$$

where the slack variable $\xi_i > 0$ denotes the difference of sample i to the required functional margin. The sum of ξ_i can be seen as an upper bound of the empirical risk. And the regularization constant $C > 0$ determines the trade-off between $\frac{1}{2}\|\boldsymbol{\beta}\|^2$ (the complexity term) and the sum of ξ_i.

By introducing non-negative Lagrangian multipliers α_i, the above optimization problem is equivalent to maximizing the dual Lagrangian function with respect to α_i in Equation (4.21):

$$\begin{aligned} L_D(\alpha) &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j, \\ s.t. \quad & \forall i \ 0 \le \alpha_i \le C \qquad . \\ & \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{4.21}$$

This is a quadratic programming problem and the solution to Equation (4.21) gives $\boldsymbol{\beta} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$ and b can be simply computed with any training point such that equality holds in problem (4.20).

4.3.2. Network-constrained SVM (netSVM)

Consider a gene network that is represented by a graph $G = (V, E, W)$, where V is a set of vertices that correspond to p genes, $E = \{u \sim v\}$ is a set of edges indicating that gene u and v are linked on the network and W is the weights of the edges. The degree of a vertex v is defined as $d_v = \sum_u w(u,v)$, where $w(u, v)$ indicates the weight of edge $u{\sim}v$. For this application, the weights could represent the probabilities of having edges between two vertices. Following Chung *et al.* [165], we define the Laplacian matrix **L** of G with the uv^{th} element to be:

$$\mathbf{L}(u,v) = \begin{cases} 1 - w(u,v)/d_u & \textit{if } u = v \textit{ and } d_u \ne 0 \\ -w(u,v)/\sqrt{d_u d_v} & \textit{if } u \textit{ and } v \textit{ are adjacent} \\ 0 & \textit{otherwise} \end{cases} . \tag{4.22}$$

This matrix is symmetric and non-negative definite and its corresponding eigenvalues or spectra reflect many properties of the graph as detailed in [165].

We define the network-constrained SVM given non-negative parameter λ as follows:

$$\min_{\beta,b,\xi} \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T L\boldsymbol{\beta} + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i(\boldsymbol{\beta}\cdot\mathbf{x}_i + b) \ge 1 - \xi_i, \xi_i \ge 0 \cdot \tag{4.23}$$

Compared to Equation (4.20), the only difference is that we add one more regularization term $\lambda\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$ into the objective function. We already know that the first regularization term is designed to maximize the width of the margin between two classes. We will thus focus on discussing the meaning of the second regularization term.

Note that **L** can be written as **L** = $\mathbf{SS}^T$, where **S** is the matrix whose rows are indexed by the vertices and whose columns are indexed by the edges of G such that each column (corresponding to an edge e = {*u*, *v*}) has an entry $\sqrt{w(u,v)}/\sqrt{d_u}$ in the row corresponding to *u*, an entry $-\sqrt{w(u,v)}/\sqrt{d_u}$ in the row corresponding to *v*, and zero entries elsewhere. Therefore we can see that $\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$ can be re-written as

$$\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta} = \sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v). \tag{4.24}$$

From this representation we can understand that the added regularization term $\lambda\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$ imposes the smoothness of parameters (coefficients) $\boldsymbol{\beta}$ over the network via penalizing the weighted sum of squares of the scaled difference of coefficients between neighboring vertices in the network.

It is worth noting that the network-constrain SVM is different from Laplacian SVM [166]. Network-constrained SVM imposes smoothness for weight vector $\boldsymbol{\beta}$, while Laplacian SVM imposes smoothness for Lagrangian multipliers *α*. In Laplacian SVM, it assumes that the data of each class, which follow a manifold and decision function must avoid passing through the manifold. In network-constrained SVM, the underlying assumption is that the genes highly connected in a network have synergistic effect and they should be considered together rather than individually.

Next, we will discuss how to solve the problem of Equation (4.20). Here we propose a simple algorithm by reducing it to a conventional SVM optimization problem. Since **L** is symmetric and semi-positive definite, Equation (4.23) can be represented as

$$\min_{\boldsymbol{\beta},b,\xi} \frac{1}{2}\boldsymbol{\beta}^T\mathbf{L}^*\boldsymbol{\beta} + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i(\boldsymbol{\beta}\cdot\mathbf{x}_i + b) \geq 1-\xi_i, \xi_i \geq 0, \tag{4.25}$$

where,

$$\begin{aligned} \mathbf{L}^* &= (\mathbf{I} + 2\lambda\mathbf{L}) = \mathbf{U\Gamma}\,\mathrm{U}^T \\ &= \mathbf{U\Gamma}^{1/2}\mathbf{\Gamma}^{1/2}\mathbf{U}^T = \mathbf{PP}^T, \quad \mathbf{P} = \mathbf{U\Gamma}^{1/2} \end{aligned}. \tag{4.26}$$

Further with the definition of $\boldsymbol{\beta}^* = \mathbf{P}^T\boldsymbol{\beta}$, the problem in (4.23) can be reduced to

$$\min_{\boldsymbol{\beta}^*,b,\xi} \frac{1}{2}\boldsymbol{\beta}^{*T}\boldsymbol{\beta}^* + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i\left(\boldsymbol{\beta}^* \cdot \mathbf{x}_i^* + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \tag{4.27}$$

where $\mathbf{x}_i^* = ((\mathbf{P}^T)^{-1})^T\mathbf{x}_i$. Therefore, this optimization problem could be solved by its corresponding dual problem similar with Equation (4.21). The solution gives $\boldsymbol{\beta}^* = \sum_{i=1}^{l}\alpha_i y_i \mathbf{x}_i^*$ and we can recover $\boldsymbol{\beta}$ through $\boldsymbol{\beta} = (\mathbf{P}^T)^{-1}\boldsymbol{\beta}^*$.Note that λ is a parameter and it can be optimized through cross validation in practice.

4.3.3. Significance analysis of subnetworks in netSVM

From the input network, we want to know which parts of the network are significantly contributing to the decision boundary for classification. As is shown in the Equation (4.23), the larger the absolute value of an element in coefficient vector **β**, the more important the corresponding gene is. Based on the clinical outcome information, we design a significance test to evaluate the significance of each gene in the network and then significant subnetworks can be determined by those genes whose p-values are less than some predefined threshold. For each gene *i* in the network, we take its absolute value of coefficient β_i as a summary statistic. To form a null distribution, we randomly permute training sample labels, and learn the coefficient vector $\boldsymbol{\beta}^0$ using network-constrained SVM on the training samples with permuted labels. The procedure is repeated *B* times, and all the corresponding absolute values of β_i^0 will be used to form the null distribution. The p-value of gene *i* can be calculated as follows:

$$p_i = \Pr_{H_0}(\left|\beta_i^0\right| > \left|\beta_i\right|) = \frac{\#\{b : \left|\beta_i^{0b}\right| > \left|\beta_i\right|, b = 1, \cdots, B\}}{B}. \tag{4.28}$$

4.4. Simulation study

4.4.1. Simulation data

We used two models to simulate the microarray gene expression data under two conditions considering the dependency of genes along a network. First, an MRF model is utilized to determine the states of genes (i.e., differentially expressed (DE) or non-differentially (equally) expressed (EE)) given the ground truth subnetwork. Then a Gamma-Gamma (GG) model [167] is utilized for modeling the gene expression levels in two conditions based on the states of genes.

We modified the MRF model in [149] to embed differentially expressed subnetwork/genes in the PPI network given a ground truth subnetwork. Let *S* is a binary vector indicating the differential expressed states of genes in a PPI network G, 0 means ‘equally expressed’ and 1 means ‘differentially expressed’;, and assume that the ground truth differential subnetwork is G_0, which means $S_{\{G0\}}$=1 and $S_{\{G\text{-}G0\}}$=0. We sample the gene state according to the following probability based on Markov random field model:

$$p_i(k\,|\,\cdot) \propto \exp(\gamma_k - \chi\mu_i(1-k))\,. \tag{4.29}$$

In the original model, μ_i(1-*k*) denotes the number of neighbors of gene *i* having state 1-*k*, *k* = 0, 1. γ_k and χ are the parameters predefined. In order to introduce different levels of false positives in the sampled differential subnetwork, we added one parameter to control the probability of keeping initial states of ground truth DE genes and background EE genes. Here we define μ_i(1-*k*) as a function of parameter ω as follows:

$$\mu_i(1-k) = \frac{\omega\cdot(1-S_i^{1-k}) + \sum_{j\in N_i}(1-S_j^{1-k})}{\omega + \sum_{j\in N_i}(S_j^{1-k}+S_j^{k})}, \text{ where } \quad S^1 = S \text{ and } S^0 = 1-S\,. \tag{4.30}$$

The larger ω *is*, the more consistent the simulated DE genes and ground truth genes are. Therefore we can vary ω to generate different simulation gene expression data sets with different levels of consistency.

Then, we simulated gene expression data *X* given *S* using a Gamma-Gamma model [167, 168]. In the GG model, the observed variable *x* (gene expression level) is a

Gamma distribution having shape parameter α>0 and scale parameter χ_g, with a mean value $\mu_g = \alpha\chi_g$. Its probability density function is:

$$p(x \mid \alpha, \chi_g) = \frac{x^{\alpha-1}\exp\{-x/\chi_g\}}{\chi_g^{\alpha}\Gamma(\alpha)}. \qquad (4.31)$$

In the above equation, the scale parameter χ_g has a Gamma distribution with shape parameter α_0 and scale parameter v. Given these three parameters, we can simulate the gene expression levels in two conditions with multiple replicates. In this experiment, we further assume that equally expressed gene has same expected mean for all samples and differentially expressed gene has different expected mean values in different conditions. To finally generate simulation data, we fist sampled the scale parameter χ_g based on Gamma distribution (α_0, v) and then sampled gene expression levels using parameters (α, χ_g) given the states of genes.

4.4.2. Simulation study for BMRF-based subnetwork identification

We conducted simulation studies on an estrogen receptor (ER) focused network that contains 365 genes and 1825 interactions, from which an ER-signaling pathway is considered as ground truth subnetwork including 35 genes and 89 interactions. To form an ER focused network for this simulation study, we collected 470 genes around ERα, ERβ, estrogen related receptors α (ESRRA) and γ (ESRRG) and aromatase from a set of bioinformatics resources including protein-protein interactions (PPIs), canonical pathways and protein-protein complex. We developed a strategy (as described below) for selecting the highest confidence genes to be part of this gene set by examining genes from various databases and different types of bioinformatics data (Figure 4.4). For the protein-protein interactions, STRING database [33] was mined for the first and second neighbors of the 5 aforementioned genes/proteins, the network seeds using the highest confidence score (>0.900), which resulting in 300 proteins. Next, a search for protein complexes with any of the 5 seeds was performed in the CORUM database [169]. 34 proteins were indentified from this search and were added to the gene set. We searched many canonical pathway sources and found Biocarta, Linnea and STKE to have ER pathways. Totaling all of the proteins from these three pathway sources, there were 232 unique genes classified as part of ER pathway. Finally, ER protein-protein interactions, ER pathway genes and ER complexes were grouped together to form 470 ER-related genes. After having overlaid these ER-related genes on Human Protein Reference Database (HPRD [32]), we finally have 365 genes and 1825 interactions for this simulation study. To form the ground truth subnetwork, 35 ER-signaling pathway genes were determined from pathway resources by at least 2 out of the 4 pathway sources to be part of the ER pathway, which resulted in 89 interactions in the HPRD database (Figure 4.5).

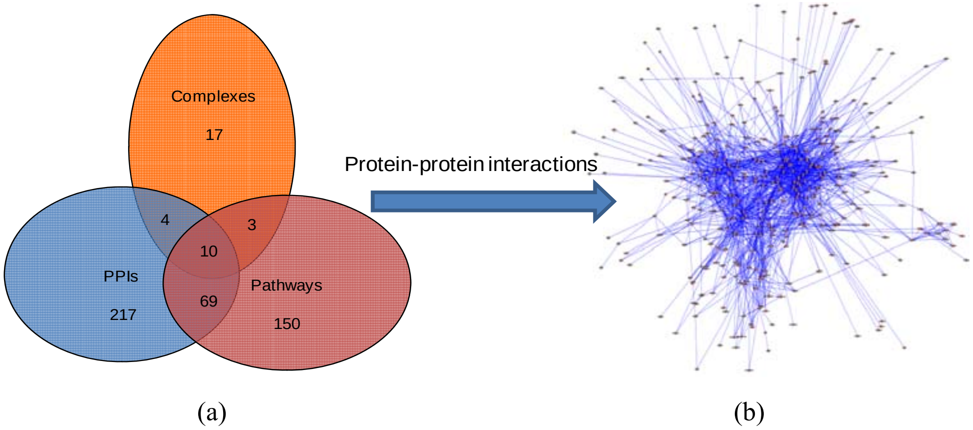


Figure 4.4. (a) Venn Diagram of PPIs, Pathway and Complexes to form ER-related genes. (b) ER-focused network on HPRD PPI network.

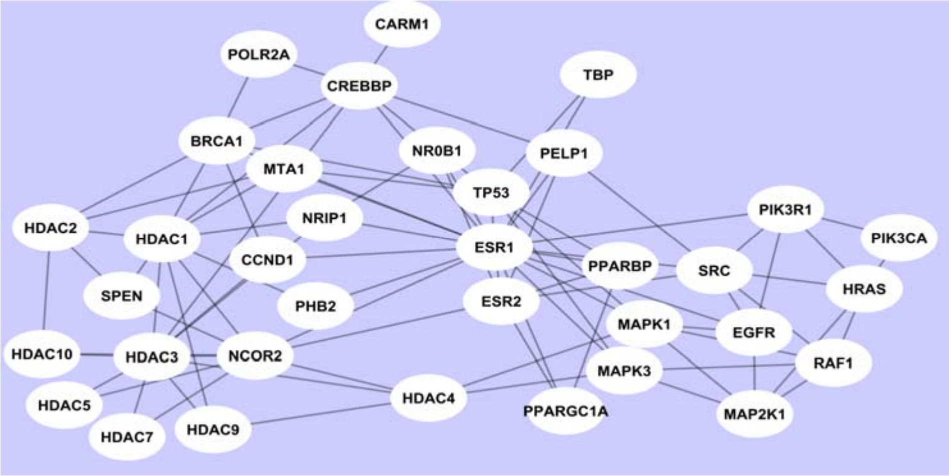


Figure 4.5. Ground truth ESR1 Subnetwork.

We first set the same parameters in GG model as the ones in Newton *et al.* (α = 10, α_0 = 0.9 and v = 0.5). We then chose w to be 0, 10, 20,…, 90, and for each parameter of w we generated 10 simulated gene expression data sets. For each data set, we used BMRF for subnetwork identification. As a comparison, we also performed jActiveModules as proposed in [42] and HEINZ in [146] on the same simulation data. jActiveModules is a subnetwork identification method that scores the subnetwork using the aggregated z-score derived from each individual gene's significance score (p-value). HEINZ (featuring a module scoring function) is a decomposition based method using mixture models, where integer-linear programming is used to find the optimal or suboptimal solution to the maximally-scoring subnetwork.

We used precision-recall curve [170] and percentage of identified hub genes (degree > 5) as the metrics to evaluate the performance. The precision and recall are defined as follows:

$$Precision = |\text{intersect } (S_{\text{recoved}}, S_{\text{ground}})|/|S_{\text{recoved}}|, \tag{4.32}$$

and

$$Recall = |intersect\ (S_{recoved}, S_{ground})|/|\ S_{ground}|, \tag{4.33}$$

where S_{recoved} indicates the recovered subnetwork after applying the MRF-based subnetwork identification method (or any other method in this comparison study) and S_{ground} indicates the ground truth subnetwork. To generate precision and recall curve, we ranked genes in the identified subnetwork according to their t-statistic p-values and then calculated precision and recall points by running down genes one by one on the ranked gene list. Mean Average Precision (MAP) of the precision-recall curve was also calculated in order to have an overall performance assessment.

Figure 4.6 shows the average precision-recall curves of identified subnetworks by BMRF-based method with different weights. From the figure we can see that MAP increases with weight *w*, which means that the BMRF-based method performs better when the genes in the subnetwork are more differentially expressed than background genes.

The performance comparisons for the BMRF-based method, jActiveModules and HEINZ are shown in Figure 4.7 and Figure 4.8. Figure 4.7 shows the mean average precision for the three methods at different weights. We also calculated the false positive rate of DE genes in the simulated gene expression data as listed in the figure. From the figure we can see that the BMRF-based method gives the best precision results consistently, and HEINZ performs a little better than jActiveModules. Figure 4.8 gives the comparison of percentage of identified hub genes for the three methods at different weights. We can see that the BMRF-based method outperforms other two methods and it can identify 80% of hub genes when false positive rate is at as high as 40%. The scoring function and simulated annealing search boost the probability of selecting hub genes in the BMRF-based method. jActiveModules is better than HEINZ because it used simulated annealing search, while integer-linear programming in HEINZ only focused on optimal solution for maximally scoring subnetworks, which may have ignored the hub genes when they are not significantly differentially expressed.

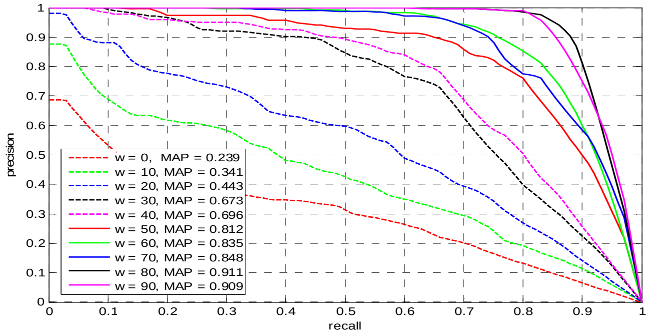


Figure 4.6. Precision-recall curve of identified subnetworks by BMRF-based method at different weights.

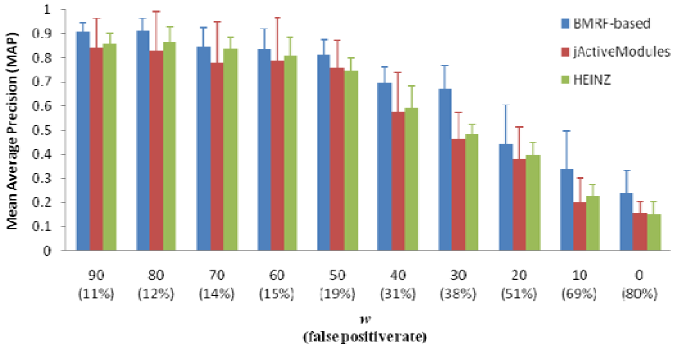


Figure 4.7. Comparison of MAP for BMRF-based method, jActiveModules and HEINZ at different weights.

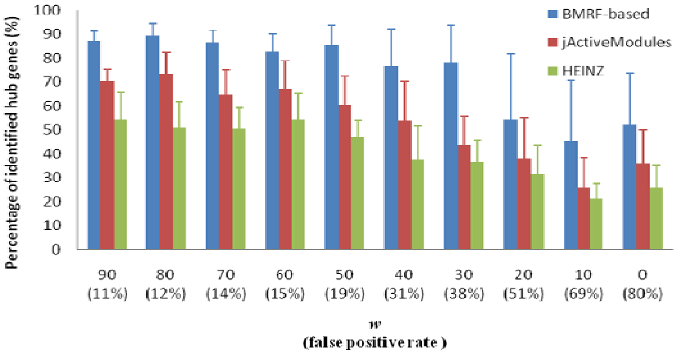


Figure 4.8. Comparison of percentage of identified hub genes for BMRF-based method, jActiveModules and HEINZ at different weights.

4.4.3. Simulation study for network-constrained SVM

We also conducted simulation studies on a breast cancer-related network that contains 584 genes and 2,280 interactions. The genes are either breast cancer related [171] or involved in estrogen signaling pathways collected from Ingenuity Pathway Analysis (Ingenuity® Systems; www.ingenuity.com). The interactions are extracted from the HPRD database [32]. Here the weights in the network are set as 1 if there are connections between two genes and 0 otherwise. We set the same parameters in the GG model as the ones in Newton *et al.* [167] ($\alpha = 10$, $\alpha_0 = 0.9$ and $\nu = 0.5$). We added different levels of noise and adjusted parameter w to control false positive rate in the sampled DE subnetworks to generate simulation data sets. For each scenario, we randomly generated 100 training and test data sets. Each data set included 100 training samples and 100 test samples.

We implemented both network-constrained SVM (netSVM) and conventional SVM for training and testing, respectively. A 10-fold cross validation was conducted on the training dataset to select the optimal value of parameter λ. We then computed the accuracy, sensitively and specificity for classification performance evaluation on the test data. In addition, we also assessed the performance of recovering ground truth subnetwork genes in classifier through receiver operating characteristic (ROC) analysis [71] for ranked gene list. Genes were ranked by their absolute coefficients in weight vector β. True positive rate and false positive rate were calculated in the ranked gene list and the Area under the ROC Curve (AUC) were computed for an overall performance evaluation.

We first fixed weight ($\omega = 10$) and added different levels of Gaussian noise to the simulated gene expression data. Figure 4.9 shows the AUC values of prediction performance on test data sets at different signal-to-noise ratios for netSVM and conventional SVM, respectively. From the figure we can see that when signal-to-noise ratio is relative high (>4db), both methods can achieve good prediction results. However, when signal-to-noise ratio is low, which reflects a more likely scenario in real microarray gene expression data, network-constrained SVM gives better classification performance compared to conventional SVM. Furthermore, AUC values for subnetwork identification are compared in Figure 4.10. We can see that netSVM outperforms SVM significantly in identifying the relevant ground truth subnetwork/genes.

Then we further evaluated the performance in uncovering underlying network/genes with different weights, with which to control the false positive rate of sampled subnetworks. With a fixed signal-to-noise ratio (SNR = 0 db), the prediction performance of two methods are comparable (results are not shown). However, the performance in identifying underlying subnetworks is becoming apparently different for the two methods when false positive rate increases, which is shown in Figure 4.11. From the figure we can see that network-constrained SVM outperforms conventional SVM significantly, especially when false positive rate of sampled subnetwork is high.

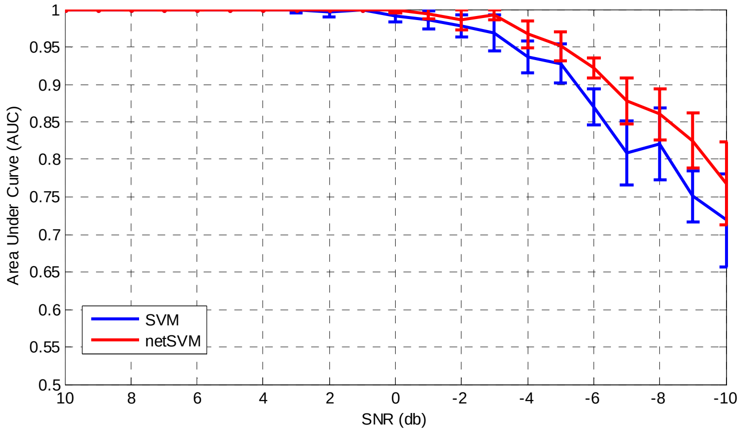


Figure 4.9. Comparison Results of prediction AUC values on simulation data sets with different signal-to-noise levels for network-constrained SVM and conventional SVM. AUC values are calculated based on 100 simulations, where standard deviations are shown in the error bar.

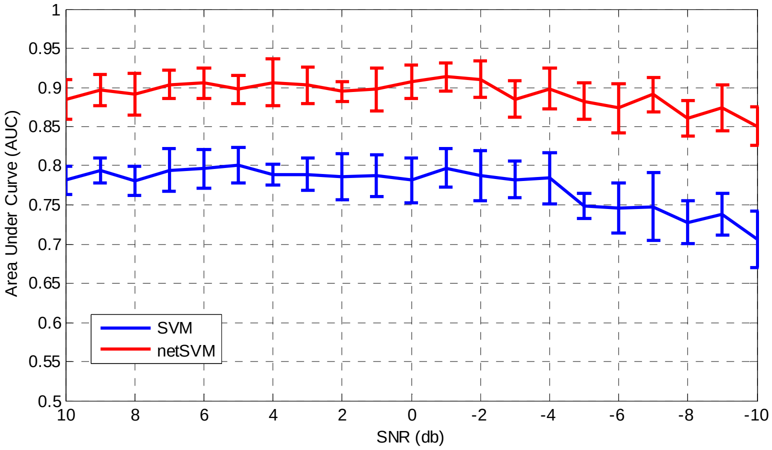


Figure 4.10. Comparison Results of subnetwork identification AUC values on simulation data sets with different signal-to-noise levels for network-constrained SVM and conventional SVM. AUC values are calculated based on 100 simulations, where standard deviations are shown in the error bar.

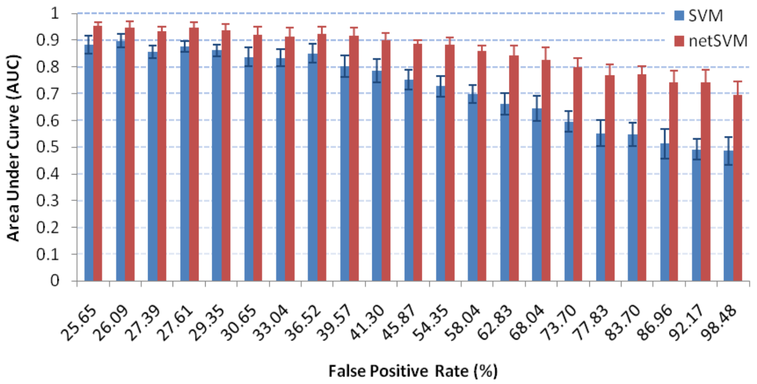


Figure 4.11. Comparison Results of subnetwork identification AUC values on simulation data sets with different false positive rates (w) for network-constrained SVM and conventional SVM. AUC values are calculated based on 100 simulations, where standard deviations are shown in the error bar.

4.5. Breast cancer Study

4.5.1. Drug untreated breast cancer data

In this study, we first applied the proposed BMRF-based method for subnetwork identification from two gene expression data sets of breast cancer patients without drug treatment, as previously reported by van de Vijver *et al.* [172] and Wang *et al.* [173] respectively. Particularly we focused on ER positive patients in our study. Among them, 78 patients in van de Vijver *et al.* [172] and 80 in Wang *et al.* [173] have been detected metastasis during follow-up visits within 5 years of surgery, which were assigned to 'recurrence' group. For the remaining 217 and 129 patients were labeled 'non-recurrence'.

Since there are many false positives in the PPI network, we assumed that genes highly connected are more confident than those sparsely connected. Therefore, we determined the seed genes from ER focused network (Figure 4.4) according to the node degree in the ER focused network. In our experiments, we set a threshold of 5 to select 202 seed genes for this breast cancer study. Subnetworks were identified from Protein-protein interaction network from HPRD [32], which contains about 9,000 genes and 35,000 interactions. We converted gene expression data from probe set IDs to Entrez gene IDs. The probe set ID with largest variance across patients' samples was used if multiple probe set IDs are linked to one Entrez gene ID. After mapping the PPI and two data sets, there were 7,249 genes with 27,885 interactions to be investigated.

From 202 bootstrapping subnetworks we determined the significant subnetworks according to network size and network score. A network is considered as significant if the network size is greater than 5 and the network score is larger than 1.65 (p-value = 0.05, normal distribution). 27 significant subnetworks are detected on Wang *et al.* [173] by the BMRF-based method. We trained a classifier using netSVM [174] based on these subnetworks from Wang *et al.* [173] and predicted on van de Vijver *et al.* [172]. The ROC curves for five-fold cross validation and independent test are shown in Figure 4.12. Specifically, the accuracy of five-fold cross validation is 72.58% with 74.61% sensitivity and 72.09% specificity. For independent test we achieved 69.14% accuracy with 73.13% sensitivity and 60.26% specificity. Similarly 14 significant subnetworks are identified on van de Vijver *et al.* [172] by the BMRF-based method. The accuracy of five-fold cross validation on van de Vijver *et al.* [172] is 70.20% with 72.22% sensitivity and 70.05% specificity. The independent test on Wang *et al.* [173] gives 63.16% accuracy with 72.50% sensitivity and 57.36% specificity. Figure 4.13 shows the ROC curves for cross validation and independent test.

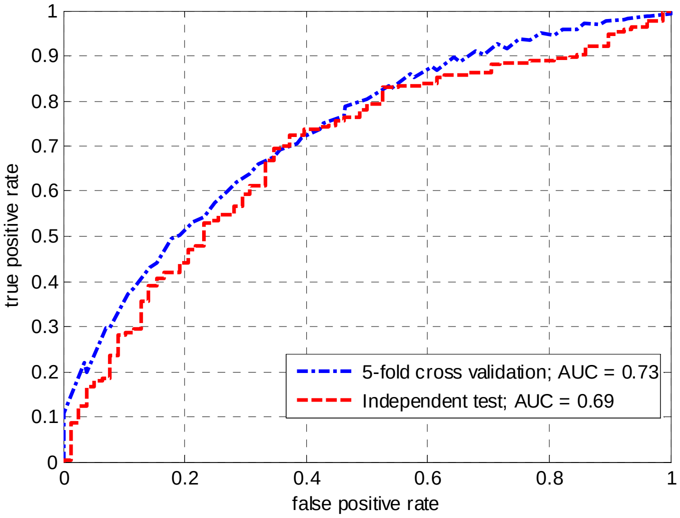


Figure 4.12. ROC curves for 5-fold cross validation on Wang *et al.* [173] and independent test on van de Vijver *et al.* [172] .

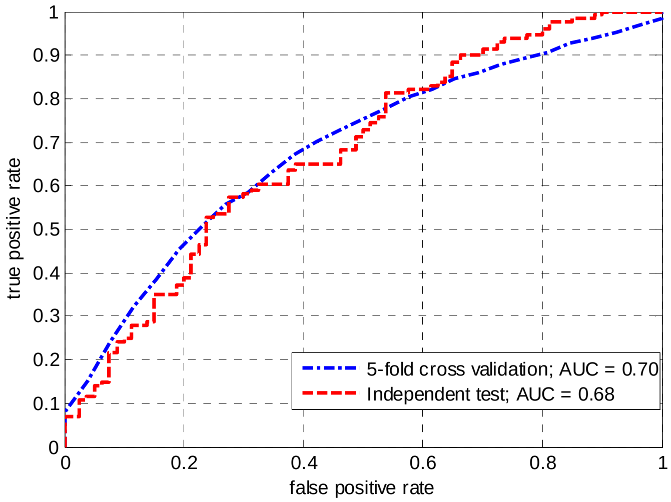


Figure 4.13. ROC curves for 5-fold cross validation on van de Vijver *et al.* [172] and independent test on Wang *et al.* [173].

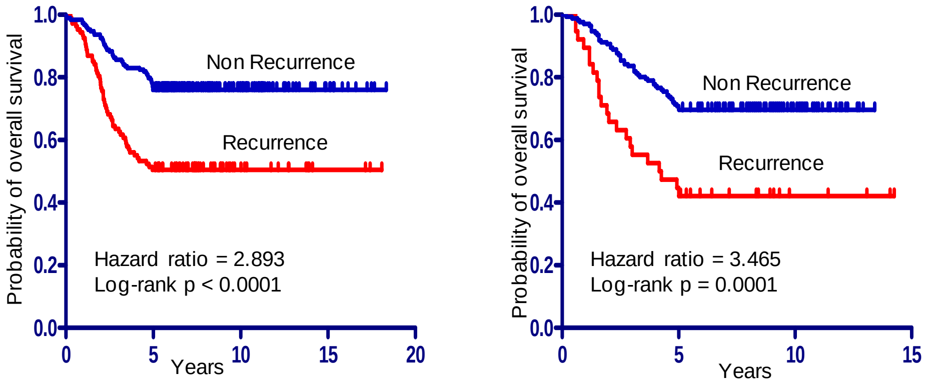
The cross validation results are comparable with the reported ones and our method achieved better prediction performance on independent data set. The detailed comparable results are shown in Table 4.1 for cross validation and Table 4.2 for independent test, respectively. The Kaplan-Meier analysis of independent test on two data sets (Figure 4.14) also showed highly significant differences in terms of overall survival between the groups predicted as 'recurrence' and 'non-recurrence', respectively.

Table 4.1. Comparison of prediction performance of five-fold cross validation on Wang *et al.* [173] and van de Vijver *et al.* [172] for bootstrapping MRF-based method and network-based method [41].

Data set	Method	Accuracy	Sensitivity	Specificity
Wang *et al.* [173]	BMRF-based	72.58%	74.61%	72.09%
	Network-based	72.2%	90%	61.7%
van de Vijver *et al.* [172]	BMRF-based	70.20%	72.22%	70.05%
	Network-based	70.1%	90%	63.1%

Table 4.2 Comparison of independent test performance on Wang *et al.* [173] and van de Vijver *et al.* [172] for bootstrapping MRF-based method and network-based method [41].

Data set	Method	AUC	Accuracy	Sensitivity	Specificity
Wang *et al.* [173]	BMRF-based	68%	63.16%	72.50%	57.36%
	Network-based	63%	48.8%	90%	24.4%
van de Vijver *et al.* [172]	BMRF-based	69%	69.14%	73.13%	60.26%
	Network-based	72%	55.8%	90%	43.3%

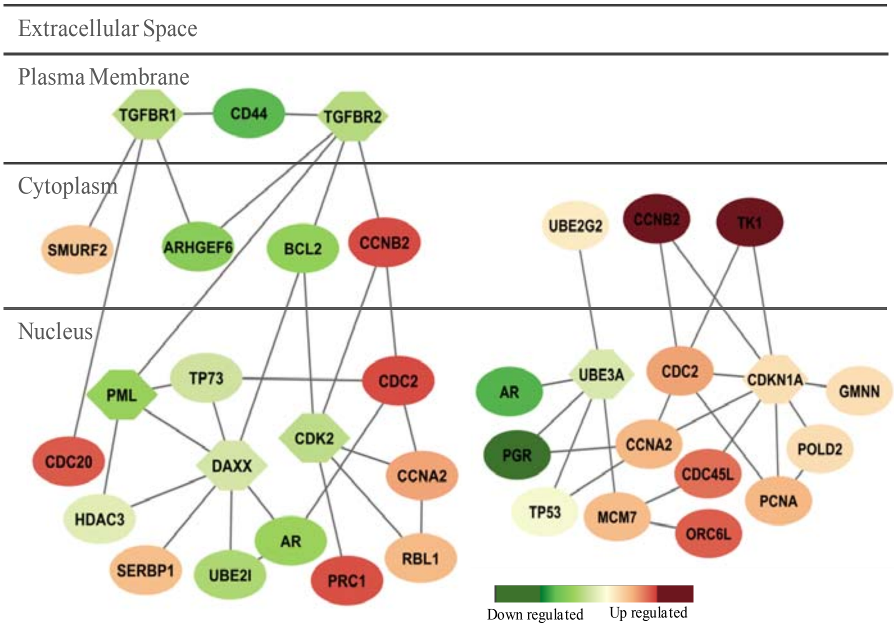


(a) (b)

Figure 4.14. Kaplan-Meier analysis for overall survival of independent test on (a) van de Vijver *et al.* [172] and (b) Wang *et al.* [173].

We further checked the overlapped genes in subnetworks as identified from two data sets. There are in total 128 genes from Wang *et al.* [173] and 77 genes from van de Vijver *et al.* [172]. 16 genes are shown in the subnetworks from both data sets. Among them, many genes are known to be related to breast cancer estrogen signaling. For instance, although the functional role of androgen receptor (AR) is still unclear, its expression was shown to be a prognostic indicator in breast cancer [175]; steroid hormones and their receptors (PGR) are involved in the regulation of eukaryotic gene expression and affect cellular proliferation and differentiation in target tissues [176]; BCL2 is an independent predictor of breast cancer outcome and can be useful as a prognostic marker [177]. AR, BCL2, CCNA2 and CCNB2 are involved in subnetworks identified from Wang *et al.* [173] (Figure 4.15(a)); AR, CCNA2, CCNB2 and PGR are involved in subnetworks identified from Vijver *et al.* [172] (Figure 4.15 (b)). Although the genes in these two subnetworks in Figure 4.15 are not exactly same, their enriched pathways and GO functions annotated by MsigDB database [48] are similar, which are cycle cell pathway and breast cancer signaling.

The experimental results support that our method can identify some hub genes that may not be significantly differentially expressed between 'recurrence' and 'non-recurrence' groups, for examples, DAXX (Figure 4.15 (a)), TP53 and CDKN1A (Figure 4.15(b)). The t-test p-values of these genes are larger than 0.05 between two groups, however, they are included in the identified subnetworks because their neighboring genes are significantly differentially expressed.



(a) (b)

Figure 4.15. Subnetwork identified from (a) Wang *et al.* [173] and (b) van de Vijver *et al.* [172]. Subnetworks are merged if more than 2 genes are common. Node shape indicates the seed gene (hexagon) or non-seed gene (ellipse). Node color indicates the fold change between 'recurrent' and 'non recurrent' group. Red means over-expressed in 'recurrent' group and green means over-expressed in 'non-recurrent' group. Enriched pathways and GO functional annotations are (a) Cell cycle pathway: 2.91e-10; Breast cancer estrogen signaling: 4.3e-05 (b) Cell cycle: 3.18E-15; Breast cancer estrogen signaling: 4.02e-07.

For the other identified subnetworks, we also conducted pathway enrichment analysis and functional annotation based on the MsigDB database. Many pathways or biochemical activities were identified. Among them, cell cycle pathway or cell cycle process is highly enriched in two data sets (Figure B.5, Figure B.6, Figure B.7and Figure B.8 in Appendix B). Apoptosis (Figure B.5) and signaling transduction (Figure B.1) were shown in Wang *et al.* [173]. Insulin receptor pathway was shown in van de Vijver *et al.* [172] (Figure B.6 and Figure B.7)). Detailed networks and annotations are shown in Appendix B. These results show that our findings are consistent with the ones reported in the original studies [172, 173].

4.5.2. Drug treated breast cancer data

As an exploration, we also applied our method onto an in-house microarray data set for anti-estrogen resistance study of breast cancer. The study was designed to find estrogen-related networks or pathways to help understand the recurrence of breast cancer after drug treatment (Tamoxifen). 64 samples have been profiled with Affymetrix GeneChip U133 Plus 2.0 Array. Among them, 24 samples were labeled as 'early recurrence' (< 3 years) and 40 samples were labeled as 'non-recurrence' (> 10 years) according to their relapse-free time. We used one public data set by Loi *et al.* [178] for cross data study, which also collected tamoxifen-treated samples and arrayed on the same microarray platform. Among them, 12 samples were labeled as 'early recurrence' and 12 samples were labeled as 'non-recurrence' according to the same relapse-free time division. We used Probe Logarithmic Intensity Error (PLIER) algorithm with quantile normalization to preprocess the original intensity data for gene expression measurements [12]. After the preprocessing, we obtained expression measurements of 54,675 probe sets in each sample. We used same seed genes as defined in the previous section to identify subnetworks by integrating the gene expression and network data sets. After mapping the PPI data and two gene expression data sets, there were 9,247 genes with 34,883 interactions remained in this experiment.

From 202 bootstrapping subnetworks we determined the significant subnetworks according to network size and network score. A network is considered as significant if the network size is greater than 5 and the network score is larger than 1.65 (p-value = 0.05, normal distribution). 19 significant subnetworks were identified from the in-house data set by the BMRF-based method. We trained a classifier using netSVM based on these subnetworks on our in-house data set and predicted on Loi *et al.* [178]. The ROC curves of cross validation and independent test are shown in Figure 4.16. Specifically, the accuracy of five-fold cross validation is 80.42% with 67.21% sensitivity and 88.35% specificity. The classifier achieved 75% accuracy with 75% sensitivity and 75% specificity for independent test. Similarly, 13 significant subnetworks are identified from Loi *et al.* [178] by our BMRF-based method. The accuracy of five-fold cross validation on Loi *et al.* [178] is 87.22% with 83.58% sensitivity and 96% specificity. The independent test on the in-house data set gives 73.44% accuracy with 70.83% sensitivity and 75% specificity. The corresponding ROC curves are shown in Figure 4.17. The Kaplan-Meier analysis of independent test on two data sets (Figure 4.18) showed highly significant differences in terms of overall survival between the groups as predicted to be either 'early recurrence' or 'non-recurrence'.

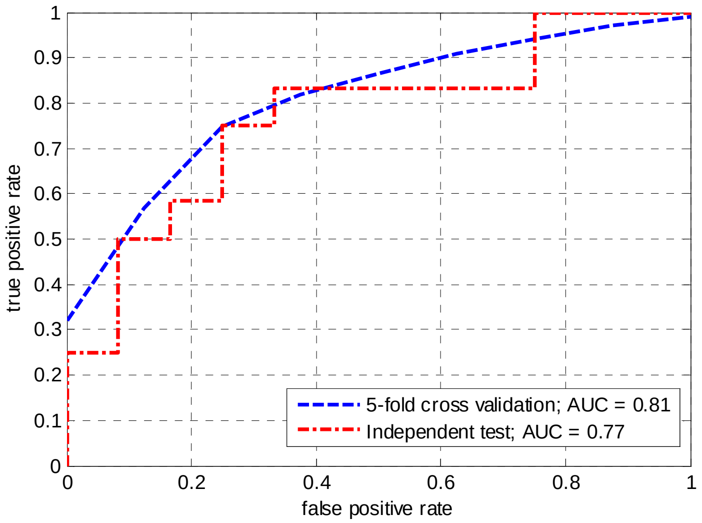


Figure 4.16. ROC curves for 5-fold cross validation on in-house data set and independent test on Loi *et al.* [178] .

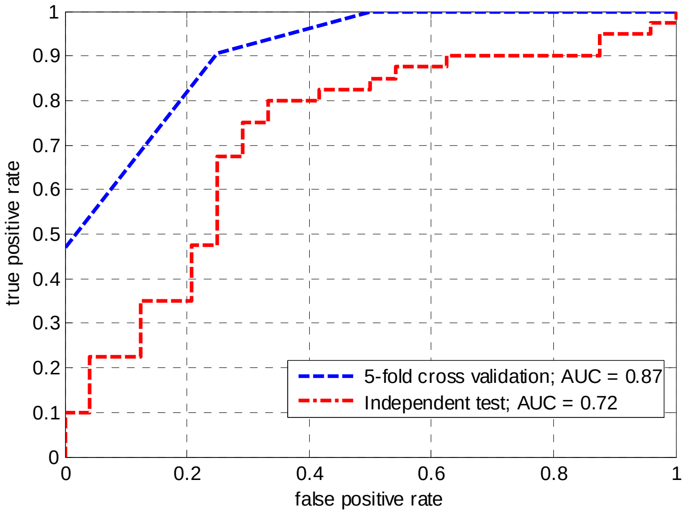
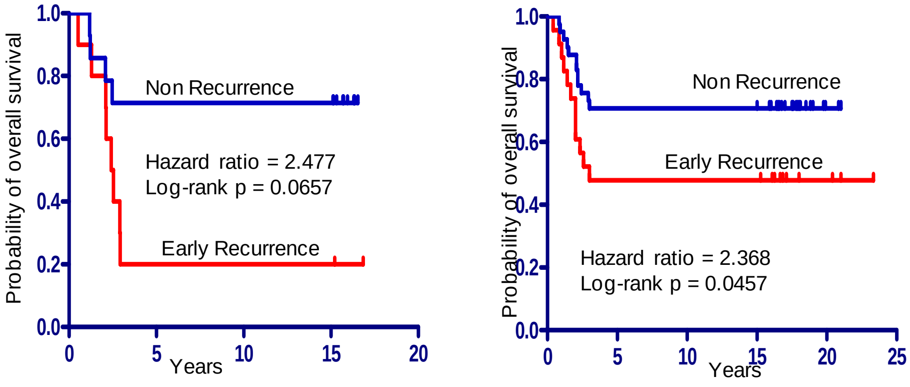


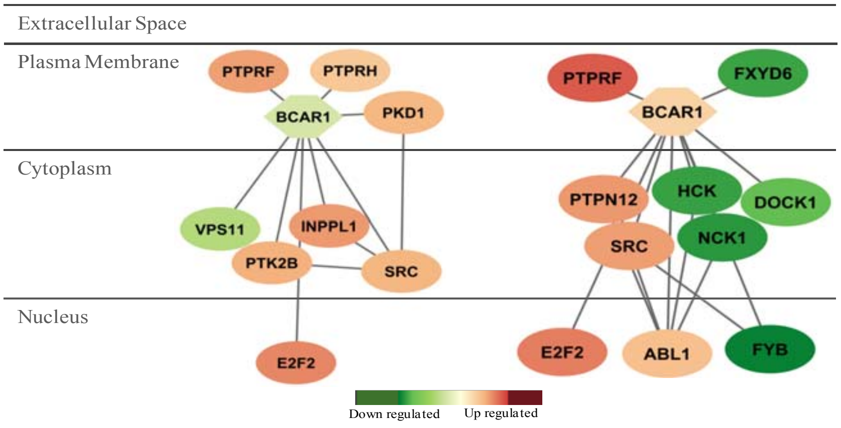
Figure 4.17. ROC curves for 5-fold cross validation on Loi *et al.* [178] and independent test on in-house data set.



(a) (b)

Figure 4.18. Kaplan-Meier analysis for overall survival of independent test on (a) Loi *et al.* [178] and (b) In-house data set.

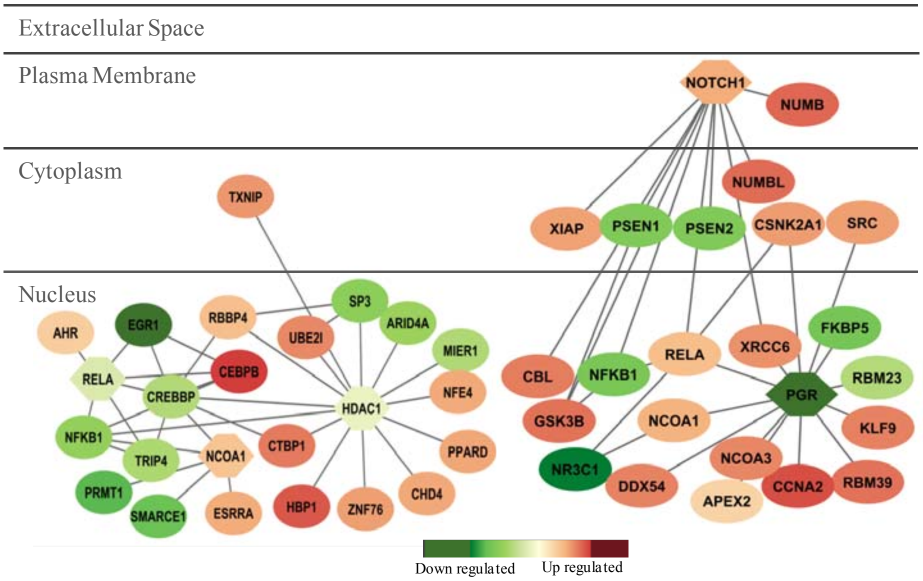
We then checked overlapped genes in the subnetworks identified from two data sets. There are in total 127 genes from in-house data set and 134 genes from Loi *et al.* data [178]. 24 genes are shown in common from both data sets. Among them, many genes are known to be associated with cell cycle/ cell proliferation or signaling pathway. For examples, E2F1, E2F2, SKP1, SKP2, CDC25A, CCNA2, CCNE2, CDK2 and RB1 are cell cycle related genes; RASA1, MAP2K3, NFKB1 and RELA are related to MAPK signaling pathway; PTK2B, BCAR1 and SRC are involved in epidermal growth factor receptor signaling pathway. Interesting, breast cancer anti-estrogen resistance 1 (BCAR1) network is shown in the networks extracted from both data sets (Figure 4.19) with overlapped genes PTPRF, SRC and E2F2, where cell to cell adhesion signaling and down-stream transduction signaling are enriched. BCAR1 has been reported that the alteration of BCAR1 is responsible for the development of anti-estrogen resistance and overexpression of BCAR1 was observed in anti-estrogen resistant human breast cancer cells [179, 180]. This may suggest that these networks identified by our method are associated with drug resistance in breast cancer. From the figure we can see that BCAR1 itself is not significantly expressed between recurrence and non-recurrence groups and reversely expressed in two data sets. Nevertheless, our method can highlight this hub gene according to the expression patterns of its neighboring genes.



(a) (b)

Figure 4.19. Subnetwork identified from (a) In-house data set and (b) Loi *et al.* [178]. Subnetworks are merged if more than 2 genes are common. Node shape indicates the seed gene (hexagon) or non-seed gene (ellipse). Node color indicates the fold change between ‘recurrent’ and ‘non recurrent’ group. Red means over-expressed in ‘recurrent’ group and green means over-expressed in ‘non-recurrent’ group. Enriched pathways and GO functional annotations are (a) Links between Pyk2 and Map Kinases: 1.64e-06; Cell to Cell Adhesion Signaling: 6.07e-05. (b) Integrin Signaling Pathway: 9.42e-07; Genes involved in Down-stream signal transduction: 5.72e-06; ErbB signaling pathway: 8.98e-05; Cell to Cell Adhesion Signaling: 9.26e-05.

We then conducted functional annotation and pathway analysis using MsigDB database. Cell cycle pathway is enriched in both data sets (Figure B.13, Figure B.15 and Figure B.19, Appendix B). Nucleus is also enriched in two data sets (Figure B.10 and Figure B.21). Apoptosis is shown in Loi *et al.* [178] (Figure B.20). The functional annotation and pathways of these networks are similar with the ones obtained from drug untreated data sets [172, 173], which may indicates that the subnetworks are related to the development of breast cancer. Besides, the results on these two data sets show more signaling information compared with the results from drug untreated microarray data sets. For example, Figure 4.20 shows the two subnetworks identified on in-house and Loi *et al.* [178], in which notch signaling pathway is enriched in both subnetworks. Specifically, in highly conserved Notch signaling pathway, some genes located in plasma member and cytoplasm were identified from Loi *et al.* [178]**,** which include Notch1, PSEN1, PSEN2, NUMB and NUMBL (Figure 4.20(b)); some genes located in nucleus were identified from our in-house data set, including CREBBP, HDAC1 and CTBP1 (Figure 4.20(a)). The results indicate that different data sets may reveal different active parts of one common underlying mechanism. Notch signaling pathway has been studied in many papers [181-183] and it is suggested that the inhibition of Notch signaling may be a therapeutic strategy for breast cancer [181]. In [184], the study has showed that the combination of tamoxifen and Notch inhibitors could eliminate the emergence of Tamoxifen resistance.



(a) (b)

Figure 4.20. Subnetwork identified from (a) In-house data set and (b) Loi *et al.* [178]. Subnetworks are merged if more than 2 genes are common. Node shape indicates the seed gene (hexagon) or non-seed gene (ellipse). Node color indicates the fold change between 'recurrent' and 'non recurrent' group. Red means over-expressed in 'recurrent' group and green means over-expressed in 'non-recurrent' group. Enriched pathways and GO functional annotations are (a) WNT signaling pathway: 2.67E-07; Notch signaling pathway: 1.45E-04 (b) Nucleus: 2.99E-06; Notch signaling pathway: 6.18E-08; Signal transduction: 7.94E-05.

Signaling pathway is more complicated and diverse. We can see that TGF-beta signaling pathway (Figure B.11) and WNT signaling pathway (Figure B.16) are enriched in the networks from our in-house data set, while WNT signaling pathway (Figure B.20), Notch signaling pathway (Figure B.21) and ErbB signaling pathway (Figure B.22) are enriched in the networks from Loi *et al.* [178]. It has been observed that WNT proteins are sometimes overexpressed and the downstream components of the WNT signaling pathway are activated in a significant proportion of human breast tumors [185, 186]. Transforming growth factor-beta (TGF-beta) is a tumor suppressor, which is involved in many types of human cancer, including breast cancer [187]. The recent findings through the in vitro cell line study provided evidence that TGFbeta-dependent mechanism could help induce immnunosuppression in the tumor microenvironment, which may contribute to the development of antiestrogen resistance in breast cancer [188]. Researchers have shown that dysregulation of human epidermal growth factor receptor (ErbB/HER) pathways by over-expression or constitutive activation can promote tumor processes and is associated with poor prognosis in many human breast tumors [189]. In a recent study by Schiff et al.[190], cross-talk between estrogen receptor and ErbB pathways was discussed as a molecular target to potentially overcome endocrine resistance. These subnetworks showed an important difference between drug untreated and drug treated data sets and may provide insights into drug resistance in breast cancer.

Finally, Figure 4.21 shows the comparison of AUC values of independent test on four data sets for different feature selection and classification methods. We compared network-constrained SVM classification method with conventional SVM method, with the BMRF-based subnetwork identification method or traditional t-test based gene selection method. For traditional t-test based method, we selected same number of genes as the ones identified by the BMRF-based method. From the figure we can see that subnetwork identification method can improve the prediction performance on independent data set compared with traditional t-test based method. Furthermore, the performance can be further improved if we use our network-constrained classification method. The results indicate that our network-based methods can improve the reproducibility of identification of significant genes/subnetworks across different data sets.

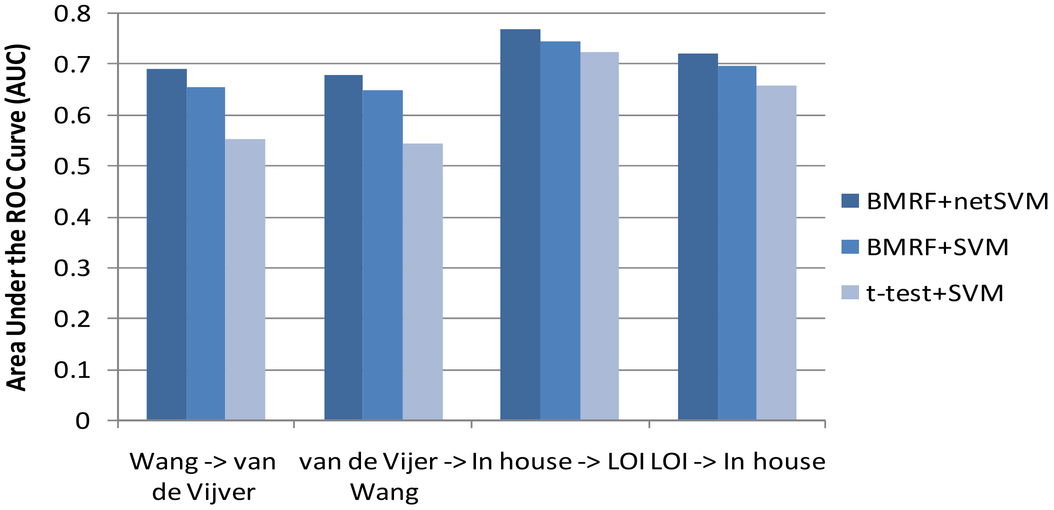


Figure 4.21. Comparison of AUC values of independent test on four data sets for different feature selection and classification methods.

4.6. Discussion

A common characteristic of the methods that we proposed for subnetwork identification and network-based prediction is that the dependency among genes in the network is explicitly represented in the objective functions that we designed. Specifically, we penalized the weighted difference between connected gene pairs to impose the smoothness of estimated parameters. In this way, the genes with larger number of neighbors are emphasized in terms of their underlying contribution to the subnetwork. Such consideration is biologically meaningful in that many hub genes play important and central role in the underlying biological process compared to their downstream genes. By capturing the underlying hub genes, it becomes possible to identify important subnetworks with common functional annotation among different data sets, hence to improve the reproducibility of prediction across different data sets.

For the BMRF-based subnetwork identification, there are several issues to be further investigated. First, the BMRF-based subnetwork identification method searches for subnetworks along a pre-defined PPI network, where the interactions in the PPI network are fixed. However, the PPI network is far from being complete, and it includes a large degree of noise and false positive interactions. Moreover, it is observed that the interactions among genes are tissue-specific or condition-specific. Therefore, it is important and necessary to address the specificity of interactions in the PPI network for the subnetwork identification problem. Second, besides PPI information, pathway information provides more direct source for pathway analysis. Different from PPI network, pathway network can be represented as a directed graph. To incorporate pathway information, our method needs to be further extended to take into account the direction information. Finally, more sophisticated statistical tests such as the ones proposed in Chuang et al. [41] need to be carried out in our experiments for significant subnetwork identification.

4.7. Conclusion

We have proposed a novel subnetwork identification method by integrating microarray data and protein-protein interaction data. We formulated a new network score through an MRF-MAP framework. Modified simulated annealing algorithm is utilized to search for subnetworks with maximal network scores and bootstrapping scheme is implemented to find the most reliable subnetworks. The simulation experiments have demonstrated the effectiveness of our proposed method. Furthermore, we have studied two types of breast cancer data (i.e., drug untreated and drug treated) and the results showed that our method can successfully identify many significant subnetworks, which could lead to good prediction performance and are associated with estrogen signaling in the development of breast cancer and drug resistance. We have further proposed a network-constrained support vector machine (netSVM) for classification by incorporating protein-protein interactions as prior knowledge. The network structure is explicitly formulated in the objective function of SVM through the Laplacian matrix of network. Through this way, we have imposed the smoothness of estimated coefficients over network to emphasize the role of hub genes in defining classification hyperplanes. Extensive simulation studies have demonstrated the effectiveness of this proposed method by providing more accurate prediction performance as well as more significant features related to the underlying network. The study of breast cancer on real microarray data sets has further demonstrated that netSVM can provide better reproducibility of prediction performance across different data sets than conventional SVM can.

5 Contribution, Future work and Conclusion

5.1. Summary of original contribution

In this dissertation, we have developed novel computational methods for integrative analysis of multiple genomic data including microarray gene expression data, protein-DNA interaction data and protein-protein interaction data, aiming to reveal and understand disease related biological functions and processes. We summarize the original contribution of this dissertation as follows.

5.1.1. Knowledge-guided multi-scale ICA for biomarker identification on time course microarray data

We have developed a novel method, knowledge-guided multi-scale independent component analysis, for biomarker identification on time course microarray data. We aim to infer knowledge-relevant regulatory signals and identify corresponding biomarkers beyond partial prior knowledge, which is constructed from multiple knowledge sources related to disease-specific biological functions. We first cluster the whole gene population into multiple sub-populations in which only a few biological processes are involved. By applying ICA to multi-level gene clusters, an examination of the revealed regulatory modes can uncover knowledge of the underlying biological regulatory mechanisms. The optimal number of clusters is determined by cross validation based on knowledge genes. Finally, disease-specific biomarkers are extracted according to the strength of their association with the extracted regulatory modes. In addition, we have designed a statistical test procedure to measure the transcription factor enrichment of a selected gene set based on motif information.

We have applied the proposed method to two gene expression data sets to identify biomarkers: yeast cell cycle microarray data and Rsf-1-induced microarray data. The experimental results show that our method can extract apparently biologically meaningful and condition-related biomarkers. The performance of the proposed method significantly outperforms several baseline methods for biomarker identification. More importantly, the proposed method has notable potential to discover novel biomarkers beyond any partial prior knowledge.

5.1.2. Transcriptional regulatory network identification by multi-level support vector regression by integration of binding information and mRNA gene expression profiles

We have developed a multi-level support vector regression method for transcriptional regulatory network identification based on motif binding information and mRNA gene expression profiles. The method aims to reduce false positives in discovery of regulatory modules due to the noises from binding and expression data. We have formulated the relationship of gene expression levels and the binding strength of their corresponding transcription factors using linear equations and proposed two-stage support vector regression to solve and refine the estimation of transcription factor activity and the binding strength. The convergence of two-stage SVR is mathematically proved. A statistical test is used to find the significant regulatory modules. A multi-level strategy based on clustering is applied to reduce the noise effect of binding information and find stable and significant regulatory modules. Finally, a weighted voting scheme is implemented for target gene identification, taking into account the entire multi-level analysis.

We have first assessed the reliability of knowledge information on ml-SVR method for transcriptional network identification through simulation experiment. We have applied the proposed method to simulation data and yeast cell cycle data to assess its performance for transcriptional regulatory module identification, in comparison with those of existing benchmark methods. The comparison results clearly demonstrate that the proposed ml-SVR method notably outperforms other methods, which show the effectiveness of identifying condition-specific regulatory modules of our methods. We have also applied the method to two breast cancer cell line data sets to identify breast cancer related transcriptional regulatory module associated with the conditions of estrogen treatment. The results show that our method can successfully identify significant condition-specific transcription factors and target genes associated with estrogen signaling in breast cancer.

5.1.3. BMRF-based subnetwork identification and network-based prediction on microarray gene expression profiles by integrating protein-protein interaction network

We have developed a BMRF-based subnetwork identification method to discover significant subnetworks between two phenotypes based on microarray gene expression data and protein-protein interaction network. We derive a novel network score based on a Markov random field (MRF) – maximum a posterior (MAP) framework, taking into account the dependency among the genes within a subnetwork. The MRF-based network score emphasizes the significance of hub genes even though they often do not have significant differential power compared with their downstream genes. A modified simulated annealing search algorithm is utilized to find the optimal/suboptimal subnetworks with maximal network scores. Finally, bootstrapping scheme is implemented to help identify confident subnetworks.

We have also developed a network-constrained support vector machine for classification and prediction. We derive the method by adding a Laplacian matrix of network constraint in the objective function to impose the smoothness of estimated coefficients along a network. Statistical significance test is then carried out to assess the importance of input features in a network.

We have applied the proposed methods to simulated data and compared the performance with those of existing benchmark methods, for both BMRF-based subnetwork identification and network-constrained support vector machine. The experimental results show that the proposed methods outperform other methods in terms of identifying underlying significant subnetworks and obtaining more accurate prediction results. We have also studied four microarray data sets acquired from breast cancer patients with or without drug treatment. We have identified significant subnetworks that can differentiate the survival times of patients. The experimental results show that the method can not only achieve an improved prediction performance for independent test, but it can also identify biologically meaningful subnetworks related to the development of breast cancer and drug resistance.

5.2. Future work

There are several remaining problems/topics that can be further studied. We discuss some of the extensions in this section. The topics are related to what we have presented above and can be summarized as follows.

5.2.1. Transcriptional regulatory network identification

- ***Enhancing the ml-SVR method to address uncertainty from both sources***

In the current ml-SVR method, we have proposed two-stage support vector regression and multi-level strategy to reduce the noise effect coming from microarray data and binding data. However, besides noises, there are many uncertainties from both data sources due to the limited knowledge. We have shown in our simulation study that inaccurate prior knowledge may lead to unreliable results. Therefore the justification of prior knowledge needs to be addressed in future study. Furthermore, the current method assumes that co-expressed genes should be co-regulated to some degree; hence, genes are clustered based on their expression profiles alone. However, this may not be true in some cases [38]. Clustering method based on gene expression profiles alone may not be appropriate to form sub-clusters. Therefore, the method can be further extended to address the noise effect of microarray gene expression data and uncertainty of forming gene clusters. Recently, Brynildsen et al. [191] developed a Gibbs sampling technique to identify genes that have consistent expression and binding information. Gong et al. [137] proposed to cluster genes based on their gene expression data and binding motif information together, which may provide more accurate gene clusters for transcriptional regulatory network analysis. These proposed methods may provide some directions and insights to enhance our method.

- ***Extending ml-SVR method for co-regulatory module identification***

Currently our ml-SVR method identifies transcriptional regulatory network only focusing on each individual transcription factor and module. However, finding cooperative transcription factors is also of importance for many biological studies. The method can be extended to find cooperative transcription factors and then determine their target genes using regression analysis. However, there are several issues that need to be considered carefully. An important issue is how to determine an appropriate motif set for SVR fitting. In our preliminary study [138], we have developed a stepwise forward greedy search strategy using a modified loss function to find the co-operative motifs in a given gene set. However, more sophisticated methods such as optimal search algorithms [152, 154, 158] need to be developed and analyzed to address the problem of co-regulatory module identification. A recent study [192] analyzed the scaling of partnerships between regulators with the number of target genes through differential equations, which may provide some guidance for further study. Another one is how to conduct biological validation since currently ChIP-on-chip technology is still in an initial development stage (with a limited number of available antibodies) and it can be only used for validating of target genes of single transcription factor at current stage.

- ***Extending the ml-SVR method to patient microarray data***

For the real application, we have applied ml-SVR method on cell line data to identify regulatory networks. The experiments have shown that our method is effective to identify significant transcription factors and their target genes since the expression pattern in cell line data is quite consistent and clear. However, it is not easy to directly extend the method to patient data for regulatory network identification since the expression pattern in patient data is much diverse due to the heterogeneity and noisy in patients. Therefore, new information based on the characteristics of patient data such as survival time needs to be incorporated, and correspondingly the algorithm needs to be modified to identify transcriptional regulatory networks from patient microarray data. For example, Cheng et al. [193] proposed to use a Kolmogorov-Smirnov test-like method to identify transcription factors associated with patient survival in cancers. The method avoided correlation between gene expression and binding affinity but emphasized the difference of transcription factor activities between two phenotype groups. Their method may provide us some directions on how to further extend our ml-SVR method for patient data analysis.

5.2.2. BMRF-based subnetwork identification and network-based prediction

- ***Enhancing structure learning for BMRF-based subnetwork identification***

The BMRF-based subnetwork identification method that we developed searches for subnetwork along a pre-defined PPI network, where the interactions in the PPI network are fixed. As we pointed out before, the PPI network is far from being complete, currently with noises and false positive interactions included. Additionally, the interactions among genes are tissue-specific or condition-specific in nature. Therefore, addressing the specificity of interactions in the PPI network becomes an important and necessary task for subnetwork identification. The BMRF-based method should be extended to evaluate the confidence of a specific interaction based on the characteristics of PPI network, thus the method could be enhanced for condition-specific subnetwork identification by addressing the influence of interactions to the subnetwork.

- ***Statistical significance analysis for the subnetworks identified by the BMRF-based method***

In the BMRF-based subnetwork identification method, we applied a simple criterion to identify significant subnetworks based on network score. More rigorous statistical tests such as the ones in Chuang et al. [41] need to be carried out. The three statistical tests in [41] measure the significance of identified subnetworks jointly. The first two tests are designed for network significance analysis to calculate global network p-value and local network p-value, respectively. The third test is a conventional significance analysis by permutation test. The significant subnetworks are then determined if all p-values pass the thresholds of three significance tests. Our method could be extended to incorporate these three significance tests to better identify significant subnetworks.

Another issue that we need to consider for significance test is the false discovery rate (FDR) to correct for multiple hypothesis testing. In the multiple hypothesis testing, it is very important to control the false discovery rate, which is the expected proportion of incorrectly rejected null hypotheses (type I errors), rather than the significance level (p-value). In the above-mentioned three statistical tests each one is a multiple hypothesis testing problem, therefore, FDR correction needs to be considered and we should calculate q value that is the FDR analogue of the p-values. Furthermore, there exists dependency among the test statistics in subnetwork identification since the network score is derived from genes that may be overlapped in different subnetworks. The control of the false discovery rate in multiple testing under dependency would be another problem to be addressed. It has been proven that the FDR controlling procedure proposed by Benjamini and Hochberg [72] also controls the false discovery rate when the test statistics have positive regression dependency on each of the test statistics corresponding to the true null hypotheses. For other forms of dependency, a simple conservative modification is proposed in [194] to control the false discovery rate. It is important to analyze the correlation of dependency and propose a conservative modification procedure in the future study.

- ***Incorporate pathway information into the BMRF-based subnetwork identification method***

Pathway information is an important source for biological data analysis. Different from PPI network, pathway network can be represented as a directed graph and extending the BMRF-based method to model directed edge information will help pathway analysis. Recently, there are several papers discussing pathway analysis and building through linear programming [195, 196], however, the existing methods are limited in that they only provide a deterministic way to identify pathways associated with biological processes. We believe that a biological process is much more complicated and multiple pathways need to work together via cross-talk in order to accomplish a specific biological process. Therefore, instead of using deterministic analysis for pathway analysis, it is necessary to develop statistical ways to identify several possible pathways with certain probability. However, it is not easy to directly extend the BMRF-based method for pathway analysis since the method is mainly focusing on subnetwork identification through a growing strategy. For pathway analysis, people are more interested in identifying several possible paths passing through some known key genes. Therefore a path search algorithm is more suitable for pathway analysis. However, in order to emphasize those hub genes that may not be significantly differentially expressed between two phenotypes/conditions, we can still utilize the MRF-MAP framework to estimate posterior gene scores instead of observed gene scores for pathway analysis and building.

5.3. Conclusion

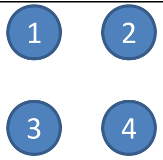
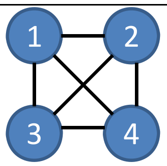
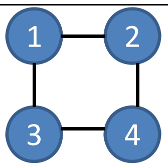
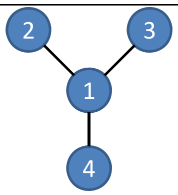
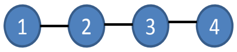
In conclusion, we have conducted research on three major topics for biological data analyses, focusing on integrating microarray gene expression data and other different types of biological knowledge data sources. We have first introduced the characteristics of multiple biological data sources including microarray gene expression data, protein-DNA interaction data and protein-protein interaction data. We have then developed different mathematical models to integrate different types of data to solve specific biological problems. For each topic, we have discussed the problems in detail, transformed them into mathematical models and derived algorithms to solve the problems. Then we have conducted simulation experiments and comparison studies to demonstrate the feasibility of the proposed methods. Finally we have applied the proposed methods to real biological data for cancer study. The results show the potential of these methods to improve the understanding of underlying biological mechanisms related to cancer development and treatment, hoping to help generate novel hypotheses for further biological study.

Appendix A. Case study for MAP estimator f

In order to have a better understanding of MAP estimator of **f** and its influence on the network score, we conducted a case study on small networks that include 4 nodes and have different network topologies. The result is shown in Table A.1. From the table we can see that given the same observed score, different network topologies result in different network scores. We normalized the observation **z** score and estimated **f** score vectors in order to compare different cases. Note that for non-connected case (case 1 in the table), Equation (4.12) does not exist when *k* equals to 0. So we define the estimated score equals to the observed score and therefore the network score is the average of observed scores in the graph. In the real case, based on our search algorithm, we can guarantee the network is always connected and so MAP estimator could be calculated.

We can observe that the network scores of connected cases are larger than the one of non-connected case (case 1). The estimated score of hub gene is more affected by other genes than the one of non-hub genes. For example, the change between estimated score and observed score of node 1 in case 4 is larger than the ones in other cases since node 1 in case 4 is a typical hub gene compared with others. This indicates that our method could promote the role of hub genes in the subnetwork identification. Furthermore, we observed that node with large degree may have large bounded variance, since its estimated discriminative score is determined and affected by more adjacent genes.

Table A.1. Case study of MRF-based network score and performance bounds for MAP estimator f ($\lambda = \gamma = 1$).

Case	Network	Observed **z** score	Estimated **f** score	Network score	Bounded variance
1		[0 1/3 1/3 1/3]	[0 1/3 1/3 1/3]	0.75	var(1) = var(2) = var(3) = var(4)
2		[0 1/3 1/3 1/3]	[0.22 0.26 0.26 0.26]	1.19	var(1) = var(2) = var(3) = var(4)
3		[0 1/3 1/3 1/3]	[0.21 0.26 0.26 0.27]	1.18	var(1) = var(2) = var(3) = var(4)
4		[0 1/3 1/3 1/3]	[0.33 0.22 0.22 0.22]	0.93	var(1) > var(2) = var(3) = var(4)
5		[0 1/3 1/3 1/3]	[0.18 0.28 0.30 0.23]	1.21	var(1) =var(4) < var(2) = var(3)

Appendix B. Addendum of Experimental Results for Chapter 4

In this appendix, we include the all subnetworks identified on four breast cancer data sets which are not reported in the main text of Chapter 4.

Identified subnetworks on Wang et al. (2005)

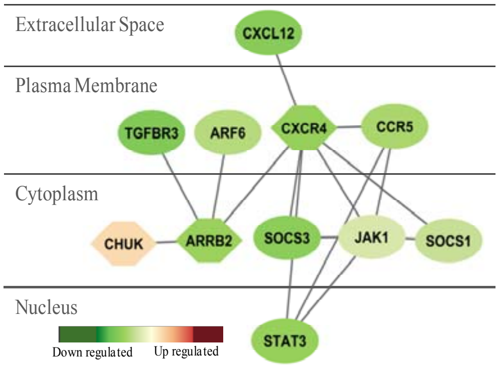


Figure B.1. Enriched pathway and GO functional annotation: Signal transduction: 3.63e-06.

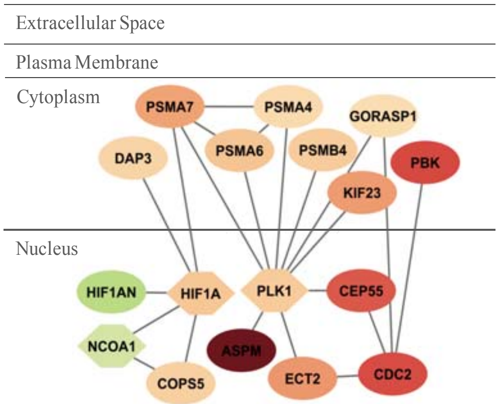


Figure B.2. Enriched pathway and GO functional annotation: Proteasome: 1.42e-08.

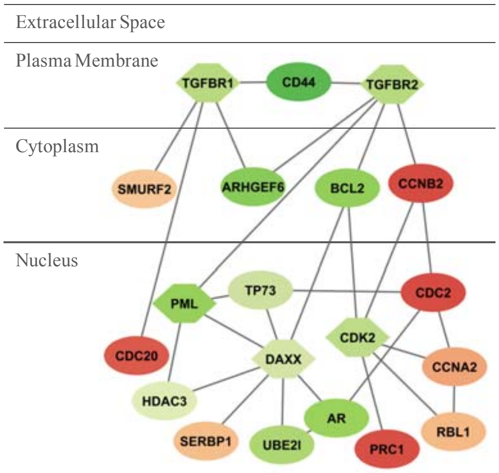


Figure B.3. Enriched pathway and GO functional annotation: Cell cycle pathway: 2.91e-10; Breast cancer estrogen signaling: 4.3e-05.

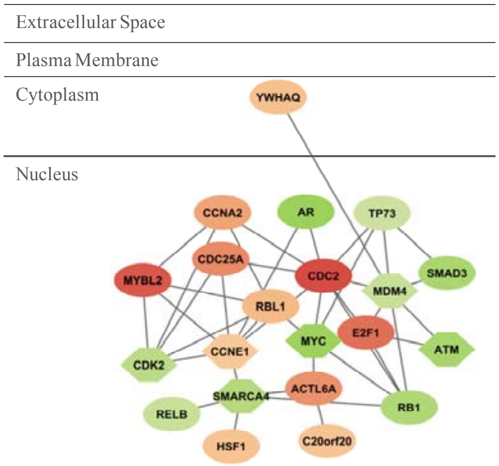


Figure B.4. Enriched pathway and GO functional annotation: Cell cycle pathway: 1.56e-14.

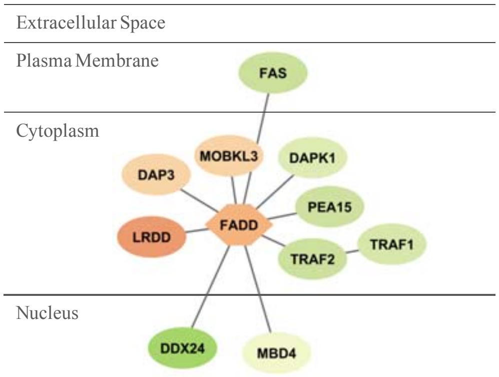


Figure B.5. Enriched pathway and GO functional annotation: Apoptosis: 1.20E-07.

Identified subnetworks on van de Vijver et al. (2002)

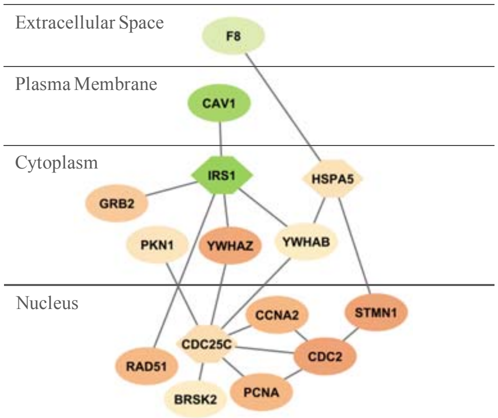


Figure B.6. Enriched pathways and GO functional annotations: Cell cycle: 1.00E-08; Insulin receptor pathway: 8.25e-07.

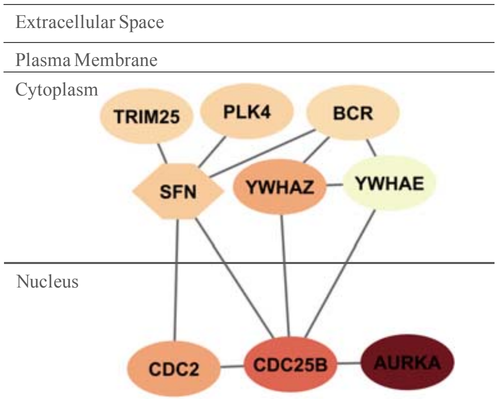


Figure B.7. Enriched pathways and GO functional annotations: Cell cycle: 2.39E-08; Insulin receptor pathway in cardiac myocytes: 1.07E-05.

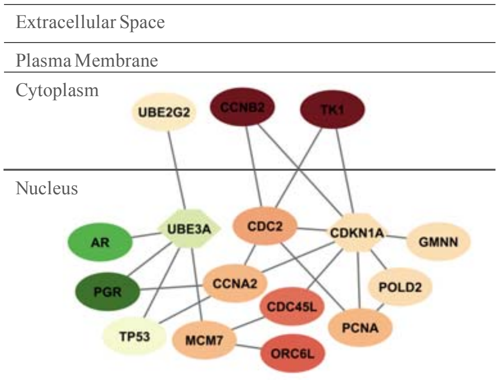


Figure B.8. Enriched pathways and GO functional annotations: Cell cycle: 3.18E-15; Breast cancer estrogen signaling: 4.02e-07.

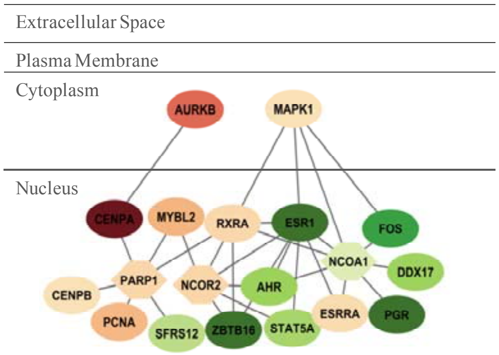


Figure B.9. Enriched pathways and GO functional annotations: Nulcear receptors: 8.55E-07.

Identified subnetworks on in-house data set

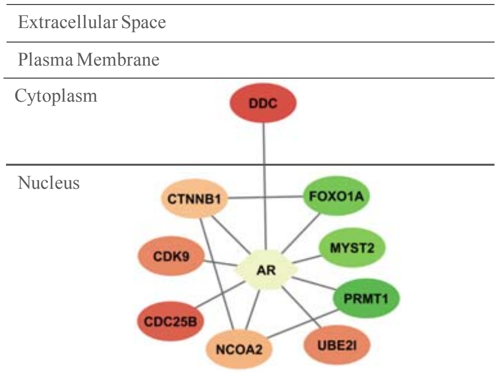


Figure B.10. Enriched pathways and GO functional annotations: Nucleus: 6.81E-04.

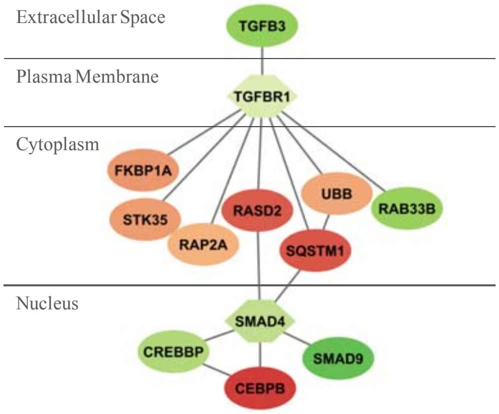


Figure B.11. Enriched pathways and GO functional annotations: TGF-beta signaling pathway: 5.17E-09;WNT signaling pathway: 4.64E-04.

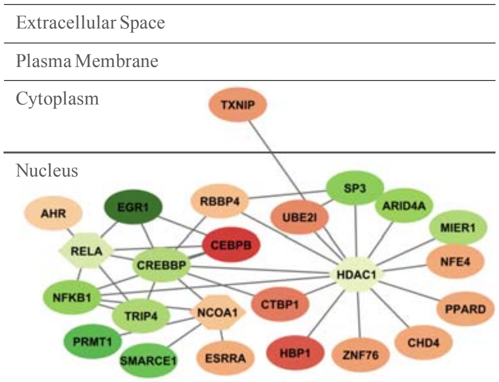


Figure B.12. Enriched pathways and GO functional annotations: WNT signaling pathway: 2.67E-07; Notch signaling pathway: 1.45E-04.

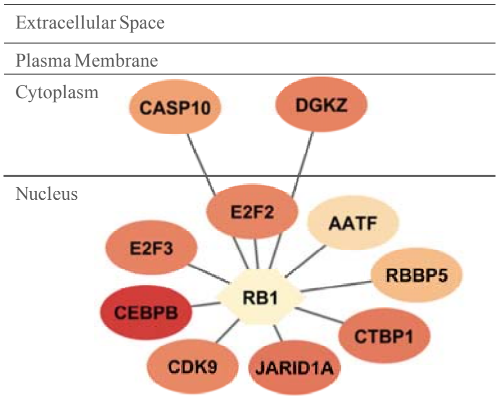


Figure B.13. Enriched pathways and GO functional annotations: Cell cycle: 2.82E-04.

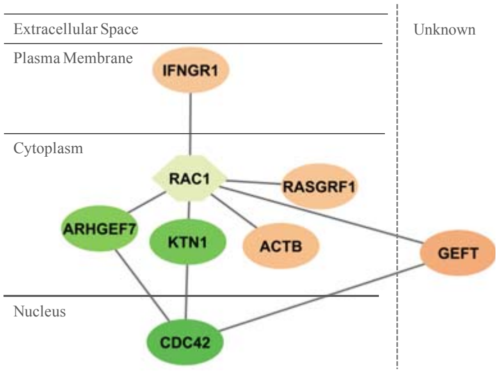


Figure B.14. Enriched pathways and GO functional annotations: Genes involved in G alpha (12/13) signaling events: 4.54E-08; Regulation of actin cytoskeleton: 1.20E-05.

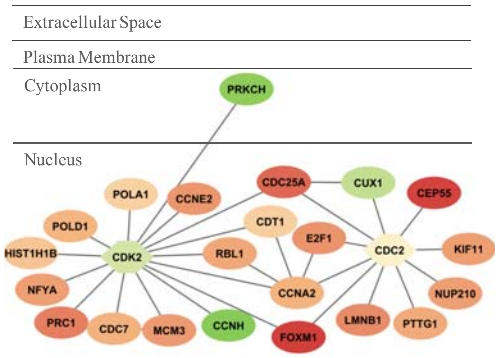


Figure B.15. Enriched pathways and GO functional annotations: Cell cycle: 5.55E-015.

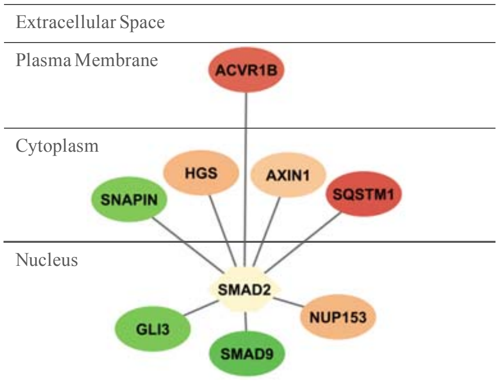


Figure B.16. Enriched pathways and GO functional annotations: Regulation of Signaling Transduction: 1.14E-05; TGF-beta signaling pathway: 1.85E-03.

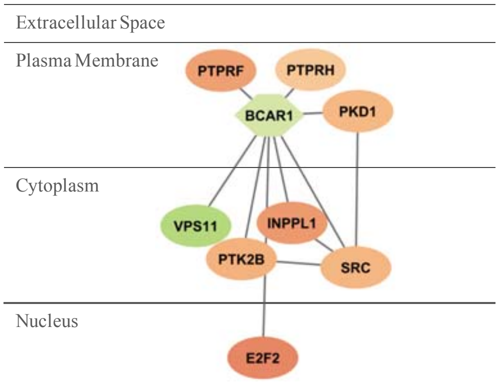


Figure B.17. Enriched pathways and GO functional annotations: Links between Pyk2 and Map Kinases: 1.64e-06; Cell to Cell Adhesion Signaling: 6.07e-05.

Identified subnetworks on Loi et al. (2008)

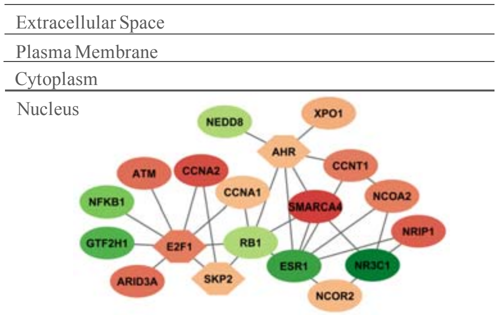


Figure B.18. Enriched pathways and GO functional annotations: Methyltransferase CARM1 methylates CBP and co-activates estrogen receptors via Grip1: 2.36E-05; Nucleus: 1.46E-06; Cell cycle pathway: 1.19E-08.

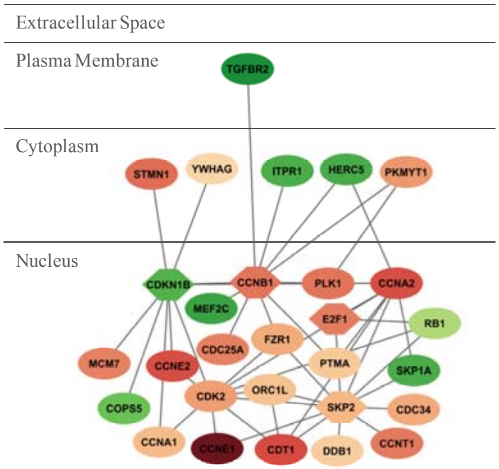


Figure B.19. Enriched pathways and GO functional annotations: Cell cycle pathway: 9.29E-16.

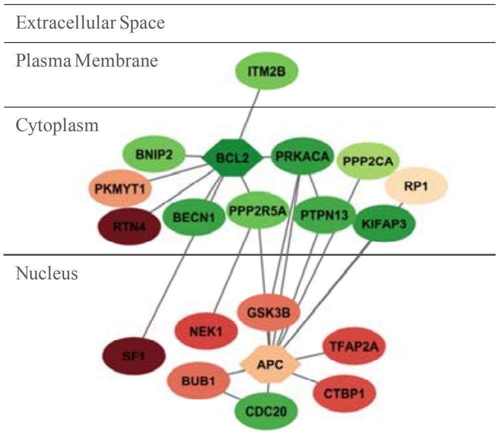


Figure B.20. Enriched pathways and GO functional annotations: Apoptosis: 4.00E-05; Negative regulation of cellular process: 1.62E-04; WNT pathway: 1.78E-07.

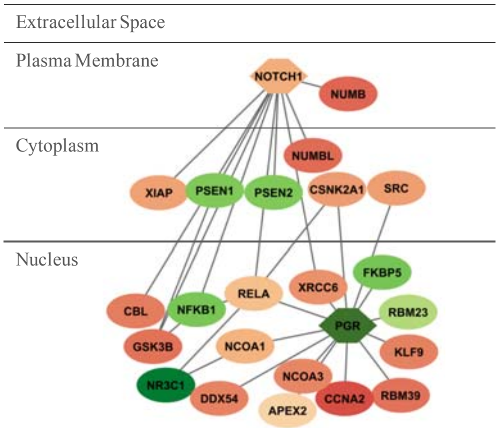


Figure B.21. Enriched pathways and GO functional annotations: Nucleus: 2.99E-06; Notch signaling pathway: 6.18E-08; Signal transduction: 7.94E-05.

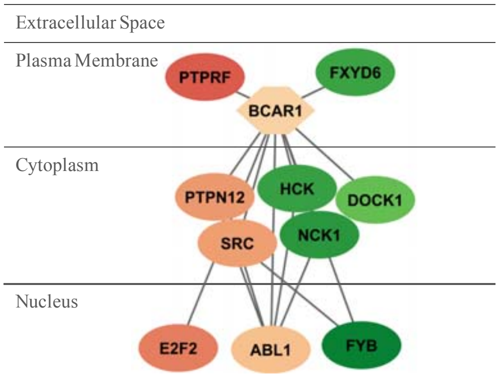


Figure B.22. Enriched pathways and GO functional annotations: Integrin Signaling Pathway: 9.42e-07; Genes involved in Downstream signal transduction: 5.72e-06; ErbB signaling pathway: 8.98e-05; Cell to Cell Adhesion Signaling: 9.26e-05.

Bibliography

[1] R. Blossey, *Computational Biology: A Statistical Mechanics Perspective*: Chapman & Hall/Crc 2006.

[2] S. Hanash, "Integrated global profiling of cancer," *Nat Rev Cancer,* vol. 4, pp. 638-44, Aug 2004.

[3] G. Parmigiani, E. S. Garett, R. A. Irizarry, and S. L. Zeger, *The Analysis of Gene Expression Data: Methods and Software*: Springer, 2003.

[4] D. A. Kulesh, D. R. Clive, D. S. Zarlenga, and J. J. Greene, "Identification of interferon-modulated proliferation-related cDNA sequences," *Proc Natl Acad Sci U S A,* vol. 84, pp. 8453-7, Dec 1987.

[5] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis, "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proc Natl Acad Sci U S A,* vol. 94, pp. 13057-62, Nov 25 1997.

[6] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467-70, Oct 20 1995.

[7] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science,* vol. 278, pp. 680-6, Oct 24 1997.

[8] "In the mainstream: microarray rivulets found campus-wide," in *The NIH CATALYST (July - August)*, 2001.

[9] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat Biotechnol,* vol. 14, pp. 1675-80, Dec 1996.

[10] Affymetrix, "Affymetrix Microarray Suite Users Guide," *Santa Clara, CA, version 5.0 edition,* 2001.

[11] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res,* vol. 31, p. e15, Feb 15 2003.

[12] Affymetrix, "Guide to Probe Logarithmic Intensity Error (PLIER) Estimation," *Edited by Affymetrix I. Santa Clara, CA,* 2005.

[13] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, pp. 185-93, Jan 22 2003.

[14] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.

[15] J. Wang, H. Li, Y. Zhu, M. Yousef, M. Nebozhyn, M. Showe, L. Showe, J. Xuan, R. Clarke, and Y. Wang, "VISDA: an open-source caBIG analytical tool for data clustering and beyond," *Bioinformatics,* vol. 23, pp. 2024-7, Aug 1 2007.

[16] S. R. Gunn, "Support Vector Machines for Classifications and Regression," 1997.

[17] T. I. Lee, S. E. Johnstone, and R. A. Young, "Chromatin immunoprecipitation and microarray-based analysis of protein location," *Nat Protoc,* vol. 1, pp. 729-48, 2006.

[18] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young, "Genome-wide location and function of DNA binding proteins," *Science,* vol. 290, pp. 2306-9, Dec 22 2000.

[19] M. Zheng, L. O. Barrera, B. Ren, and Y. N. Wu, "ChIP-chip: data, model, and analysis," *Biometrics,* vol. 63, pp. 787-96, Sep 2007.

[20] J. S. Song, W. E. Johnson, X. Zhu, X. Zhang, W. Li, A. K. Manrai, J. S. Liu, R. Chen, and X. S. Liu, "Model-based analysis of two-color arrays (MA2C)," *Genome Biol,* vol. 8, p. R178, 2007.

[21] W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu, "Model-based analysis of tiling-arrays for ChIP-chip," *Proc Natl Acad Sci U S A,* vol. 103, pp. 12457-62, Aug 15 2006.

[22] M. J. Buck, A. B. Nobel, and J. D. Lieb, "ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data," *Genome Biol,* vol. 6, p. R97, 2005.

[23] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras, "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs," *Cell,* vol. 116, pp. 499-509, Feb 20 2004.

[24] P. Zhao, J. Seo, Z. Wang, Y. Wang, B. Shneiderman, and E. P. Hoffman, "In vivo filtering of in vitro expression data reveals MyoD targets," *C R Biol,* vol. 326, pp. 1049-65, Oct-Nov 2003.

[25] Q. Zhou and W. H. Wong, "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling," *Proc Natl Acad Sci U S A,* vol. 101, pp. 12114-9, Aug 17 2004.

[26] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford, "Computational discovery of gene modules and regulatory networks," *Nat Biotechnol,* vol. 21, pp. 1337-42, Nov 2003.

[27] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat Rev Genet,* vol. 5, pp. 276-87, Apr 2004.

[28] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res,* vol. 34, pp. D108-10, Jan 1 2006.

[29] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics,* vol. 16, pp. 16-23, Jan 2000.

[30] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc Natl Acad Sci U S A,* vol. 98, pp. 4569-74, Apr 10 2001.

[31] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature,* vol. 415, pp. 141-7, Jan 10 2002.

[32] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H. G. Kumar, M. Nagini, G. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. Prasad, and A. Pandey, "Human protein reference database--2006 update," *Nucleic Acids Res,* vol. 34, pp. D411-4, Jan 1 2006.

[33] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic Acids Res,* vol. 31, pp. 258-61, Jan 1 2003.

[34] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *Proteomics,* vol. 6, pp. 35-40, Jan 2006.

[35] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science,* vol. 298, pp. 824-7, Oct 25 2002.

[36] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proc Natl Acad Sci U S A,* vol. 100, pp. 12123-8, Oct 14 2003.

[37] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *J Comput Biol,* vol. 10, pp. 947-60, 2003.

[38] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet,* vol. 34, pp. 166-76, Jun 2003.

[39] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics,* vol. 18, pp. 261-74, Feb 2002.

[40] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells," *Genetics,* vol. 149, pp. 1633-48, Aug 1998.

[41] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Mol Syst Biol,* vol. 3, p. 140, 2007.

[42] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics,* vol. 18 Suppl 1, pp. S233-40, 2002.

[43] C. Auffray, "Protein subnetwork markers improve prediction of cancer outcome," *Mol Syst Biol,* vol. 3, p. 141, 2007.

[44] W. F. Symmans, J. Liu, D. M. Knowles, and G. Inghirami, "Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions," *Hum Pathol,* vol. 26, pp. 210-6, Feb 1995.

[45] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics,* vol. 21, pp. 171-8, Jan 15 2005.

[46] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science,* vol. 310, pp. 644-8, Oct 28 2005.

[47] T. Bo and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol,* vol. 3, p. RESEARCH0017, 2002.

[48] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A,* vol. 102, pp. 15545-50, Oct 25 2005.

[49] R. K. Curtis, M. Oresic, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *Trends Biotechnol,* vol. 23, pp. 429-35, Aug 2005.

[50] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data*. CA Duxbury Press, 1997.

[51] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proc Natl Acad Sci U S A,* vol. 102, pp. 12837-42, Sep 6 2005.

[52] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talon, "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments," *Bioinformatics,* vol. 22, pp. 1096-102, May 1 2006.

[53] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *App. Statist.,* vol. 28, pp. 100 - 108, 1978.

[54] T. Kohonen, *Self-Organizing Maps*. NY: Springer, 1997.

[55] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat Rev Cancer,* vol. 8, pp. 37-49, Jan 2008.

[56] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, "Reverse engineering of regulatory networks in human B cells," *Nat Genet,* vol. 37, pp. 382-90, Apr 2005.

[57] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics,* vol. 18, pp. 51-60, Jan 2002.

[58] G. Hori, M. Inoue, S. Nishimura, and H. Nakahara, "Blind gene classification on ICA of microarray data," in *ICA*, San Diego, CA, 2001, pp. 332 - 336.

[59] S. I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biol,* vol. 4, p. R76, 2003.

[60] S. A. Saidi, C. M. Holland, D. P. Kreil, D. J. MacKay, D. S. Charnock-Jones, C. G. Print, and S. K. Smith, "Independent component analysis of microarray data in the study of endometrial cancer," *Oncogene,* vol. 23, pp. 6677-83, Aug 26 2004.

[61] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*: John Wiley & Sons, 2001.

[62] T. Gong, J. Xuan, C. Wang, H. Li, E. Hoffman, R. Clarke, and Y. Wang, "Gene module identification from microarray data using nonnegative independent component analysis," *Gene Regulation and Systems Biology,* vol. 1, pp. 349-363, 2007.

[63] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," *Proc Natl Acad Sci U S A,* vol. 100, pp. 15522-7, Dec 23 2003.

[64] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proc Natl Acad Sci U S A,* vol. 100, pp. 3339-44, Mar 18 2003.

[65] J. G. Joung, D. Shin, R. H. Seong, and B. T. Zhang, "Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation," *Bioinformatics,* vol. 22, pp. 2005-11, Aug 15 2006.

[66] C. Wang, L. Chen, P. Zhao, E. Hoffman, Y. Wang, R. Clarke, and J. Xuan, "Motif-directed network component analysis for regulatory network inference," in *Sixth International Conference on Bioinformatics*, Hong Kong, China, 2007.

[67] Hyvarinen A and O. E, "A fast fixed-point algorithm for independent component analysis," *Neural Compuatation,* vol. 9, pp. 1483-1492, 1997.

[68] A. Frigyesi, S. Veerla, D. Lindgren, and M. Hoglund, "Independent component analysis reveals new and biologically significant structures in micro array data," *BMC Bioinformatics,* vol. 7, p. 290, 2006.

[69] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database," *Nucleic Acids Res,* vol. 31, pp. 51-4, Jan 1 2003.

[70] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, "MATCH: A tool for searching transcription factor binding sites in DNA sequences," *Nucleic Acids Res,* vol. 31, pp. 3576-9, Jul 1 2003.

[71] Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufmann, 2000.

[72] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *J. R. Statist. Soc. B,* vol. 57, No. 1, pp. 289-300, 1995.

[73] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B,* pp. 479-498, 2002.

[74] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell,* vol. 9, pp. 3273-97, Dec 1998.

[75] M. Shih Ie, J. J. Sheu, A. Santillan, K. Nakayama, M. J. Yen, R. E. Bristow, R. Vang, G. Parmigiani, R. J. Kurman, C. G. Trope, B. Davidson, and T. L. Wang, "Amplification of a chromatin remodeling gene, Rsf-1/HBXAP, in ovarian carcinoma," *Proc Natl Acad Sci U S A,* vol. 102, pp. 14004-9, Sep 27 2005.

[76] C. Wang, J. Xuan, T. Gong, R. Clarke, E. Hoffman, and Y. Wang, "Stability Based Dimension Estimation of ICA with Application to Microarray Data Analysis," in *The International Conference on Bioinformatics & Computational Biology*, 2007.

[77] J. Y. Huang, B. J. Shen, W. H. Tsai, and S. C. Lee, "Functional interaction between nuclear matrix-associated HBXAP and NF-kappaB," *Exp Cell Res,* vol. 298, pp. 133-43, Aug 1 2004.

[78] M.-L. Karin, "The Fos family of transcription factors and their role in tumourigenesis, European journal of cancer," *European journal of cancer,* vol. 41, pp. 2449-2461, 2005.

[79] S. C. Sharma and J. S. Richards, "Regulation of AP1 (Jun/Fos) factor expression and activation in ovarian granulosa cells. Relation of JunD and Fra2 to terminal differentiation," *J Biol Chem,* vol. 275, pp. 33718-28, Oct 27 2000.

[80] L. F. Lee, R. P. Hellendall, Y. Wang, J. S. Haskill, N. Mukaida, K. Matsushima, and J. P. Ting, "IL-8 reduced tumorigenicity of human ovarian cancer in vivo due to neutrophil infiltration," *J Immunol,* vol. 164, pp. 2769-75, Mar 1 2000.

[81] L. Xu, "Ovarian cancer angiogenesis, biology and therapy," in *Bimedical*. vol. Ph.D.: University of Texas, 2000.

[82] P. Topilko, S. Schneider-Maunoury, G. Levi, A. Trembleau, D. Gourdji, M. A. Driancourt, C. V. Rao, and P. Charnay, "Multiple pituitary and ovarian defects in Krox-24 (NGFI-A, Egr-1)-targeted mice," *Mol Endocrinol,* vol. 12, pp. 107-22, Jan 1998.

[83] R. Hayami, K. Sato, W. Wu, T. Nishikawa, J. Hiroi, R. Ohtani-Kaneko, M. Fukuda, and T. Ohta, "Down-regulation of BRCA1-BARD1 ubiquitin ligase by CDK2," *Cancer Res,* vol. 65, pp. 6-10, Jan 1 2005.

[84] J. Ihmels, S. Bergmann, and N. Barkai, "Defining transcription modules using large-scale gene expression data," *Bioinformatics,* vol. 20, pp. 1993-2003, Sep 1 2004.

[85] W. Wang, J. M. Cherry, Y. Nochomovitz, E. Jolly, D. Botstein, and H. Li, "Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation," *Proc Natl Acad Sci U S A,* vol. 102, pp. 1998-2003, Feb 8 2005.

[86] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R. M. Karp, "CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments," *Bioinformatics,* vol. 19 Suppl 1, pp. i283-91, 2003.

[87] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor, "Computational detection of cis -regulatory modules," *Bioinformatics,* vol. 19 Suppl 2, pp. ii5-14, Oct 2003.

[88] Y. Qi and H. Ge, "Modularity and dynamics of cellular networks," *PLoS Comput Biol,* vol. 2, p. e174, Dec 29 2006.

[89] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science,* vol. 298, pp. 799-804, Oct 25 2002.

[90] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, "Predicting expression patterns from regulatory sequence in Drosophila segmentation," *Nature,* vol. 451, pp. 535-40, Jan 31 2008.

[91] F. Gao, B. C. Foat, and H. J. Bussemaker, "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data," *BMC Bioinformatics,* vol. 5, p. 31, Mar 18 2004.

[92] J. Ruan and W. Zhang, "A bi-dimensional regression tree approach to the modeling of gene expression regulation," *Bioinformatics,* vol. 22, pp. 332-40, Feb 1 2006.

[93] D. Das, Z. Nahle, and M. Q. Zhang, "Adaptively inferring human transcriptional subnetworks," *Mol Syst Biol,* vol. 2, p. 2006 0029, 2006.

[94] T. Yu and K. C. Li, "Inference of transcriptional regulatory network by two-stage constrained space factor analysis," *Bioinformatics,* vol. 21, pp. 4033-8, Nov 1 2005.

[95] D. H. Nguyen and P. D'Haeseleer, "Deciphering principles of transcription regulation in eukaryotic genomes," *Mol Syst Biol,* vol. 2, p. 2006 0012, 2006.

[96] G. Chen, S. T. Jensen, and C. J. Stoeckert, Jr., "Clustering of genes into regulons using integrated modeling-COGRIM," *Genome Biol,* vol. 8, p. R4, 2007.

[97] O. Aparicio, J. V. Geisberg, and K. Struhl, "Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo," *Curr Protoc Cell Biol,* vol. Chapter 17, p. Unit 17 7, Sep 2004.

[98] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Res,* vol. 34, pp. W369-73, Jul 1 2006.

[99] D. S. Chekmenev, C. Haid, and A. E. Kel, "P-Match: transcription factor binding site search by combining patterns and weight matrices," *Nucleic Acids Res,* vol. 33, pp. W432-7, Jul 1 2005.

[100] S. Mahony, D. L. Corcoran, E. Feingold, and P. V. Benos, "Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome," *Genome Biol,* vol. 8, p. R84, 2007.

[101] J. S. Carroll, C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu, and M. Brown, "Genome-wide analysis of estrogen receptor binding sites," *Nat Genet,* vol. 38, pp. 1289-97, Nov 2006.

[102] H. J. Bussemaker, H. Li, and E. D. Siggia, "Regulatory element detection using correlation with expression," *Nat Genet,* vol. 27, pp. 167-71, Feb 2001.

[103] A. J. Smola and B. Scholkopf, "A Tutorial on Support Vector Regression," 1998.

[104] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support Vector Regression Machines*: MIT Press, 1997.

[105] R. G. Lomax, *Statistical Concepts: A Second Course*, 3 ed.: Mahwah, N.J. : Lawerence Erlbaum Associates, 2007.

[106] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics,* vol. 7, p. 43, 2006.

[107] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.,* vol. 58, pp. 267-288, 1996.

[108] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, pp. 14863-14868, December 8, 1998 1998.

[109] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks," *Bioinformatics,* vol. 21, pp. 3448-9, Aug 15 2005.

[110] C. J. Creighton, K. E. Cordero, J. M. Larios, R. S. Miller, M. D. Johnson, A. M. Chinnaiyan, M. E. Lippman, and J. M. Rae, "Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors," *Genome Biol,* vol. 7, p. R28, 2006.

[111] L. Bjornstrom and M. Sjoberg, "Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes," *Mol Endocrinol,* vol. 19, pp. 833-42, Apr 2005.

[112] Z. Gu, R. Y. Lee, T. C. Skaar, K. B. Bouker, J. N. Welch, J. Lu, A. Liu, Y. Zhu, N. Davis, F. Leonessa, N. Brunner, Y. Wang, and R. Clarke, "Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappaB, and cyclic AMP response element binding with acquired resistance to Faslodex (ICI 182,780)," *Cancer Res,* vol. 62, pp. 3428-37, Jun 15 2002.

[113] B. P. Gomez, R. B. Riggins, A. N. Shajahan, U. Klimach, A. Wang, A. C. Crawford, Y. Zhu, A. Zwart, M. Wang, and R. Clarke, "Human X-box binding protein-1 confers both estrogen independence and antiestrogen resistance in breast cancer cell lines," *FASEB J,* vol. 21, pp. 4013-27, Dec 2007.

[114] C. Imbriano, A. Gurtner, F. Cocchiarella, S. Di Agostino, V. Basile, M. Gostissa, M. Dobbelstein, G. Del Sal, G. Piaggio, and R. Mantovani, "Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters," *Mol Cell Biol,* vol. 25, pp. 3737-51, May 2005.

[115] T. Kakizawa, T. Miyamoto, K. Ichikawa, T. Takeda, S. Suzuki, J. Mori, M. Kumagai, K. Yamashita, and K. Hashizume, "Silencing mediator for retinoid and thyroid hormone receptors interacts with octamer transcription factor-1 and acts as a transcriptional repressor," *J Biol Chem,* vol. 276, pp. 9720-5, Mar 30 2001.

[116] W. Xing and T. K. Archer, "Upstream stimulatory factors mediate estrogen receptor activation of the cathepsin D promoter," *Mol Endocrinol,* vol. 12, pp. 1310-21, Sep 1998.

[117] P. Hayward, K. Brennan, P. Sanders, T. Balayo, R. DasGupta, N. Perrimon, and A. Martinez Arias, "Notch modulates Wnt signalling by associating with

Armadillo/beta-catenin and regulating its transcriptional activity," *Development,* vol. 132, pp. 1819-30, Apr 2005.

[118] N. Brunner, B. Boysen, S. Jirus, T. C. Skaar, C. Holst-Hansen, J. Lippman, T. Frandsen, M. Spang-Thomsen, S. A. Fuqua, and R. Clarke, "MCF7/LCC9: an antiestrogen-resistant MCF-7 variant in which acquired resistance to the steroidal antiestrogen ICI 182,780 confers an early cross-resistance to the nonsteroidal antiestrogen tamoxifen," *Cancer Res,* vol. 57, pp. 3486-93, Aug 15 1997.

[119] R. Clarke, N. Brunner, B. S. Katzenellenbogen, E. W. Thompson, M. J. Norman, C. Koppi, S. Paik, M. E. Lippman, and R. B. Dickson, "Progression from hormone dependent to hormone independent growth in MCF-7 human breast cancer cells," *Proc Natl Acad Sci,* vol. 86, pp. 3649-3653, 1989.

[120] M. A. Pratt, T. E. Bishop, D. White, G. Yasvinski, M. Menard, M. Y. Niu, and R. Clarke, "Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence," *Mol Cell Biol,* vol. 23, pp. 6887-900, Oct 2003.

[121] R. B. Riggins, J. P. Lan, Y. Zhu, U. Klimach, A. Zwart, L. R. Cavalli, B. R. Haddad, L. Chen, T. Gong, J. Xuan, S. P. Ethier, and R. Clarke, "ERRgamma mediates tamoxifen resistance in novel models of invasive lobular breast cancer," *Cancer Res,* vol. 68, pp. 8908-17, Nov 1 2008.

[122] R. B. Riggins, A. Zwart, R. Nehra, and R. Clarke, "The nuclear factor kappa B inhibitor parthenolide restores ICI 182,780 (Faslodex; fulvestrant)-induced apoptosis in antiestrogen-resistant breast cancer cells," *Mol Cancer Ther,* vol. 4, pp. 33-41, Jan 2005.

[123] Y. Zhou, C. Yau, J. W. Gray, K. Chew, S. H. Dairkee, D. H. Moore, U. Eppenberger, S. Eppenberger-Castori, and C. C. Benz, "Enhanced NF kappa B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer," *BMC Cancer,* vol. 7, p. 59, 2007.

[124] K. Kim, N. Thu, B. Saville, and S. Safe, "Domains of estrogen receptor alpha (ERalpha) required for ERalpha/Sp1-mediated activation of GC-rich promoters by estrogens and antiestrogens in breast cancer cells," *Mol Endocrinol,* vol. 17, pp. 804-17, May 2003.

[125] K. E. Luker and G. D. Luker, "Functions of CXCL12 and CXCR4 in breast cancer," *Cancer Lett,* vol. 238, pp. 30-41, Jul 8 2006.

[126] A. Pattarozzi, M. Gatti, F. Barbieri, R. Wurth, C. Porcile, G. Lunardi, A. Ratto, R. Favoni, A. Bajetto, A. Ferrari, and T. Florio, "17beta-estradiol promotes breast cancer cell proliferation-inducing stromal cell-derived factor-1-mediated epidermal growth factor receptor transactivation: reversal by gefitinib pretreatment," *Mol Pharmacol,* vol. 73, pp. 191-202, Jan 2008.

[127] A. Sala, B. Saitta, P. De Luca, M. N. Cervellera, I. Casella, R. E. Lewis, R. Watson, and C. Peschle, "B-MYB transactivates its own promoter through SP1-binding sites," *Oncogene,* vol. 18, pp. 1333-9, Feb 11 1999.

[128] P. T. Pennanen, N. S. Sarvilinna, and T. J. Ylikomi, "Gene expression changes during the development of estrogen-independent and antiestrogen-resistant growth in breast cancer cell culture models," *Anticancer Drugs,* vol. 20, pp. 51-8, Jan 2009.

[129] S. D. Andrew, P. J. Delhanty, L. M. Mulligan, and B. G. Robinson, "Sp1 and Sp3 transactivate the RET proto-oncogene promoter," *Gene,* vol. 256, pp. 283-91, Oct 3 2000.

[130] A. Boulay, M. Breuleux, C. Stephan, C. Fux, C. Brisken, M. Fiche, M. Wartmann, M. Stumm, H. A. Lane, and N. E. Hynes, "The Ret receptor tyrosine kinase pathway functionally interacts with the ERalpha pathway in breast cancer," *Cancer Res,* vol. 68, pp. 3743-51, May 15 2008.

[131] N. Hay, M. Takimoto, and J. M. Bishop, "A FOS protein is present in a complex that binds a negative regulator of MYC," *Genes Dev,* vol. 3, pp. 293-303, Mar 1989.

[132] D. G. DeNardo, H. T. Kim, S. Hilsenbeck, V. Cuba, A. Tsimelzon, and P. H. Brown, "Global gene expression analysis of estrogen receptor transcription factor cross talk in breast cancer: identification of estrogen-induced/activator protein-1-dependent genes," *Mol Endocrinol,* vol. 19, pp. 362-78, Feb 2005.

[133] T. Kordula, M. Bugno, R. E. Rydel, and J. Travis, "Mechanism of interleukin-1- and tumor necrosis factor alpha-dependent regulation of the alpha 1-antichymotrypsin gene in human astrocytes," *J Neurosci,* vol. 20, pp. 7510-6, Oct 15 2000.

[134] R. Lu, P. A. Moore, and P. M. Pitha, "Stimulation of IRF-7 gene expression by tumor necrosis factor alpha: requirement for NFkappa B transcription factor and gene accessibility," *J Biol Chem,* vol. 277, pp. 16592-8, May 10 2002.

[135] K. L. Wright, L. C. White, A. Kelly, S. Beck, J. Trowsdale, and J. P. Ting, "Coordinate regulation of the human TAP1 and LMP2 genes from a shared bidirectional promoter," *J Exp Med,* vol. 181, pp. 1459-71, Apr 1 1995.

[136] T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, C. Hopf, B. Huhse, R. Mangano, A. M. Michon, M. Schirle, J. Schlegl, M. Schwab, M. A. Stein, A. Bauer, G. Casari, G. Drewes, A. C. Gavin, D. B. Jackson, G. Joberty, G. Neubauer, J. Rick, B. Kuster, and G. Superti-Furga, "A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway," *Nat Cell Biol,* vol. 6, pp. 97-105, Feb 2004.

[137] T. Gong, J. Xuan, R. B. Riggins, Y. Wang, E. P. Hoffman, and R. Clarke, "Exploring Transcriptional Modules by Integrative Gene Clustering Guided by Transcription Factor Binding Information," in *The 2008 Intl Conference on Bioinformatics & Computational Biology* Las Vegas, Nevada,, 2008.

[138] L. Chen, J. Xuan, R. B. Riggins, Y. Wang, E. P. Hoffman, and R. Clarke, "Identification of Condition-specific Regulatory Modules by Multi-level Motif and mRNA Expression Analysis," in *The 2008 Intl Conference on Bioinformatics & Computational Biology* Las Vegas, Nevada, 2008.

[139] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, 1995.

[140] Richard O. Duda, Peter E. Hart, and D. G. Stork, *Pattern classification (2nd edition)*: Wiley, New York, 2001.

[141] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters,* vol. 15, pp. 1119-1125, November 1994.

[142] P. Somol, P. Pudil, Novovicova, and J. P. Paclik, "Adaptive floating search methods in feature selection," *Pattern Recognition Letters,* vol. 20, pp. 1157-1163, November 1999.

[143] J. Kittler, *Pattern Recognition and Signal Processing, chapter Feature set search algorithms*: Sijthoff and Noordhoff, Alphen aan den Rijn, 1978.

[144] I. Guyon, J. Weston, S. Barnihill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[145] D. Rajagopalan and P. Agarwal, "Inferring pathways from gene lists using a literature-derived network of biological relationships," *Bioinformatics,* vol. 21, pp. 788-93, Mar 2005.

[146] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Muller, "Identifying functional modules in protein-protein interaction networks: an integrated exact approach," *Bioinformatics,* vol. 24, pp. i223-31, Jul 1 2008.

[147] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics,* vol. 24, pp. 1175-82, May 1 2008.

[148] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC Bioinformatics,* vol. 10 Suppl 1, p. S21, 2009.

[149] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data," *Bioinformatics,* vol. 23, pp. 1537-44, Jun 15 2007.

[150] J. Hammersley and P. Clifford, *Markov fields on finite graphs and lattices*, 1971.

[151] H. V. Poor, *An Introduction to Signal Detection and Estimation*: Springer, 1994.

[152] J. Hollan, *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press, 1975.

[153] F. Glover, "Tabu search - part I," *ORSA Journal on computing,* vol. 1, pp. 190-206, 1989.

[154] F. Glover, "Tabu Search - Part II," *ORSA Journal on computing,* vol. 2, pp. 4 - 32, 1990.

[155] M. Dorigo, "Optimization, Learning and Natural Algorithms." vol. PhD thesis: Politechnico di Milano, 1992.

[156] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*: Wiley, 2003.

[157] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proceedings of IEEE International Conference on Neural Networks*. vol. 5, 1995, pp. 1942-1948.

[158] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, pp. 671-680, May 13 1983.

[159] V. V. Černý, "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm," *Journal of Optimization Theory and Applications,* vol. 45, pp. 41-51, 1985.

[160] L. Franconi and C. Jennison, "Comparison of a genetic algorithm and simulated annealing in an application to statistical image reconstruction," *Statistic and Computing,* vol. 7, pp. 193 - 207, 1997.

[161] C. Jennison and N. Sheehan, "Theoretical and empirical properties of the genetic algorithm as a numerical optimizer," *Journal of Computational and Graphical Statistics,* vol. 4, pp. 296 - 318, 1995.

[162] V. Granville, M. Krivanek, and J.-P. Rasson, "Simulated Annealing: A Proof of Convergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 16, pp. 652-656, 1994.

[163] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. London: Chapman and Hall, 1993.

[164] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J Comput Biol,* vol. 7, pp. 601-20, 2000.

[165] F. Chung, *Spectral Graph Theory* vol. 92: American Mathematical Society, Providence, 1997.

[166] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from label and unlabeled examples," *Journal of Machine Learning Research,* vol. 1, pp. 1-48, 2006.

[167] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui, "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *J Comput Biol,* vol. 8, pp. 37-52, 2001.

[168] C. M. Kendziorski, M. A. Newton, H. Lan, and M. N. Gould, "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles," *Stat Med,* vol. 22, pp. 3899-914, Dec 30 2003.

[169] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stumpflen, and H. W. Mewes, "CORUM: the comprehensive resource of mammalian protein complexes," *Nucleic Acids Res,* vol. 36, pp. D646-50, Jan 2008.

[170] C. V. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworth, 1979.

[171] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nat Rev Cancer,* vol. 4, pp. 177-83, Mar 2004.

[172] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med,* vol. 347, pp. 1999-2009, Dec 19 2002.

[173] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet,* vol. 365, pp. 671-9, Feb 19-25 2005.

[174] L. Chen, J. Xuan, Y. Wang, R. B. Riggins, and R. Clarke, "Network-constrained SVM for Classification," in *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications* San Diego, CA: IEEE Computer Society, 2008, pp. 60-65.

[175] L. O. Gonzalez, M. D. Corte, J. Vazquez, S. Junquera, R. Sanchez, A. C. Alvarez, J. C. Rodriguez, M. L. Lamelas, and F. J. Vizoso, "Androgen receptor expresion in breast cancer: relationship with clinicopathological characteristics of the

tumors, prognosis, and expression of metalloproteases and their inhibitors," *BMC Cancer,* vol. 8, p. 149, 2008.

[176] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: integrating information about genes, proteins and diseases," *Trends Genet,* vol. 13, p. 163, Apr 1997.

[177] G. M. Callagy, P. D. Pharoah, S. E. Pinder, F. D. Hsu, T. O. Nielsen, J. Ragaz, I. O. Ellis, D. Huntsman, and C. Caldas, "Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index," *Clin Cancer Res,* vol. 12, pp. 2468-75, Apr 15 2006.

[178] S. Loi, B. Haibe-Kains, C. Desmedt, P. Wirapati, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, K. Ryder, J. F. Reid, M. G. Daidone, M. A. Pierotti, E. M. Berns, M. P. Jansen, J. A. Foekens, M. Delorenzi, G. Bontempi, M. J. Piccart, and C. Sotiriou, "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen," *BMC Genomics,* vol. 9, p. 239, 2008.

[179] A. Brinkman, S. van der Flier, E. M. Kok, and L. C. Dorssers, "BCAR1, a human homologue of the adapter protein p130Cas, and antiestrogen resistance in breast cancer cells," *J Natl Cancer Inst,* vol. 92, pp. 112-20, Jan 19 2000.

[180] L. C. Dorssers, T. van Agthoven, A. Dekker, T. L. van Agthoven, and E. M. Kok, "Induction of antiestrogen resistance in human breast cancer cells by random insertional mutagenesis using defective retroviruses: identification of bcar-1, a common integration site," *Mol Endocrinol,* vol. 7, pp. 870-8, Jul 1993.

[181] H. Al-Husaini, D. Subramanyam, M. J. Reedijk, and S. S. Sridhar, "Notch Signaling Pathway as a Therapeutic Target in Breast Cancer," *Mol Cancer Ther,* Oct 22 2010.

[182] K. Brennan and A. M. Brown, "Is there a role for Notch signalling in human breast cancer?," *Breast Cancer Res,* vol. 5, pp. 69-75, 2003.

[183] S. Stylianou, R. B. Clarke, and K. Brennan, "Aberrant activation of notch signaling in human breast cancer," *Cancer Res,* vol. 66, pp. 1517-25, Feb 1 2006.

[184] P. Rizzo, H. Miao, G. D'Souza, C. Osipo, L. L. Song, J. Yun, H. Zhao, J. Mascarenhas, D. Wyatt, G. Antico, L. Hao, K. Yao, P. Rajan, C. Hicks, K. Siziopikou, S. Selvaggi, A. Bashir, D. Bhandari, A. Marchese, U. Lendahl, J. Z. Qin, D. A. Tonetti, K. Albain, B. J. Nickoloff, and L. Miele, "Cross-talk between notch and the estrogen receptor in breast cancer suggests novel therapeutic approaches," *Cancer Res,* vol. 68, pp. 5226-35, Jul 1 2008.

[185] A. M. Brown, "Wnt signaling in breast cancer: have we come full circle?," *Breast Cancer Res,* vol. 3, pp. 351-5, 2001.

[186] S. Y. Lin, W. Xia, J. C. Wang, K. Y. Kwong, B. Spohn, Y. Wen, R. G. Pestell, and M. C. Hung, "Beta-catenin, a novel prognostic marker for breast cancer: its roles in cyclin D1 expression and cancer progression," *Proc Natl Acad Sci U S A,* vol. 97, pp. 4262-6, Apr 11 2000.

[187] M. Kretzschmar, "Transforming growth factor-beta and breast cancer: Transforming growth factor-beta/SMAD signaling defects and cancer," *Breast Cancer Res,* vol. 2, pp. 107-15, 2000.

[188] C. M. Joffroy, M. B. Buck, M. B. Stope, S. L. Popp, K. Pfizenmaier, and C. Knabbe, "Antiestrogens induce transforming growth factor beta-mediated

immunosuppression in breast cancer," *Cancer Res,* vol. 70, pp. 1314-22, Feb 15 2010.

[189] G. Lurje and H. J. Lenz, "EGFR signaling and drug discovery," *Oncology,* vol. 77, pp. 400-10, 2009.

[190] R. Schiff, S. A. Massarweh, J. Shou, L. Bharwani, S. K. Mohsin, and C. K. Osborne, "Cross-talk between estrogen receptor and growth factor pathways as a molecular target for overcoming endocrine resistance," *Clin Cancer Res,* vol. 10, pp. 331S-6S, Jan 1 2004.

[191] M. P. Brynildsen, L. M. Tran, and J. C. Liao, "A Gibbs sampler for the identification of gene expression and network connectivity consistency," *Bioinformatics,* vol. 22, pp. 3040-6, Dec 15 2006.

[192] N. Bhardwaj, M. B. Carson, A. Abyzov, K. K. Yan, H. Lu, and M. B. Gerstein, "Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets," *PLoS Comput Biol,* vol. 6, p. e1000755, May.

[193] C. Cheng, L. M. Li, P. Alves, and M. Gerstein, "Systematic identification of transcription factors associated with patient survival in cancers," *BMC Genomics,* vol. 10, p. 225, 2009.

[194] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist,* vol. 29, pp. 1165-1188, 2001.

[195] E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel, "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity," *Nat Genet,* vol. 41, pp. 316-23, Mar 2009.

[196] A. Mitsos, I. N. Melas, P. Siminelakis, A. D. Chairakaki, J. Saez-Rodriguez, and L. G. Alexopoulos, "Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data," *PLoS Comput Biol,* vol. 5, p. e1000591, Dec 2009.